ACTAS

XIII CONGRESO GALEGO DE ESTATÍSTICA E INVESTIGACIÓN DE OPERACIÓNS

Ferrol, do 26 ao 28 de outubro de 2017

XIII SGAPEIO

ACTAS

XIII Congreso Galego de Estatística e Investigación de Operacións

Ferrol, do 26 ao 28 de outubro de 2017

Departamento de Matemáticas - Universidade da Coruña

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

ISBN: 978-84-697-6999-7

Comité organizador

PRESIDENTE
Salvador Naya Fernández
MODES. Departamento de Matemáticas. Universidade da Coruña.

SECRETARIO
Javier Tarrío Saavedra
MODES. Departamento de Matemáticas. Universidade da Coruña.

VOGAIS

Germán Aneiros Pérez
MODES. Departamento de Matemáticas. Universidade da Coruña.

Manuel Carballeira Romero
Empresa NAVANTIA. Ferrol.

María Luisa Carpente Rodríguez
MODES. Departamento de Matemáticas. Universidade da Coruña.

Julián Costa Bouzas
MODES. Departamento de Matemáticas. Universidade da Coruña.

Graciela Estévez Pérez
MODES. Departamento de Matemáticas. Universidade da Coruña.

Rubén Fernández Casal
MODES. Departamento de Matemáticas. Universidade da Coruña.

Mario Francisco Fernández
MODES. Departamento de Matemáticas. Universidade da Coruña.

María José Ginzo Villamayor
Departamento de Estatística, Análise Matemática e Optimización.
Universidade de Santiago de Compostela.

María Amalia Jácome Pumar
MODES. Departamento de Matemáticas. Universidade da Coruña.

María José Lombardía Cortiña
MODES. Departamento de Matemáticas. Universidade da Coruña.

Silvia María Lorenzo Freire
MODES. Departamento de Matemáticas. Universidade da Coruña.

Manuel Antonio Presedo Quindimil
MODES. Departamento de Matemáticas. Universidade da Coruña.

José Antonio Vilar Fernández
MODES. Departamento de Matemáticas. Universidade da Coruña.

Comité científico

Ricardo Cao Abad
MODES. Departamento de Matemáticas. Universidade da Coruña.

María Jesús Casado Barrio
I.E.S Francisco Daviña Rey de Monforte de Lemos. Xunta de Galicia.

Rosa María Crujeiras Casais
Departamento de Estatística, Análise Matemática e Optimización.
Universidade de Santiago de Compostela.

Jacobo de Uña Álvarez
Departamento de Estatística e Investigación Operativa. Universidade de Vigo.

Ignacio García Jurado
MODES. Departamento de Matemáticas. Universidade da Coruña.

Wenceslao González Manteiga
Departamento de Estatística, Análise Matemática e Optimización.
Universidade de Santiago de Compostela.

Luis Alberto Ramil Novo
Departamento de Estatística, Análise Matemática e Optimización.
Universidade de Santiago de Compostela.

Covadonga Rodríguez Moldes
I.E.S. de Mugardos. Xunta de Galicia.

Antonio Vaamonde Liste
Departamento de Estatística e Investigación Operativa. Universidade de Vigo.

Juan Manuel Vilar Fernández
MODES. Departamento de Matemáticas. Universidade da Coruña.

Entidades colaboradoras





























Índice

Conferencias plenarias

BIG DATA, MÁS ALLÁ DE LA CLASIFICACIÓN Y LA REGRESIÓN 3
Amparo Alonso Betanzos

A ESTATÍSTICA iO'h NO PARADIGMA DA INDUSTRIA 4.0 4
Ricardo Cao

OPTIMIZACIÓN MATEMÁTICA AL SERVICIO DEL ANÁLISIS DE DATOS Y SU VISUALIZACIÓN 5
Emilio Carrizosa

INFERENCE BASED ON DATA FROM SUPERPOSITION OF IDENTICAL RENEWAL PROCESSES 6
Luis A. Escobar

Comunicaciones

OVERCOMING INFEASIBILITY IN BBC MODELS UNDER DELETED DOMAIN, WITH AN APPLICATION 9
Pilar Alberca, Luís Cavique, Jorge Santos

ALLOCATING THE COSTS OF CLEANING A RIVER: ESTIMATING RESPONSIBILITIES VERSUS INCENTIVE COMPATIBILITY 14
Jorge Alcalde-Unzu, María Gómez-Rúa, Elena Molis

CONSTRUCTING OPTIMAL TESTS FOR CIRCULAR SYMMETRY 17
Jose Ameijeiras-Alonso, Christophe Ley, Arthur Pewsey, Thomas Verdebout

O BOOTSTRAP PARA A ESTIMACIÓN NON PARAMÉTRICA DA FUNCIÓN DE RAZÓN DE FALLO 25
Inés Barbeito, Ricardo Cao

INTERVALOS DE CONFIANZA BOOTSTRAP PARA A DENSIDADE CONDICIONAL EN PROCESOS DE MARKOV 37
Inés Barbeito, Ricardo Cao, Dimitris Politis

HART AND MAS-COLELL CONSISTENCY IN PERT PROBLEMS 44
Gustavo Bergantiños, Alfredo Valencia-Toledo, Juan Vidal-Puga

USING FRACTIONAL REGRESSION MODELS TO ANALYSE EFFICIENCY OF EUROPEAN MANUFACTURING FIRMS 56
Bhaswar Chakma, Joaquim J.S. Ramalho, Jorge M. Azevedo Santos

ESTIMACIÓN NON PARAMÉTRICA DA MEDIA PARA DATOS NESGADOS DE GRAN VOLUME 66
Laura Borrajo, Ricardo Cao

CHARACTERISATION OF POINT PROCESSES WITH COVARIATES THROUGH THE FIRST-ORDER INTENSITY FUNCTION 77
M.I. Borrajo, W. González-Manteiga, M.D. Martínez-Miranda

ESTUDIO DE LOS INCENDIOS FORESTALES EN GALICIA MEDIANTE UN MODELO MIXTO DE POISSON ESPACIO-TEMPORAL 89
Miguel Boubeta, María José Lombardía, Manuel Marey-Pérez, Domingo Morales

STATISTICS IN "PERCEPTUAL DIALECTOLOGY": FROM THE BEGINNING TO THE FUTURE 90
Laura Calaza Díaz, Rosa M. Crujeiras, Soraya Suárez Quintas

UN PROBLEMA DE FLOTA DE VEHÍCULOS CON REPLANIFICACIONES PERIÓDICAS 98
Luisa Carpente, Julián Costa, Manuel Fontenla, Ignacio García-Jurado, Silvia Lorenzo-Freire, Guido Novoa

A PLUG-IN BANDWIDTH SELECTOR FOR LOCAL LINEAL QUANTILE REGRESSION 102
Mercedes Conde-Amboage, César Sánchez-Sellero

PLUGIN DE R-COMMANDER PARA O ENSINO DE ESTATÍSTICA BÁSICA 114
Tomás R. Cotos-Yáñez, Manuel Alfredo Mosquera, Ana Pérez-González, Benigno Reguengo Lareo

CLASIFICACIÓN DE ACEITES COMESTIBLES SEGÚN SU COMPOSICIÓN Y ESTABILIDAD A LA OXIDACIÓN 117
Ana María Díaz Díaz, Jorge López Beceiro, Ramón Artiaga, Javier Tarrío Saavedra

A COMPOSITIONAL APPROACH TO PREDICT SMALL AREA PROPORTIONS 124
María Dolores Esteban, María José Lombardía, Esther López-Vizcaíno, Domingo Morales

A TRIVARIATE FAY-HERRIOT MODEL TO PREDICT SMALL AREA POVERTY INDICATORS 126
María Dolores Esteban, María José Lombardía, Esther López-Vizcaíno, Domingo Morales, Agustín Pérez

COMPARING ROC CURVES IN THE PRESENCE OF COVARIATES 128
Arís Fanjul Hevia, Wenceslao González Manteiga,
Juan Carlos Pardo Fernández

MOBILIDADE E COÑECEMENTO DE IDIOMAS DOS TITULADOS NO SISTEMA UNIVERSITARIO DE GALICIA NOS CURSOS 2011-2012 E 2012-2013 132
Mª Carmen Fernández Montes, José Eduardo López Pereira, Mª Patrocinio Morrondo Pelayo, Pedro Faraldo Roca, Rosa María Crujeiras Casais, María José Lombardía Cortiña, Salvador Naya Fernández, José María Matías Fernández, Verónica Carreira Figueiras, Belén María Fernández de Castro, Adela Martínez Calvo

A INSERCIÓN LABORAL DOS TITULADOS EN MÁSTER: MOTIVACIÓN E VALORACIÓN DO TÍTULO 138
Mª Carmen Fernández Montes, José Eduardo López Pereira, Mª Patrocinio Morrondo Pelayo, Jose Carlos de Miguel Domínguez, David Rodríguez González, Beatriz Valcárcel Aguiar, Alexandre Vecino Aguirre, Luís Pedro Pereira Andrade, Manuel Guisado Tato

ILS: UN PAQUETE EN R PARA REALIZAR ESTUDIOS INTERLABORATORIO 146
Miguel Flores, Salvador Naya, Javier Tarrío-Saavedra,
Rubén Fernández-Casal

ANÁLISIS ESTADÍSTICO DE LOS INCENDIOS DE ALBERTA 158
Laura Freijeiro Fernández, María José Ginzo Villamayor

COORDENADAS XEOGRÁFICAS DE UNIDADES ESTATÍSTICAS E A SÚA ASIGNACIÓN: UN CASO TÍPICO 167
J.R. García Olano, C.L. Iglesias Patiño, M.E. López Vizcaíno

ANÁLISE ESPAZO-TEMPORAL DOS APELIDOS TOPONÍMICOS EN GALICIA 171
María José Ginzo Villamayor, Rosa M. Crujeiras Casais

UN NOVO INDICADOR DE DESEMPEÑO CON BASE NUNHA VELLA IDEA 181
Carlos L. Iglesias Patiño

CONCORDANCIA ENTRE AS CUALIFICACIÓNS DO ALUMNADO E DO PROFESOR NOS OBRADOIROS DE MOODLE 185
María del Carmen Iglesias Pérez, Juan Vidal Puga

GOODNESS-OF-FIT TESTS FOR DISORDER DETECTION IN NGS EXPERIMENTS 192
Norman Jiménez Otero, Jacobo de Uña Álvarez

A XESTIÓN DE RISCOS EN PROXECTOS DE EÓLICA MARIÑA MEDIANTE SIMULACIÓN DE EVENTOS DISCRETOS E SIMULACIÓN DE MONTECARLO 199
Adolfo Lamas Rodriguez, David Chas Alvarez, José Antonio Muiña Dono

NONPARAMETRIC COVARIATE SIGNIFICANCE TESTS IN MIXTURE CURE MODELS 208
Ana López Cheda, Ricardo Cao, M. Amalia Jácome

COMPARACIÓN DE MODELOS DE REGRESIÓN PARA LA DETERMINACIÓN DE LA VIDA ÚTIL DE PESCADO CONGELADO 219
M.M. López García, L.A. Ramil Novo, M.A. Romero Rodríguez, M.L. Vázquez Odériz

ESTRATEXIA DE XEORREFERENCIACIÓN DA POBOACIÓN DE GALICIA EMPREGANDO TÉCNICAS ESTATÍSTICAS 221
Mª Esther López Vizcaíno, Carlos L. Iglesias Patiño

A TESTING METHOD FOR LINEAR SPATIAL TRENDS 227
Andrea Meilán-Vila, Mario Francisco-Fernández, Rosa M. Crujeiras

ESTUDIO DE UN PROBLEMA DE PLANIFICACIÓN DE TAREAS EN UNA EMPRESA DE AYUDA A DOMICILIO 232
Isabel Méndez-Fernández, Luisa Carpente, Julián Costa-Bouzas, Ignacio García-Jurado, Silvia Lorenzo-Freire, Guido Novoa-Flores

BIG DATA EN PROBLEMAS DE CLASIFICACIÓN: APRENDENDO A COMPARTIMENTAR OS DATOS 236
Pablo Montero-Manso, José A. Vilar

THE PAINTING RULE IN MINIMUM COST SPANNING TREE PROBLEMS WITH MULTIPLE SOURCES 237
Adriana Navarro Ramos, Gustavo Bergantiños

OPTIMIZACIÓN DE LA PRODUCCIÓN EN UNA FÁBRICA DE PIENSOS 241
Guido Ignacio Novoa-Flores, Luisa Carpente, Silvia Lorenzo-Freire

PREDICTION INTERVALS FOR QUANTILE AUTOREGRESSION 249
Silvia Novo-Díaz, César Sánchez-Sellero

MODELIZACIÓN ESTADÍSTICA DEL PROCESO DE ENHEBRADO EN CONSTRUCCIÓN NAVAL 256
Déborah Otero, Ricardo Cao, Vicente Blasco, Javier Tarrío-Saavedra, Salvador Naya

DETECCIÓN DE ERRORES DE MEDIDA EN CONSTRUCCIÓN NAVAL 262
Déborah Otero, Ricardo Cao, Álvaro Brage, Javier Tarrío-Saavedra, Salvador Naya

A NOVEL DISTRIBUTED MULTIMETHOD TO SOLVE CHALLENGING GLOBAL OPTIMIZATION PROBLEMS IN COMPUTATIONAL SYSTEMS BIOLOGY 272
David R Penas, Patricia González, Ramón Doallo, Julio R Banga

NUEVAS NECESIDADES DE PREVISIÓN EN LOS ACTUALES MARCOS PRESUPUESTARIOS 274
Marcos Rodríguez Rey

UN MODELO DE PROGRAMACIÓN MATEMÁTICA PARA LA SELECCIÓN Y ASIGNACIÓN TEMPORAL DE RECURSOS AÉREOS DESTINADOS A LA CONTENCIÓN DE UN INCENDIO FORESTAL 281
Jorge Rodríguez-Veiga, María José Ginzo-Villamayor, Balbina Casas-Méndez

LOS VALORES PERDIDOS EN EL MUESTREO DE POBLACIONES FINITAS. TÉCNICAS DE IMPUTACIÓN 293
Gabriel Román Radío

ANALYSIS OF INCIDENCE OF CHRONIC GRANULOCYTIC LEUKEMIA AND PRIMARY BILIARY CIRRHOSIS USING LEVEL SET ESTIMATION 299
Paula Saavedra-Nieves, Alberto Rodríguez-Casal

OPTIMAL DESIGNS IN ENZYME REACTIONS WITH INHIBITION BY HIGH SUBSTRATE CONCENTRATIONS 301
M.Teresa Santos-Martín, M. Jesús Rivas-López

CARGA DE POBOACIÓN ESTACIONAL DOS CONCELLOS DE GALICIA 306
Pablo Torrado Solar, María Esther López Vizcaíno

AUTO-DUALIDADE EN PROBLEMAS DE ALUGUER DE TERREOS 315
Alfredo Valencia-Toledo, Juan Vidal-Puga

IMPUTACIÓN DE DATOS DE PRECIPITACIÓN DIEZMINUTALES EN LA RED DE ESTACIONES DE METEOGALICIA 324
José Vidal Paz, Benigno Rodríguez-Gómez

Conferencias plenarias

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

BIG DATA, MÁS ALLÁ DE LA CLASIFICACIÓN Y LA REGRESIÓN

Amparo Alonso Betanzos[1]

[1]Departamento de Computación, Facultad de Informática, Universidad de A Coruña, Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA)

RESUMEN

Los avances en las TIC han contribuido a la proliferación de grandes bases de datos, de manera que por ejemplo, en 2012 se generaban unos 2.5 exabytes de datos/día, y se estima que esta cantidad se duplicará cada 40 meses. Este notable crecimiento del volumen de datos, junto con otros problemas relacionados con el llamado Big Data (variedad, variabilidad, veracidad, etc.), hace que muchos de los algoritmos de Aprendizaje Computacional (AC) clásicos no sean capaces de funcionar adecuadamente, bien porque ofrecen un pobre rendimiento o porque tardarían en ser ejecutados una cantidad de tiempo inasumible para muchas aplicaciones. Debido a esto, la escalabilidad de los algoritmos ha pasado de ser una cualidad deseable, a ser una característica prácticamente imprescindible si se pretende que sean útiles en los escenarios reales que se avecinan. Los primeros esfuerzos de la comunidad que trabaja en Big Data se centraron en los dos problemas más tradicionales del aprendizaje supervisado: la clasificación y la regresión. De hecho, a pesar de que las técnicas de Big Data son relativamente recientes, se dispone ya de librerías que incluyen versiones escalables de algoritmos de AC comunes, incluyendo algoritmos de clasificación, regresión y clustering. En los últimos años, existen otros algoritmos con una alta aplicabilidad, que probablemente serán la siguiente "ola" de escalabilidad, como puede ser el caso de los algoritmos de ranking (capaces de ordenar total o parcialmente conjuntos de objetos); algoritmos de una sola clase (aplicables, por ejemplo, a la detección de anomalías); algoritmos de cuantificación, en los que lo que se pretende es estimar la prevalencia de ciertas clases de objetos en una muestra; y algoritmos de selección de características adaptados a los tipos de problemas antes citados. Sobre estas nuevos tipos de problemas versará esta charla.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

A ESTATÍSTICA iO'h NO PARADIGMA DA INDUSTRIA 4.0

Ricardo Cao[1]

[1]Grupo de investigación de Modelización, Optimización e Inferencia Estatística (MODES), Departamento de Matemáticas, Facultade de Informática, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC) e Instituto Tecnolóxico de Matemática Industrial (ITMATI), Universidade da Coruña

RESUMO

O termo "Industria 4.0" utilízase para aludir á dixitalización das cadeas de valor industriais. Trátase de facer uso das tecnoloxías emerxentes facilitadoras (FET) para implementar o Internet das Cousas (IoT) e os servizos co obxecto de integrar procesos de enxeñaría e negocio. Así a produción pode levarse a cabo eficientemente e de xeito flexible, con baixos custes e alta calidade. Neste paradigma da chamada cuarta revolución industrial, a Estatística (e tamén a Investigación de Operacións!) está chamada a xogar un papel clave, como ferramenta indispensable para converter datos en información útil para a mellora dos procesos produtivos e para engadir valor aos produtos. Nesta conferencia percorreranse algúns problemas reais existentes en empresas no ámbito industrial, comercial e dos servizos, que foron abordados nos últimos anos polos membros do grupo de investigación MODES mediante o uso de modelos e técnicas estatísticas axeitados. É o que poderiamos bautizar como a Estatística iO'h.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

OPTIMIZACIÓN MATEMÁTICA AL SERVICIO DEL ANÁLISIS DE DATOS Y SU VISUALIZACIÓN

Emilio Carrizosa[1]

[1]IMUS - Instituto de Matemáticas de la Universidad de Sevilla

RESUMEN

Es bien conocido que los procedimientos usuales en Análisis de Datos acaban de una manera u otra en un problema de optimización: análisis en componentes principales, escalado multidimensional, análisis de conglomerados, clasificación supervisada son solo algunos ejemplos. El espectacular avance en los métodos de la Optimización Matemática (en particular, de la Programación No Lineal en Números Enteros) permite abordar exitosamente nuevos problemas, o abordar problemas tradicionales de modo mucho más eficiente que hace una década. En esta charla haremos una revisión de algunos de los campos del Análisis de Datos y su visualización en los que el tradicional papel de la Optimización nos ofrece estimulantes soluciones.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

INFERENCE BASED ON DATA FROM SUPERPOSITION OF IDENTICAL RENEWAL PROCESSES

Luis A. Escobar[1]

[1]Department of Experimental Statistics, Louisiana State University, Baton Rouge, Louisiana USA, luis@lsu.edu

ABSTRACT

Maintenance data can be used to make inferences about the lifetime distribution of system components. Typically a fleet contains multiple systems. Within each system there is a set of nominally identical replaceable components of particular interest (e.g., two automobile headlights, eight DIMM modules in a computing server, sixteen cylinders in a locomotive engine). For each component replacement event, there is system-level information that a component was replaced, but not information on which particular component was replaced. Thus the observed data is a collection of superposition of identical renewal processes (SRP), one for each system in the fleet. This paper proposes a procedure for estimating the component lifetime distribution using the aggregated event data from a fleet of systems. We show how to compute the likelihood function for the collection of SRPs and provide suggestions for efficient computations. We compare performance of this incomplete-data ML estimator with the complete-data ML estimator and study the performance of confidence interval methods for estimating quantiles of the lifetime distribution of the component.
This joint work with Ye Tian, Wei Zhang, and William Q. Meeker.

Keywords: Maintenance Data; Maximum likelihood; Recurrence data; Reliability; Weibull.

Comunicaciones

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

OVERCOMING INFEASIBILITY IN BCC MODELS UNDER DELETED DOMAIN, WITH AN APPLICATION.

Pilar Alberca[1] Luís Cavique[2] Jorge Santos[3]

[1] Departamento de Economía de la Empresa y Contabilidad, UNED, palberca@cee.uned.es
[2] MAS-BioISI FCUL and Universidade Aberta, luis.cavique@uab.pt
[3] Departamento de Matemática, CIMA, Escola de Ciências e Tecnologia, Universidade de Évora, jmas@uevora.pt

ABSTRACT

It is a well-known fact that the Banker Charnes and Cooper models under deleted domain can originate unfeasible linear programs for extreme observations. In this paper, we introduce a way to overcome this issue introducing some extra constraints in the multiplier model. Finally, we apply these models to data of the Spanish hotel industry.

Keywords: Data Envelopment Analysis, Efficiency Analysis, Variable Returns to Scale, Linear Programming.

1. INTRODUCTION

Data Envelopment Analysis is a non-parametric efficiency measurement technique introduced in 1978 by Charnes, Cooper and Rhodes (Charnes et al, 1978), although its roots may be found as early as 1957 in Farrell's seminal work (Farrell, 1957). It deserves special attention also the work of the Dutch Nobel Prized Tjalling Koopmans' and its "Activity Analysis Concepts" (Koopmans, T., 1951).

The Charnes, Cooper and Rhodes model is a Constant Returns to Scale and only in 1984 Banker Charnes and Cooper (Banker et al, 1984) developed a model for Variable Returns to Scale. Both models are always feasible, but they do not give much information about efficient units usually referred to as DMUs, the acronym of Decision Making Units.

This issue was addressed by Andersen and Petersen from Odense University in 1993 (Andersen and Petersen, 1993). Their technique known as super efficiency or deleted domain, is rather appealing because they allow efficiency to be greater than one, discriminating between efficient DMUs that otherwise are all ranked equal and avoiding ambiguity on weights of those DMUs.

This super efficiency or deleted domain technique when applied to the Banker Charnes and Cooper model results in unfeasible problems when assessing extreme DMUs, this can be solved with weights restrictions as illustrated by Alberca and Santos in 2015. Those weight restriction techniques are somehow subjective and difficult to justify, especially in the multiple input, multiple output situation. This communication presents a related technique but that comes from the interpretation of the unbounded extra variable in the variable returns to scale.

This paper unfolds as follows: In Section 2 the Banker Charnes and Cooper multiplier problem under input minimization is defined and compared to the Charnes, Cooper and Rhodes homologous one. It will be exemplified by the simplest model with just one input and one output and 12 DMUs.

In section 3 the interpretation of the extra variable is introduced only under the minimization of inputs and the extra constraints described.

Next, in section 4 the new technique is applied to a sample data of 572 audited Spanish hotel firms for the period 2011 and compared to the results previously published using the traditional weight restriction techniques.

Finally, in section 5 we draw conclusions, discussing the advantages of the super-efficiency technique and comparing this new approach to the traditional ones, namely weights restrictions. We also make some final considerations about the computational implementation of the new model since, so far, traditional packages do not allow for constraints on the variable returns to scale extra variable.

2. THE MULTIPLIER PROBLEM UNDER INPUT MINIMIZATION

In the Banker, Charnes and Cooper model, when assessing a set of J organizations, efficiency (Ef_a) of a specific decision making unit (DMU a) under **a**nalysis is defined by:

$$\text{Max} \quad Ef_a = \sum_{r=1}^{s} \mu_{ra} y_{ra} + \upsilon_a \qquad (1)$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \nu_{ia} x_{ia} = 1 \qquad (2)$$

$$-\sum_{i=1}^{m} \nu_{ia} x_{ik} + \sum_{r=1}^{s} \mu_{ra} y_{rk} + \upsilon_a \le 0 \qquad j=1...J \qquad (3)$$

$$\mu_{ra} \ \varepsilon > 0 \qquad m=1...M \qquad (4)$$

$$\nu_{ia} \ \varepsilon > 0 \qquad n=1...N \qquad (5)$$

Where X_{ik} stands for the ith input of the kth DMU, with a similar meaning for Y_{rk} , the weights μ_{ra}' and ν_{ia}', in Eq (1), are chosen for each DMU_j' under evaluation as those that maximize its efficiency.

We present an example graphically with 12 DMUs in figure 1 and the results in table 1.

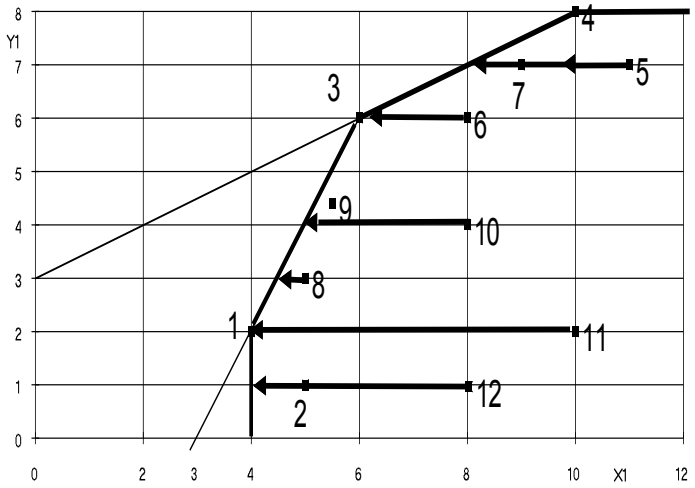


Figure 1: VRS Production Possibility Set and Efficiency Frontier for the example.

Comparing the Banker Charnes and Cooper model with its predecessor we see that it has an extra unbounded variable υ_a that is negative in the Decreasing Returns to Scale region and positive in the Increasing Returns to Scale region.

	Variables		**Constant**	**Multipliers**		**Eficiency**
DMU	x_1	y_1	υ_a	ν	μ	***Ef***
1	4.0	2.0	1.0000	0.2500	1E-06	1.000000
2	5.0	1.0	0.8000	0.2000	1E-06	0.799999
3	6.0	6.0	0.5000	0.1667	0.0833	1.000000
4	10.0	8.0	-0.6000	0.1000	0.2000	1.000000
5	11.0	7.0	-0.5455	0.0909	0.1818	0.727273
6	8.0	6.0	-0.7500	0.1250	0.2500	0.750000
7	9.0	7.0	-0.6667	0.1111	0.2222	0.888889
8	5.0	3.0	0.6000	0.2000	0.1000	0.900000
9	5.5	4.4	0.5455	0.1818	0.0909	0.945455
10	8.0	4.0	0.3750	0.1250	0.0625	0.625000
11	10.0	2.0	0.3000	0.1000	0.0500	0.400000
12	8.0	1.0	0.5000	0.1250	1E-06	0.499999

Table 1: Results for the 12 units under input minimization.

3. INTERPRETATION OF THE EXTRA VARIABLE

First, we interpret geometrically the extra variable by means of figure 2 where we show that the BCC model consists of the intersection of four cones. In other words: The BCC model can be considered as several CCR models with an arbitrary origin. By instance, DMU 4 is not CCR efficient, but if we take point (X=-6; Y=0) as the origin it will be CCR efficient. The same applies if we take point (X=0; Y=3) as the origin, or even any convex combinations of these two points.

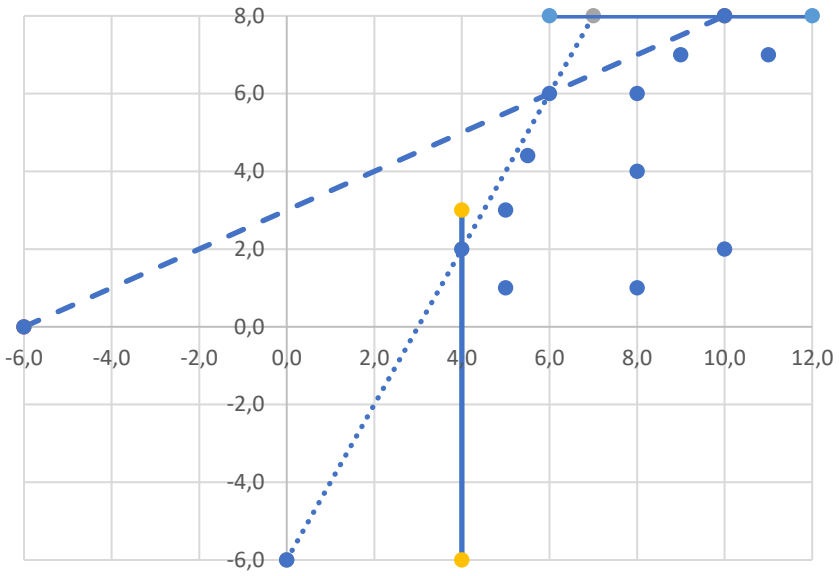


Figure 2: Decomposition of the efficiency frontier in its 5 edges.

Now we can interpret algebraically the extra variable by focusing on the shift of the vertex of the polyhedral cone.

From the previous formulation, we see that υ_a can easily be associated with a shift in output Y_r denoted as $-s_{ra}$ as follows: $\upsilon_a = \mu_{ra} s_{ra}$. and in equation (1) $\sum_{r=1}^{s} \mu_{ra} y_{ra} + \upsilon_a$ would become:

$\mu_{1a} y_{1a} + \ldots + \mu_{ra} y_{ra} + \ldots + \mu_{sa} y_{sa} + \mu_{ra} s_{ra}$ resulting in $\mu_{1a} y_{1a} + \ldots + \mu_{ra} [y_{ra} - (-s_{ra})] \ldots + \mu_{sa} y_{sa}$

so: $s_{ra} = \upsilon_a / \mu_{ra}$

We can apply this to the results in table 1 for DMU 4 by instance: $\upsilon_4 = -0.6$ $\mu_4 = 0.2$ and so $s_4 = -0.6/0.2 = -3$ and it means a shift of 3 units in the output y_1.

The interpretation we just did concerning outputs could be made easily for inputs leading to the relation: $s_{ia} = \upsilon_a / \mu_{ia}$ that relates the shift on Input x_i, when assessing DMU a s_{ia} with the extra variable υ_a and its multiplier μ_{ia}.

Exemplifying again with the results in table 1 for DMU 4 by instance we get: $\upsilon_4 = -0.6$ $\mu_4 = 0.1$ and so $s_4 = -0.6/0.1 = -6$ and it means a shift of -6 units in the input x_1.

The previous results show that DMU 4 would be CCR efficient if the origin of data was shifted to point (X=0; Y=3) or to point (X=-6; Y=0) or any convex combination of those two points.

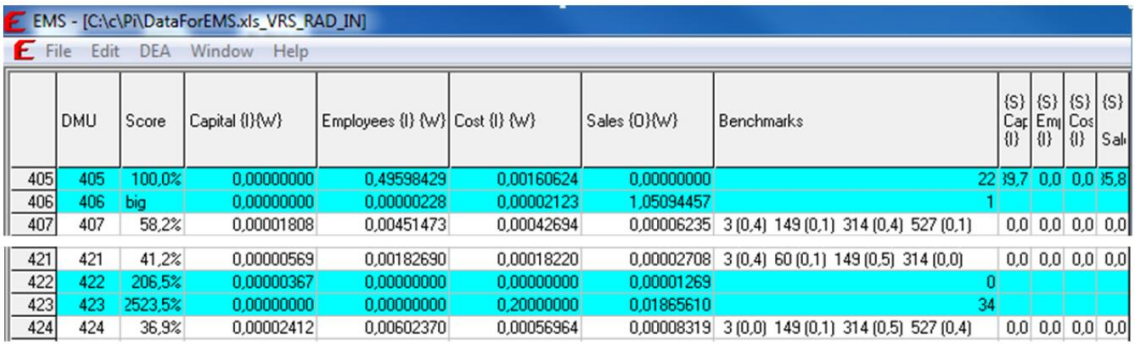
Of course, it would be possible to choose other shifts involving the two variables x_1 and y_1 simultaneously, but it would lead to more difficult interpretation of the new origin for data. This is particularly true in higher dimensionality applications, we suggest to consider always just a shift in one axis each time.

We did not refer to the super-efficiency technique for sake of clarity, but it simply means that the J constraints in equation (3) do not apply to all the J DMUs the DMU under assessment is excluded (this is the reason for the designation deleted domain) so that its efficiency score can be greater than one (this is the reason for the designation super-efficiency).

4. APPLICATION TO REAL DATA

In 2015 Alberca and Santos (Alberca and Santos,2015) analyzed the performance of 572 Spanish hotel firms for the period 2011 using one input and three output variables with the BCC model and super-efficiency technique.

In figure 3 (taken from their 2015 paper) we can see that DMU 407 originates an unfeasible LP and that DMU 423 has a very high super-efficiency.



	DMU	Score	Capital (I){W}	Employees (I) {W}	Cost (I) {W}	Sales (O){W}	Benchmarks	{S} Cap (I)	{S} Emp (I)	{S} Cos (I)	{S} Sal
405	405	100,0%	0,00000000	0,49598429	0,00160624	0,00000000	22	9,7	0,0	0,0	5,8
406	406	big	0,00000000	0,00000228	0,00002123	1,05094457	1				
407	407	58,2%	0,00001808	0,00451473	0,00042694	0,00006235	3 (0,4) 149 (0,1) 314 (0,4) 527 (0,1)	0,0	0,0	0,0	0,0
421	421	41,2%	0,00000569	0,00182690	0,00018220	0,00002708	3 (0,4) 60 (0,1) 149 (0,5) 314 (0,0)	0,0	0,0	0,0	0,0
422	422	206,5%	0,00000367	0,00000000	0,00000000	0,00001269	0				
423	423	2523,5%	0,00000000	0,00000000	0,20000000	0,01865610	34				
424	424	36,9%	0,00002412	0,00602370	0,00056964	0,00008319	3 (0,0) 149 (0,1) 314 (0,5) 527 (0,4)	0,0	0,0	0,0	0,0

Figure 3: Sample results for the BCC model with super-efficiency technique.

We applied the same model, but with constraints on the maximum allowable shift for any variable, specifically we restricted them not to exceed the maximum value of that variable, as we present in table 2.

	SALES	CAPITAL	Nr of Employees	Operational Cost
Mean	13253	46619	170	2547
Maximum	516160	2451857	6170	59017

Table 2: Mean and maximum for the 4 variables.

With this kind of constraints, we cannot use the Efficiency Measurement System software anymore, so we had to solve it with a linear solver in a standard spreadsheet. In table 3 we present the relevant values so that a comparison can be made to the results of Figure 3.

DMU	μ_1	ν_1	ν_2	ν_3	υ_a	Eff
405	1,91E-06	3,59E-08	0,499264	5,43E-06	0,986244	0,986363
406	2,75E-06	4,35E-08	0,000115	3,93E-06	-0,10656	1,311735
407	6,24E-05	1,81E-05	0,004515	0,000427	-0,01546	0,582379
421	2,71E-05	5,69E-06	0,001827	0,000182	-0,03406	0,412267
422	7,79E-06	1,86E-06	6,66E-05	6,96E-06	-0,41084	1,683675
423	2,92E-05	-6,2E-07	-0,00025	0,452153	1,525095	2,990614
424	8,32E-05	2,41E-05	0,006024	0,00057	-0,02062	0,369375

Table 3: Results for the new model.

We can remark that the unfeasibility occurring when assessing DMU 406 was removed and that all the efficiencies are lower than those presented in figure 3.

5. CONCLUSIONS

First, we must stress the advantages of the super-efficiency techniques, since they allow ranking even of the efficient DMUs, another important advantage is the unicity of the set of weights, since the solution of the envelopment problem is not degenerated anymore.

The unfeasibility that arises in some extreme circumstances can be easily removed, and we illustrated that our new model in the above example is feasible and its efficiency scores are the same for inefficient DMUs and decreases for the efficient ones, since we apply extra constraints to a maximization problem. In a previous work, Pilar and Santos removed the infeasibilities, by weights restrictions (Thompson et al,1990) but they profited that they had a single output, so bounds for weights were easily determined by a multiple linear regression as suggested by Dyson and Thanassoulis (Dyson and Thanassoulis, 1988). This new technique is easier to apply, since we can use the maximum of a variable as an upper bound for the maximum shift on its axis. This is especially interesting in a multiple input multiple output situation.

The only drawback is that there are no available packages to implement this model, so the user must rely on programming in spreadsheets or available mathematical programming language.

REFERENCES

Alberca, P., Santos, J. (2015). RESTRICTING THE WEIGHTS AND THE JUDGMENT VALUES IN DEA MODELS. XII Congreso Galego de Estatística e Investigación de Operacións. SGAPEIO e Servizo de Publicacións da Deputación de Lugo

Alberca, P., Rodríguez, A., & Parte, L. (2015). Measuring the efficiency of trade shows: A Spanish case study. Tourism Management, 47(4), 127-137.

Andersen, P., Petersen, N.C., (1993), A procedure for ranking efficient units in data envelopment analysis, Management Science, 39, (10), 1261-4.

Banker, R.D., Charnes, A. and Cooper, W. (1984) Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. Management Science, 30(9), 1078-1092.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the efficiency of decision making units. European Journal of Operational Research 2(6), 429-444.

Dyson, R. G. and Thanassoulis, E. (1988) Reducing weight flexibility in data envelopment analysis. Journal of the Operational Research Society, 563-576.

Farrell, M. J., (1957), "The measurement of productive efficiency", Journal of Royal Statistical Society A, 120, 253-281.

Koopmans, T. (1951). Analysis of production as an efficient combination of activities. In: Activity Analysis of Production and Allocation [edited by T.C. Koopmans], John Wiley & Sons, Inc., New York.

Thompson, R. G., Langemeier, L. N., Lee, C. T., Lee, E. and Thrall, R. M. (1990) The role of multiplier bounds in efficiency analysis with application to Kansas farming. Journal of econometrics, 46(1), 93-108.

Wong, Y. H., and Beasley, J. E. (1990) Restricting weight flexibility in data envelopment analysis. Journal of the Operational Research Society, 829-835.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Allocating the costs of cleaning a river: estimating responsibilities versus incentive compatibility

Jorge Alcalde-Unzu[1], María Gómez-Rúa[2] and Elena Molis[3]

[1]Universidad Pública de Navarra
[2]Universidade de Vigo
[3]Universidad de Granada

ABSTRACT

We model a river as a segment divided into subsegments, each occupied by one region, from upstream to downstream. The waste is transferred from one region to the next at some rate t. Given that t may be unknown, Alcalde-Unzu et al. (2015) proposed a method to allocate the costs of cleaning a river in which each region pays the responsibility that it would have if the transfer rate was its expected value. We prove in this paper that this allocation differs from the expected responsibility of each region and propose and characterize a new solution, the Expected Upstream Responsibility, that assigns to each region precisely its expected responsibility. However, we show that although this new solution improves the estimation of responsibilities, there is a trade-off in terms of incentive compatibility.

Keywords: Cost allocation; waste river; responsibility; incentive compatibility; characterization.

1. INTRODUCTION

Motivation and overview of results

The cleaning process of the waste present in river channels is a task that has to be undertaken frequently. Given that most rivers flow through different regions or even countries, the costs of the cleaning activities have to be divided between all the administrations involved, probably depending on the responsibilities that each one has for the waste discharged.[1]

The literature has considered four solutions that allocate the costs for these problems. On the one hand, Ni and Wang (2007) proposed the Local Responsibility Sharing (LRS) and the Upstream Equal Sharing (UES) solutions and Dong et al. (2012) proposed the Downstream Equal Sharing solution (DES). The first allocates the cost of cleaning each part of the river totally to the region located in that segment. The second considers more realistically that the regions upstream of a segment have also responsibility over the waste present in that segment and allocates the cost of cleaning each segment between all these responsible regions equally. The third solution is dual to the UES, but considering that the regions that should pay from cleaning a segment are the ones situated downstream instead of the upstreams ones. On the other hand, Alcalde-Unzu et al. (2015) showed that the social planner can infer some information on the transfer rate of the waste between two segments of the river by analyzing the cleaning costs of all segments. This information allows to deduce some limits for the responsibilities of each region over the waste discharged that is not considered in the previous solutions and, therefore, some paradoxes about the behavior of these solutions can be constructed. Then, a new solution was proposed, the Upstream Responsibility

[1]The literature has also focused on how to distribute *fairly* the benefits of the river among the different regions. See for example Ambec and Sprumont (2002), Ambec and Ehlers (2008), Wang (2011), Khmelnitskaya (2010), and van den Brink et al. (2012).

(UR), which assigns to each region its responsibility calculated with the expected value of the transfer rate.

We show in this paper that, since the responsibility is not linear in the transfer rate, the UR solution does not correspond with the expected responsibility of each region. In particular, it biases the responsibilities of the regions in the following sense: (i) the most upstream (resp. downstream) region pays less (resp. more) than its expected responsibility; and (ii) an intermediate region pays more (resp. less) than its expected responsibility if the cost of cleaning the preceding region is higher (resp. lower) than the cost of cleaning its own segment. In consequence, we propose a new solution, the Expected Upstream Responsibility (EUR), which allocates to each region its expected responsibility. We show that this new solution corresponds to the responsibility that each region would have if the transfer rate was a particular value greater than its expected value. We characterize this new solution maintaining all axioms that characterize the UR solution except one that should be adapted to the features of the new proposal.

Although the EUR solution improves its predecessor UR in terms of estimating responsibilities, we will show that there is a trade-off in terms of incentive compatibility. Meanwhile the UR solution satisfies the basic property that if a region discharges more waste, it is guaranteed that it will not pay less, this is not the case with the EUR solution. The reason is that, when a region discharges more waste, two effects appear: it increases the costs of cleaning the segments, which tends to increase the costs allocated to the regions, but, at the same time, it changes the information that the social planner can deduce on the transfer rate from the cleaning cost vector, which affects how the solution distributes all costs among the regions. If the second effect dominates the first, which can occur in the EUR solution but not in the UR one, a region discharging more waste could pay less. Moreover, we show that it is not only that the UR solution satisfies this basic incentive compatibility property, but also that there is no other solution among a general family that do it with a less bias at the time of estimating responsibilities.

Related literature

This paper is related to the literature that studies the allocation of the costs of transboundary rivers using game theoretical and/or axiomatic models, which has two main approaches. Several papers, starting with Ni and Wang (2007) and including Alcalde-Unzu et al. (2015) (both described above), assumed that the cost of cleaning each part of the river is exogenously given. In the same vein, van den Brink and van der Laan (2008) showed additional characterizations of the LRS and UES solutions and Dong et al. (2012) introduced the DES solution and extended these three solutions to the case of a river with tributaries. On the other hand, other papers such as Gengenbach et al. (2010) and van der Laan and Moes (2012) took a substantially different approach by assuming that the cost allocation method adopted may affect the decision of each region about how much waste to discharge. Our results link both approaches with the axiom of incentive compatibility.

2. CONCLUSIONS

The literature has provided some solutions for the problem of sharing the costs of cleaning a transboundary river. In a previous paper (Alcalde-Unzu et al. (2015)), we proposed the UR solution that calculates the responsibility that each region would have if the transfer rate was its expected value. This paper has shown that this solution does not estimate the expected responsibility of each region correctly and we have proposed and characterized the new EUR solution that assigns to each region precisely its expected responsibility.

On the other hand, when a social planner has to select a solution to allocate the costs, it is desirable that such a rule does not give incentives to regions to discharge more waste if this allows them to pay less for cleaning the river. We have called this property IC. We have found that the EUR solution does not satisfy IC, while the UR solution does, and that this latter one is the closest solution in a particular family to the EUR solution that satisfies this property. Therefore, the social planner has a trade-off between a better estimation of responsibilities and the fulfillment of IC. If she prefers to satisfy the IC axiom, she should select the UR solution, although she

would be committing a bias at the estimation of responsibilities: the most upstream region would pay less than its expected responsibility, the most downstream region would pay more than its expected responsibility, while the intermediate regions would pay more or less depending on the cost structure. If, however, the social planner prefers a better estimation of responsibilities, she should select the new EUR solution, although this could be giving incentives to regions to discharge more waste in some occasions.

REFERENCIAS

Alcalde-Unzu, J., Gómez-Rúa, M., Molis, E. (2015). Sharing the costs of cleaning a river: the Upstream Responsibility rule. Game Econ. Behav. 90, 134-150.

Ambec, S., Ehlers, L. (2008). Sharing a river among satiable agents. Game Econ. Behav. 64, 35-50.

Ambec, S., Sprumont, Y. (2002). Sharing a river. J. Econ. Theory 107, 453-462.

Dong, B., Ni, D., Wang, Y. (2012). Sharing a polluted river network. Environ. Resour. Econ. 53, 367-387.

Gengenbach, M.F., Weikard, H.P., Ansink, E. (2010). Cleaning a river: an analysis of voluntary joint action. Nat. Resour. Model. 23, 565-589.

Khmelnitskaya, A.B. (2010). Values for rooted-tree and sink-tree digraphs games and sharing a river. Theor. Decis., 69, 657-669.

Ni, D., Wang, Y. (2007). Sharing a polluted river. Game. Econ. Behav., 60, 176-186.

van den Brink, R., van der Laan, G. (2008). Comment on "Sharing a polluted river". Mimeo.

van den Brink, R., van der Laan, G. and Moes, N. (2012). Fair agreements for sharing international rivers with multiple springs and externalities. J. Environ. Econ. Manag. 3, 388-403.

van der Laan, G., Moes, N. (2012). Transboundary externalities and property rights: an international river pollution model. Tinbergen Discussion Paper 12/006-1, Tinbergen Institute and VU University, Amsterdam.

Wang, Y. (2011). Trading water along a river. Math. Soc. Sci., 61, 124-130.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Constructing optimal tests for circular symmetry

Jose Ameijeiras-Alonso[1], Christophe Ley[2], Arthur Pewsey[3] and Thomas Verdebout[4]

[1]Dep. de Estatística, An. Mat. e Optimización. Universidade de Santiago de Compostela.
[2]Department of Applied Mathematics, Computer Science and Statistics. Ghent University.
[3]Department of Mathematics, Escuela Politécnica. University of Extremadura.
[4]Département de Mathématique. Université libre de Bruxelles.

ABSTRACT

Most statistical procedures for circular data are based on symmetric distributions (like the von Mises, wrapped Cauchy and cardioid). It is thus of paramount importance to be able to test whether a given data set is symmetric, and hence the classical tools can be used, or whether it is asymmetric, in which case more work is required. The objective of this talk is to show a way of constructing optimal tests for circular reflective symmetry, following the Le Cam methodology. In particular, the focus will be to design an optimal test against the broad class of sine-skewed alternatives when the centre of symmetry is completely unknown.

Key phrases and keywords: circular statistics, optimal testing procedure, reflective symmetry, skewed distributions.

1. INTRODUCTION

In the classical statistical analysis of univariate continuous data, knowing whether the symmetry, or more precisely reflective symmetry, is present in a random sample is a matter of interest when one want to know if the data is equally distributed at the left and the right of some value (the centre of symmetry). Testing circular symmetry is not only important per se, also as it is one of the most frequently encountered simplifying hypothesis. In the sense that the rejection of such hypothesis leads the exploration of models more complex than the classic symmetric ones. In particular this problem is of special interest in the circular data, that is, samples that can be represented as points on the circumference of a unit circle, as "symmetrical distributions on the circle are comparatively rare" (see Mardia, 1972).

The circular data appear for one dimensional data when the periodicity must be taken into account. Examples of circular data can be found when the goal is to model orientations or a periodic phenomenon with known period. These examples are frequent in many applied fields, some of them related with orientations: the orientation of the red wood ants in reaction to different stimuli (Jander, 1957); the flight orientation in pigeons returning home (Schmidt-Koenig, 1963); the wind direction at Gorleston (England) during 1968 (Fisher, 1995, Ch. 5); or waves directions in the Adriatic sea (Jona-Lasinio et al., 2012; Lagona et al., 2015). Other examples are related with periodic phenomena: the number of soldiers deaths (along each year) during the Crimean War (originally introduced by Nightingale in 1858 and revisited in Brasseur, 2005); or the frosting and defrosting times (during the day) in glaciers Monte Alvear and Vinciguerra in Ushuaia, Argentina (Oliveira et al., 2013).

As in the real data, most of the classical models in the circular setting are symmetric. Some examples include the von Mises, cardioid, wrapped Cauchy and wrapped normal, which densities are given, respectively, by: $f_{\mathrm{VM}_\kappa}(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos\theta)$, for $\kappa > 0$, where I_k denotes the modified Bessel function of the first kind and order k; $f_{\mathrm{C}_\rho}(\theta) = \frac{1}{2\pi}(1 + 2\rho\cos\theta)$, for $0 \leq \rho < 1/2$; $f_{\mathrm{WC}_\rho}(\theta) = \frac{1}{2\pi}\left(\frac{1-\rho^2}{1+\rho^2-2\rho\cos\theta}\right)$, for $0 \leq \rho < 1$; $f_{\mathrm{WN}_\rho}(\theta) = \frac{1}{2\pi}(1 + 2\sum_{p=1}^{\infty} \rho^{p^2} \cos(p\theta))$, for $0 \leq \rho < 1$. In these densities, κ and ρ are concentration parameters, with ρ denoting the mean resultant length.

A location parameter $\mu \in [-\pi, \pi)$ can readily be introduced to change the centre of symmetry, leading to densities of the form $f_0(\theta - \mu)$ with modal direction μ.

All these classic parametric models are (reflective) symmetric, for that reason, it is important to determine when a more complex model with more parameters than its symmetric counterpart is needed. During the last years several asymmetric models appear in the circular literature, where, due its flexibility, one of the most popular ones is the *k-sine-skewed* distribution, which density is

$$f^k_{\mu,\lambda}(\theta) = f_0(\theta - \mu)[1 + \lambda \sin(k(\theta - \mu))], \tag{1}$$

where λ is the skewness parameter in the sense that for $\lambda = 0$ the density (1) is (reflectively) symmetric and when $k = 1$, in general, the density is unimodal and it is skewed to the left if $\lambda > 0$ and to the right if $\lambda < 0$. For $k \geq 2$ and $\lambda \neq 0$, the density (1) is generally multimodal, λ and k determine the number of modes as well as their heights and skewness (see Abe and Pewsey, 2011, for more information about the effect of these parameters). Regarding their applications, the k-sine-skewed distributions have been used as models for ant orientation data and the times of thunder storms, in Abe and Pewsey (2011), the CO_2 daily cycle at a rural site, in Pérez et al. (2012), and forest disturbance regimes, in Abe et al. (2012).

With the objective of determining when an asymmetric model is needed, for testing (reflective) circular symmetry when the central direction is known, there are, in the statistical literature, two methods: the b2-star test of Pewsey (2004), based on the second sine moment about the fixed central direction; and the optimal test (against k-sine-skewed alternatives) of Ley and Verdebout (2014). When the central direction is unknown, just the b2-bar proposal of Pewsey (2002) can be properly used for testing symmetry (and there is no notion of optimality in this test). The objective of this work is to present the results achieved in Ameijeiras-Alonso et al. (2017), where a new optimal proposal (against the k-sine-skewed alternatives) for testing reflective symmetry when the central direction is unknown. A new test statistic and its asymptotic distribution under the null and under k-sine-skewed alternatives will be given. It will be also seen that the b2-bar test of Pewsey (2002), asymptotically, is a particular case of this new proposal. The presented proposal will be test for reflective symmetry about an unknown central against the k-sine-skewed alternatives (1), i.e., it will be tested $H_0 : \lambda = 0$ vs. $H_1 : \lambda \neq 0$ in a parametric (when f_0 is determined) and in a semi-parametric way (when nothing about f_0 is known). The (optimal) tests will be obtained following the Le Cam methodology (an introduction to the Le Cam Theory can be found in van der Vaart, 2002). For that reason, in this manuscript, the optimality will be understood in the *maximin* sense. A test ϕ^* is maximin in the class of test at level α, $\mathcal{C}_\alpha$, for the null hypothesis H_0 against the alternative H_1, if ϕ^* has level α and its power is such that $\inf_{P \in H_1} \mathbb{E}_P(\phi^*) \geq \sup_{\phi \in \mathcal{C}_\alpha} \inf_{P \in H_1} \mathbb{E}_P(\phi)$.

2. THE ULAN PROPERTY

Let $\Theta_1, \ldots, \Theta_n$ be i.i.d. circular observations with common density (1). Let

$$\begin{aligned} \mathcal{F} \quad := \quad \Big\{ & f_0 : f_0(\theta) > 0, f_0(\theta + 2\pi k) = f_0(\theta) \forall k \in \mathbb{Z}, f_0(-\theta) = f_0(\theta), \\ & f_0 \text{ unimodal in } \theta \in [-\pi, \pi) \text{ with mode at } 0, \int_{-\pi}^{\pi} f_0(\theta) d\theta = 1 \Big\}. \end{aligned}$$

Then, for any reflectively symmetric unimodal base density $f_0 \in \mathcal{F}$ and any $k \in \mathbb{N}_0$, denote the joint distribution of the n-tuple $\Theta_1, \ldots, \Theta_n$ by $\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}; f_0, k}$, where $\boldsymbol{\vartheta} := (\mu, \lambda) \in [-\pi, \pi) \times [-1, 1]$. Since $f^k_{\mu,\lambda} = f_0$ when $\lambda = 0$, and hence does not depend on k, we drop the index k and simply write $\mathrm{P}^{(n)}_{\boldsymbol{\vartheta}; f_0}$ at $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 := (\mu, 0)$. Any pair (f_0, k) induces the *parametric* model

$$\mathcal{P}^{(n)}_{f_0,k} := \left\{ \mathrm{P}^{(n)}_{\boldsymbol{\vartheta}; f_0, k} : \boldsymbol{\vartheta} \in [-\pi, \pi) \times [-1, 1] \right\},$$

whereas any $k \in \mathbb{N}_0$ induces the *semi-parametric* model $\mathcal{P}^{(n)}_k := \cup_{f_0 \in \mathcal{F}} \mathcal{P}^{(n)}_{f_0,k}$.

The ULAN property of the parametric model $\mathcal{P}^{(n)}_{f_0,k}$ in the vicinity of unimodal reflective symmetry, i.e. around $\lambda = 0$, can be established under the following mild regularity condition.

ASSUMPTION A: *The base density $f_0(\theta)$ is $\mathcal{C}^1$ almost everywhere over $[-\pi,\pi)$, or equivalently over $\mathbb{R}$ by periodicity, with derivative $\dot{f}_0$ almost everywhere.*

Most classical reflectively symmetric densities satisfy this requirement. Note that the continuously differentiable condition over a compact manifold, combined with the fact that $f_0 > 0$, implies that the Fisher information quantity for location, $I_{f_0} := \int_{-\pi}^{\pi} \varphi_{f_0}^2(\theta) f_0(\theta) d\theta$, where $\varphi_{f_0} = -\dot{f}_0/f_0$, is finite. The ULAN property of the parametric model $\mathcal{P}_{f_0,k}^{(n)}$ with respect to $\boldsymbol{\vartheta} = (\mu,\lambda)'$, in the vicinity of unimodal reflective symmetry, then takes the following form.

Theorem 1. *Suppose $f_0 \in \mathcal{F}$, $k \in \mathbb{N}_0$ and that Assumption A holds. Then, for any $\mu \in [-\pi,\pi)$, the parametric family of densities $\mathcal{P}_{f_0,k}^{(n)}$ is ULAN at $\boldsymbol{\vartheta}_0 = (\mu,0)'$ with central sequence*

$$\begin{aligned} \boldsymbol{\Delta}_{f_0,k}^{(n)}(\mu) &:= \begin{pmatrix} \Delta_{f_0,k;1}^{(n)}(\mu) \\ \Delta_{k;2}^{(n)}(\mu) \end{pmatrix} \\ &:= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \varphi_{f_0}(\Theta_i - \mu) \\ \sin(k(\Theta_i - \mu)) \end{pmatrix}, \end{aligned}$$

and corresponding Fisher information matrix

$$\boldsymbol{\Gamma}_{f_0,k} := \begin{pmatrix} \Gamma_{f_0,k;11} & \Gamma_{f_0,k;12} \\ \Gamma_{f_0,k;12} & \Gamma_{f_0,k;22} \end{pmatrix},$$

where $\Gamma_{f_0,k;11} := I_{f_0}$, $\Gamma_{f_0,k;12} := -\int_{-\pi}^{\pi} \sin(k\theta)\dot{f}_0(\theta)d\theta$ and $\Gamma_{f_0,k;22} := \int_{-\pi}^{\pi} \sin^2(k\theta) f_0(\theta) d\theta$.

More precisely, for any $\mu^{(n)} = \mu + O(n^{-1/2})$ and for any bounded sequence $\boldsymbol{\tau}^{(n)} = (\tau_1^{(n)}, \tau_2^{(n)})' \in \mathbb{R}^2$ such that $\mu^{(n)} + n^{-1/2}\tau_1^{(n)}$ remains in $[-\pi,\pi)$ and $n^{-1/2}\tau_2^{(n)}$ in $[-1,1]$, we have

$$\begin{aligned} \Lambda^{(n)}_{(\mu^{(n)}+n^{-1/2}\tau_1^{(n)}, n^{-1/2}\tau_2^{(n)})'/(\mu^{(n)},0)';f_0,k} &:= \log\left(d\mathrm{P}^{(n)}_{(\mu^{(n)}+n^{-1/2}\tau_1^{(n)}, n^{-1/2}\tau_2^{(n)})';f_0,k} / d\mathrm{P}^{(n)}_{(\mu^{(n)},0)';f_0} \right) \\ &= \boldsymbol{\tau}^{(n)\prime} \boldsymbol{\Delta}_{f_0,k}^{(n)}(\mu^{(n)}) - (1/2)\boldsymbol{\tau}^{(n)\prime} \boldsymbol{\Gamma}_{f_0,k} \boldsymbol{\tau}^{(n)} + o_{\mathrm{P}}(1) \end{aligned} \tag{2}$$

and $\boldsymbol{\Delta}_{f_0,k}^{(n)}(\mu^{(n)}) \xrightarrow{\mathcal{D}} \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Gamma}_{f_0,k})$, both under $\mathrm{P}^{(n)}_{(\mu^{(n)},0);f_0}$ as $n \to \infty$.

The proof of Theorem 1 is given in Ley and Verdebout (2014), where a brief discussion of the minimal conditions required for the ULAN property to hold is also provided.

Remark 1. For the ULAN property to hold, the Fisher information matrix $\boldsymbol{\Gamma}_{f_0,k}$ must be non-singular. Proposition 1 of Ley and Verdebout (2014) states that this is always the case, except for when f_0 is von Mises and $k = 1$. As we shall see in the sequel, a singular information matrix is of no relevance when building tests for reflective symmetry about a *known* central direction but precludes the construction of a powerful test for reflective symmetry against von-Mises-based sine-skewed alternatives when the central direction is *unknown*.

3. OPTIMAL TESTS FOR REFLECTIVE SYMMETRY

Ley and Verdebout (2014) proposed locally and asymptotically optimal tests, in the Le Cam sense, for reflective symmetry within the k-sine-skewed family when μ is *known*. In this section, we first consider the *parametric* testing problem

$$\begin{cases} \mathcal{H}_{0;f_0} := \cup_{\mu\in[-\pi,\pi)} \mathrm{P}^{(n)}_{(\mu,0);f_0}, \\ \mathcal{H}_{1;f_0,k} := \cup_{\lambda\neq 0\in[-1,1]} \cup_{\mu\in[-\pi,\pi)} \mathrm{P}^{(n)}_{(\mu,\lambda);f_0,k}, \end{cases} \tag{3}$$

where f_0 is a specified density belonging to $\mathcal{F}$ and the unknown central direction under $\mathcal{H}_{0;f_0}$ is estimated.

A drawback of the above tests is that they are only valid under the parametric null hypothesis $\mathcal{H}_{0;f_0}$ with f_0 specified. In order to address the more general null hypothesis of reflective symmetry,

we need a test statistic whose asymptotic distribution is valid under any symmetric density $g_0 \in \mathcal{F}$. Thus, we subsequently consider the more demanding testing problem

$$\begin{cases} \mathcal{H}_0 := \cup_{\mu\in[-\pi,\pi)} \cup_{g_0\in\mathcal{F}} \mathrm{P}^{(n)}_{(\mu,0);g_0}, \\ \mathcal{H}_{1;k} := \cup_{\lambda\neq 0\in[-1,1]} \cup_{\mu\in[-\pi,\pi)} \cup_{g_0\in\mathcal{F}} \mathrm{P}^{(n)}_{(\mu,\lambda);g_0,k}, \end{cases} \tag{4}$$

in which the location parameter μ and the density g_0 both take on nuisance roles.

For both problems, we make use of the ULAN property of Theorem 1 to derive tests that (i) are valid under the null hypotheses considered and (ii) achieve local and asymptotic parametric optimality against a k-sine-skewed alternative characterized by the fixed couple $(f_0, k) \in (\mathcal{F} \times \mathbb{N}_0)$. In the semi-parametric testing problem (4), f_0 and k are chosen *a priori* by the practitioner and we derive tests $\phi_k^{(n);f_0}$ that are asymptotically optimal against the (f_0, k)-sine-skewed alternative and are such that

$$\lim \mathrm{E}[\phi_k^{(n);f_0}] \leq \alpha,$$

where the expectation is taken under any possible $\mathrm{P}^{(n)}_{(\mu,0);g_0}$ belonging to $\mathcal{H}_0$: i.e., they are valid under *any* density $g_0 \in \mathcal{F}$.

For the testing problem (3), our tests are constructed using a root-n consistent and discretized (see Assumption B below) estimator $\hat{\mu}^{(n)}$. The main reason why this testing problem is more demanding than the fixed-μ problem considered in Ley and Verdebout (2014) is because the Fisher information matrix $\mathbf{\Gamma}_{f_0,k}$ is not, in general, diagonal. If the information matrix $\mathbf{\Gamma}_{f_0,k}$ were diagonal, the substitution of $\hat{\mu}^{(n)}$ for μ would, asymptotically, have no influence on the behavior of the central sequence for departures from unimodal reflective symmetry $\Delta^{(n)}_{k;2}(\mu)$.

Remark 2. The information matrix $\mathbf{\Gamma}_{f_0,k}$ is never diagonal if $k = 1$. This can be seen by noting that $\sin(\theta)\varphi_{f_0}(\theta)f_0(\theta) > 0$ over $(-\pi, \pi)$. On the other hand, when $k > 1$ we can find densities for which $\Gamma_{f_0,k;12} = 0, \forall k \in \{2, 3, \ldots\}$. If the density function is square integrable on $[-\pi, \pi)$ this happens when $\alpha_k = \mathrm{E}[\cos k\Theta] = 0$ for $\Theta \sim f_0$, which can be proved using the Fourier expansion (see Jammalamadaka and Sengupta, 2001, Section 2.1) of density (1). A well-known example where this occurs is the cardioid density, for which $\alpha_1 = \rho$ and $\alpha_k = 0$ for $k > 1$.

From Remark 2, the covariance $\Gamma_{f_0,k;12}$ is only rarely null. Hence, a local perturbation of μ has the same asymptotic impact on $\Delta^{(n)}_{k;2}(\mu)$ as a local perturbation of $\lambda = 0$. It follows that the cost of not knowing the value of μ is strictly positive when performing inference on λ: the stronger the correlation between μ and λ, the larger that cost. The worst case occurs when the information matrix is singular (see Remark 1), which leads to asymptotic local powers equal to the nominal level α. For this scenario, the best possible test is that which ignores the data and simply rejects the null hypothesis with probability α. Henceforth we refer to such a test as the "trivial test".

We address the cost of estimating μ by removing the effect of the location central sequence $\Delta^{(n)}_{f_0,k;1}(\mu)$ from the skewness central sequence $\Delta^{(n)}_{k;2}(\mu)$. To achieve this we use a Gram-Schmidt orthogonalization approach. We project $\Delta^{(n)}_{k;2}(\mu)$ onto the subspace orthogonal to $\Delta^{(n)}_{f_0,k;1}(\mu)$, which ensures that the resulting *f_0-efficient central sequence for skewness* $\Delta^{(n)eff}_{f_0,k;2}(\mu)$ and $\Delta^{(n)}_{f_0,k;1}(\mu)$ are asymptotically uncorrelated. This new central sequence is of the form

$$\begin{aligned} \Delta^{(n)eff}_{f_0,k;2}(\mu) &:= \Delta^{(n)}_{k;2}(\mu) - \frac{\Gamma_{f_0,k;12}}{\Gamma_{f_0,k;11}}\Delta^{(n)}_{f_0,k;1}(\mu) \\ &= n^{-1/2}\sum_{i=1}^{n}\left(\sin(k(\Theta_i-\mu)) - \frac{\Gamma_{f_0,k;12}}{\Gamma_{f_0,k;11}}\varphi_{f_0}(\Theta_i-\mu)\right). \end{aligned} \tag{5}$$

Now we make use of another important consequence of the ULAN property, namely the *asymptotic linearity property*:

$$\Delta^{(n)}_{f_0,k}(\mu + n^{-1/2}\tau_1^{(n)}) - \Delta^{(n)}_{f_0,k}(\mu) = -\Gamma_{f_0,k}(\tau_1^{(n)}, 0) + o_{\mathrm{P}}(1) \tag{6}$$

under $\mathrm{P}^{(n)}_{(\mu,0);f_0}$ as $n \to \infty$, with $\tau_1^{(n)} \in \mathbb{R}$ as in Theorem 1. We refer the reader to Sections 2 and 3 of Koudou and Ley (2014) for in-depth discussions of these issues. It is not difficult to derive the

asymptotic linearity property of $\Delta^{(n)eff}_{f_0,k;2}(\mu)$ from (6), namely:

$$\Delta^{(n)eff}_{f_0,k;2}(\mu + n^{-1/2}\tau^{(n)}_1) - \Delta^{(n)eff}_{f_0,k;2}(\mu) = o_{\mathrm{P}}(1) \tag{7}$$

under $\mathrm{P}^{(n)}_{(\mu,0);f_0}$ as $n \to \infty$.

Now consider replacing the non-random bounded sequence $\tau^{(n)}_1$ with $n^{1/2}(\hat{\mu}^{(n)} - \mu)$ for some root-n consistent estimator $\hat{\mu}^{(n)}$. The latter is bounded in probability and, via Lemma 4.4 of Kreiss (1987), serves as an ideal candidate for $\tau^{(n)}_1$, provided the following assumption holds.

ASSUMPTION B: *The sequence of estimators $\hat{\mu}^{(n)}$ is (i) root-n consistent, i.e. $n^{1/2}(\hat{\mu}^{(n)} - \mu) = O_{\mathrm{P}}(1)$ as $n \to \infty$, under $\mathrm{P}^{(n)}_{(\mu,0);f_0}$, and (ii) locally asymptotically discrete, meaning that, for all $\mu \in [-\pi, \pi)$ and all $c > 0$, there exists an $M = M(c) > 0$ such that the number of possible values of $\hat{\mu}^{(n)}$ in intervals of the form $\{t \in \mathbb{R} : n^{1/2}|t - \mu| \leq c\}$ is bounded by M, uniformly as $n \to \infty$.*

Note that Assumption B(ii) is a purely technical requirement, with little practical implication. Indeed, for fixed sample size, any estimator can be considered part of a locally asymptotically discrete sequence. However, it is this assumption that enables us to replace $\tau^{(n)}_1$ by $n^{1/2}(\hat{\mu}^{(n)} - \mu)$ in (7) thanks to the aforementioned Lemma 4.4 of Kreiss (1987), yielding

$$\Delta^{(n)eff}_{f_0,k;2}(\hat{\mu}^{(n)}) - \Delta^{(n)eff}_{f_0,k;2}(\mu) = o_{\mathrm{P}}(1) \tag{8}$$

under $\mathrm{P}^{(n)}_{(\mu,0);f_0}$ as $n \to \infty$.

Our locally and asymptotically maximin f_0-parametric test, $\phi^{(n);f_0}_k$, rejects $\mathcal{H}_{0;f_0}$ at asymptotic level α whenever the statistic

$$Q^{(n);f_0}_k := \frac{|\Delta^{(n)eff}_{f_0,k;2}(\hat{\mu}^{(n)})|}{\Gamma^{1/2}_{f_0,k;22.1}}$$

exceeds the upper $\alpha/2$ quantile of the standard normal distribution, $z_{1-\alpha/2}$, where $\Gamma_{f_0,k;22.1} := \Gamma_{f_0,k;22} - \frac{\Gamma^2_{f_0,k;12}}{\Gamma_{f_0,k;11}}$ is the asymptotic variance of $\Delta^{(n)eff}_{f_0,k;2}(\mu)$ under $\mathrm{P}^{(n)}_{(\mu,0);f_0}$. Optimal properties of this test statistic are described in the semi-parametric scenario.

Consider now the testing problem in (4). Our objective is still to construct a test that is locally and asymptotically maximin for detecting an alternative characterized by a specified couple $(f_0, k) \in (\mathcal{F} \times \mathbb{N}_0)$. The main difference between the semi-parametric scenario addressed here and the parametric one is that here we aim to build a test that is asymptotically valid under $\mathcal{H}_0 := \cup_{\mu\in[-\pi,\pi)} \cup_{g_0\in\mathcal{F}} \mathrm{P}^{(n)}_{(\mu,0);g_0}$. Specifically, we need to allow for the substitution of μ by $\hat{\mu}^{(n)}$ in $\Delta^{(n)}_{k;2}(\mu)$ under $\mathrm{P}^{(n)}_{(\mu,0);g_0}$ and the distinct possibility that $g_0 \neq f_0$. The ULAN property combined with Lemma 4.4 of Kreiss (1987) leads to

$$\Delta^{(n)}_{k;2}\left(\hat{\mu}^{(n)}\right) - \Delta^{(n)}_{k;2}(\mu) = -\Gamma_{g_0,k;12}\sqrt{n}\left(\hat{\mu}^{(n)} - \mu\right) + o_{\mathrm{P}}(1), \tag{9}$$

under $\mathrm{P}^{(n)}_{(\mu,0);g_0}$ as , provided that $\hat{\mu}^{(n)}$ satisfies Assumption B. Then, the substitution of μ by $\hat{\mu}^{(n)}$ under $\mathrm{P}^{(n)}_{(\mu,0)';g_0}$ has no asymptotic cost only if $\Gamma_{g_0,k;12} = 0$. As we saw in Remark 2, this will rarely be the case. In order to circumvent this problem and eliminate the asymptotic covariance $\Gamma_{g_0,k;12}$ while keeping the (f_0, k) target in mind, we consider an *efficient central sequence*

$$\Delta^{(n);ecd}_{f_0,g_0,k;2}(\mu) := n^{-1/2}\sum_{i=1}^{n}\left(\sin(k(\Theta_i - \mu)) - \eta\varphi_{f_0}(\Theta_i - \mu)\right),$$

where $\eta := \Gamma_{g_0,k;12}/\Gamma_{f_0,g_0,k;11}$ with

$$\Gamma_{f_0,g_0,k;11} := \int_{-\pi}^{\pi} \varphi_{f_0}(\theta)\varphi_{g_0}(\theta)g_0(\theta)d\theta.$$

Since f_0 and g_0 are both periodic $\mathcal{C}^1$ functions over a bounded set and $f_0, g_0 > 0$, the cross-information quantity $\Gamma_{f_0,g_0,k;11}$ is finite. When $g_0 = f_0$, $\Gamma_{f_0,f_0,k;11} = \Gamma_{f_0,k;11}$, so that $\Delta^{(n);ecd}_{f_0,f_0,k;2}(\mu)$ will coincide with $\Delta^{(n)eff}_{f_0,k;2}(\mu)$ under $\mathrm{P}^{(n)}_{(\mu,0)',f_0}$, which, as we will see in the sequel, is key to maintaining the asymptotic optimality against (f_0, k) alternatives. Integrating by parts, we obtain

$$\begin{aligned}\Gamma_{g_0,k;12} &= \int_{-\pi}^{\pi} \sin(k\theta)\varphi_{g_0}(\theta)g_0(\theta)d\theta \\ &= k\int_{-\pi}^{\pi} \cos(k\theta)g_0(\theta)d\theta \\ &= k\mathrm{E}_{g_0}[\cos(k(\Theta_i - \mu))] \end{aligned} \tag{10}$$

and

$$\begin{aligned}\Gamma_{f_0,g_0,k;11} &= \int_{-\pi}^{\pi} \varphi_{f_0}(\theta)\varphi_{g_0}(\theta)g_0(\theta)d\theta \\ &= [-\varphi_{f_0}(\theta)g_0(\theta)]_{-\pi}^{\pi} + \int_{-\pi}^{\pi} \dot{\varphi}_{f_0}(\theta)g_0(\theta)d\theta \\ &= \int_{-\pi}^{\pi} \dot{\varphi}_{f_0}(\theta)g_0(\theta)d\theta = \mathrm{E}_{g_0}[\dot{\varphi}_{f_0}(\Theta_i - \mu)], \end{aligned} \tag{11}$$

provided that the following assumption holds.

Assumption C. The mapping $\theta \mapsto \varphi_{f_0}(\theta)$ is $\mathcal{C}^1$ almost everywhere over $[-\pi, \pi)$ with derivative $\dot{\varphi}_{f_0}(\theta)$ almost everywhere, where $f_0 \in \mathcal{F}$.

In the following lemma, we establish that

$$\hat{\Gamma}_{g_0,k;12} = n^{-1}\sum_{i=1}^{n} k\cos\left(k\left(\Theta_i - \hat{\mu}^{(n)}\right)\right)$$

and

$$\hat{\Gamma}_{f_0,g_0,k;11} = n^{-1}\sum_{i=1}^{n} \dot{\varphi}_{f_0}\left(\Theta_i - \hat{\mu}^{(n)}\right)$$

are consistent estimators of $\Gamma_{g_0,k;12}$ and $\Gamma_{f_0,g_0,k;11}$ in (10) and (11), respectively.

Lemma 1. *Suppose $k \in \mathbb{N}_0$, $f_0, g_0 \in \mathcal{F}$ and Assumptions A, B and C hold. Then $\hat{\Gamma}_{g_0,k;12} - \Gamma_{g_0,k;12} = o_{\mathrm{P}}(1)$ and $\hat{\Gamma}_{f_0,g_0,k;11} - \Gamma_{f_0,g_0,k;11} = o_{\mathrm{P}}(1)$ as $n \to \infty$ under $\mathrm{P}^{(n)}_{(\mu,0);g_0}$.*

The proof is provided in Ameijeiras-Alonso et al. (2017).

Using these estimators, our test is based on the estimated version of the efficient central sequence

$$\Delta^{*(n);ecd}_{f_0,k;2}(\mu) := n^{-1/2}\sum_{i=1}^{n}\left(\sin(k(\Theta_i - \mu)) - \hat{\eta}\varphi_{f_0}(\Theta_i - \mu)\right), \tag{12}$$

where $\hat{\eta} := \hat{\Gamma}_{g_0,k;12}/\hat{\Gamma}_{f_0,g_0,k;11}$. The test $\phi^{*(n)}_{f_0,k}$ rejects $\mathcal{H}_0$ at asymptotic level α whenever the test statistic $|Q^{*(n)}_{f_0,k}| > z_{1-\alpha/2}$, where

$$\begin{aligned} Q^{*(n)}_{f_0,k} &:= \frac{\Delta^{*(n);ecd}_{f_0,k;2}\left(\hat{\mu}^{(n)}\right)}{\left(C^{*(n)}_{f_0,k}\left(\hat{\mu}^{(n)}\right)\right)^{1/2}} \\ &:= \frac{n^{-1/2}\sum_{i=1}^{n}\left(\sin(k(\Theta_i - \hat{\mu}^{(n)})) - \hat{\eta}\varphi_{f_0}\left(\Theta_i - \hat{\mu}^{(n)}\right)\right)}{\left(n^{-1}\sum_{i=1}^{n}\left(\sin(k(\Theta_i - \hat{\mu}^{(n)})) - \hat{\eta}\varphi_{f_0}\left(\Theta_i - \hat{\mu}^{(n)}\right)\right)^2\right)^{1/2}}. \end{aligned} \tag{13}$$

The asymptotic distribution of $Q^{*(n)}_{f_0,k}$ is formally established in Theorem 2, where we also prove the optimality properties of $\phi^{*(n)}_{f_0,k}$. Before doing so, however, we first need the following result on the efficient central sequence in (12), whose proof is given in Ameijeiras-Alonso et al. (2017).

Lemma 2. *Suppose $k \in \mathbb{N}_0$, $f_0, g_0 \in \mathcal{F}$ and Assumptions A, B and C hold. Then, as $n \to \infty$ under $\mathrm{P}^{(n)}_{(\mu,0);g_0}$: (i) $\Delta^{*(n);ecd}_{f_0,k;2}\left(\hat{\mu}^{(n)}\right) - \Delta^{(n);ecd}_{f_0,g_0,k;2}(\mu) = o_{\mathrm{P}}(1)$; (ii) $C^{*(n)}_{f_0,k}\left(\hat{\mu}^{(n)}\right) - C^{(n)}_{f_0,g_0,k}(\mu) = o_{\mathrm{P}}(1)$, where*

$$C^{(n)}_{f_0,g_0,k}(\mu) := n^{-1}\sum_{i=1}^{n}\left(\sin(k(\Theta_i-\mu)) - \frac{\Gamma_{g_0,k;12}}{\Gamma_{f_0,g_0,k;11}}\varphi_{f_0}(\Theta_i-\mu)\right)^2.$$

Using Lemma 2, we can establish the optimality properties of the semi-parametric test $\phi^{*(n)}_{f_0,k}$. Given a posited base density $f_0 \in \mathcal{F}$ and value of k, in Theorem 2 we provide the asymptotic properties of the test statistic $Q^{*(n)}_{f_0,k}$ both under $\mathcal{H}_0$ and a sequence of contiguous alternatives. Theorem 2 is the main result of the paper and resolves the complicated issue of the non-null behavior of semi-parametrically efficient test procedures for circular reflective symmetry about an unknown central direction against k-sine-skewed alternatives.

Theorem 2. *Suppose $k \in \mathbb{N}_0$, the posited base density $f_0 \in \mathcal{F}$ and Assumptions (A), (B) and (C) hold. Then:*

(i) under $\mathcal{H}_0$, $Q^{*(n)}_{f_0,k} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$ as $n \to \infty$, so that the test $\phi^{*(n)}_{f_0,k}$ has asymptotic level α under $\mathcal{H}_0$;

(ii) under $\cup_{\mu\in[-\pi,\pi)}\mathrm{P}^{(n)}_{(\mu,n^{-1/2}\tau_2^{(n)})';g_0,k'}$ with $g_0 \in \mathcal{F}$, $k' \in \mathbb{N}_0$ and $\tau_2^{(n)}$ a bounded sequence as in Theorem 1, $Q^{*(n)}_{f_0,k}$ is asymptotically normal with mean $V^{g_0}_{f_0}(k)^{-1/2}C^{g_0}_{f_0}(k,k')\tau_2$ and variance 1, where $\tau_2 = \lim_{n\to\infty}\tau_2^{(n)}$,

$$V^{g_0}_{f_0}(k) = \int_{-\pi}^{\pi}\left(\sin(k\theta) - \frac{\Gamma_{g_0,k;12}}{\Gamma_{f_0,g_0,k;11}}\varphi_{f_0}(\theta)\right)^2 g_0(\theta)d\theta,$$

and

$$C^{g_0}_{f_0}(k,k') = \int_{-\pi}^{\pi}\left(\sin(k\theta) - \frac{\Gamma_{g_0,k;12}}{\Gamma_{f_0,g_0,k;11}}\varphi_{f_0}(\theta)\right)\sin(k'\theta)g_0(\theta)d\theta,$$

(both finite) ;

(iii) the test $\phi^{*(n)}_{f_0,k}$ is locally and asymptotically maximin at asymptotic level α when testing $\mathcal{H}_0$ against $\mathcal{H}_{1;f_0,k}$.

The proof is given in Ameijeiras-Alonso et al. (2017).

Theorem 2 states that $\phi^{*(n)}_{f_0,k}$ is valid under the entire null hypothesis $\mathcal{H}_0$, and so is asymptotically distribution and location free. Theorem 2(ii) provides an important result which can be used to calculate the asymptotic power of $\phi^{*(n)}_{f_0,k}$ against local alternatives of the form $\cup_{\mu\in[-\pi,\pi)}\mathrm{P}^{(n)}_{(\mu,n^{-1/2}\tau_2^{(n)});g_0,k'}$ as a function of the posited density f_0.

4. CONCLUSIONS

In this work, a new proposal for testing reflective symmetry when the central direction is unknown is developed following the Le Cam methodology. The asymptotic normal distribution is obtained under the null (centred in zero) and, also, under the k-sine-skewed alternatives (centred in a point different from zero). Showing, also, that this test is optimal (in the maximin sense) against these alternatives. Its behaviour in small- to medium-sized samples is analyzed and compared with previous purposes in Ameijeiras-Alonso et al. (2017) and its application to two different real datasets is also provied on there. This work extend previous results obtained in Ley and Verdebout (2014), where the focus was testing the symmetry hypothesis when the central direction is known.

ACKNOWLEDGEMENTS

The research work underpinning this talk was supported by grants MTM2016-76969-P (Spanish State Research Agency, AEI) and MTM2013-41383-P (Spanish Ministry of Economy, Industry and Competitiveness), both co-funded by the European Regional Development Fund (ERDF), grant

GR15013 (Junta de Extremadura and the European Union), and funding from the IAP network P7/06 StUDyS of the Belgian Government (Belgian Science Policy) and the National Bank of Belgium. Part of the research was carried out by the first author during a visit to Ghent University supported by grants BES-2014-071006 and EEBB-I-16-11503 (Spanish Ministry of Economy and Competitiveness).

BIBLIOGRAPHY

Abe, T., Kubota, Y., Shimatani, K., Aakala, T., and Kuuluvainen, T. (2012). Circular distributions of fallen logs as an indicator of forest disturbance regimes. *Ecological Indicators*, **18**, 559–566.

Abe, T., and Pewsey, A. (2011). Sine–skewed circular distributions. *Statistical Papers*, **52**, 683–707.

Ameijeiras-Alonso, J., Ley, C., Pewsey, A., and Verdebout, T. (2017). Simple optimal tests for circular reflective symmetry about an unknown central direction. *arXiv preprint arXiv:1707.09272*. Submitted.

Brasseur, L. (2005). Florence nightingale's visual rhetoric in the rose diagrams. *Technical Communication Quarterly*, **14**, 161–182.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge University Press, Cambridge, U.K.

Jammalamadaka, S. R., and Sengupta, A. (2001). *Topics in Circular Statistics*. World Scientific.

Jander, R. (1957). Die optische richtungsorientierung der roten waldameise (formica ruea l.). *Zeitschrift für vergleichende Physiologie*, **40**, 162–238.

Jona-Lasinio, G., Gelfand, A., and Jona-Lasinio, M. (2012). Spatial analysis of wave direction data using wrapped gaussian processes. *Annals of Applied Statistics*, (pp. 1478–1498).

Koudou, A. E., and Ley, C. (2014). Efficiency combined with simplicity: new testing procedures for generalized inverse gaussian models. *Test*, **23**, 708–724.

Kreiss, J. (1987). On adaptive estimation in stationary ARMA processes. *Annals of Statistics*, **15**, 112–133.

Lagona, F., Picone, M., Maruotti, A., and Cosoli, S. (2015). A hidden markov approach to the analysis of space–time environmental data with linear and circular components. *Stochastic Environmental Research and Risk Assessment*, **29**, 397–409.

Ley, C., and Verdebout, T. (2014). Simple optimal tests for circular reflective symmetry about a specified median direction. *Statistica Sinica*, **24**, 1319–1339.

Mardia, K. V. (1972). *Statistics of Directional Data*. Academic Press, London.

Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2013). Nonparametric circular methods for exploring environmental data. *Environmental and Ecological Statistics*, **20**, 1–17.

Pérez, I. A., Sánchez, M. L., García, M. A., and Pardo, N. (2012). Analysis of CO_2 daily cycle in the low atmosphere at a rural site. *Science of the Total Environment*, **431**, 286–292.

Pewsey, A. (2002). Testing circular symmetry. *Canadian Journal of Statistics*, **30**, 591–600.

Pewsey, A. (2004). Testing for circular reflective symmetry about a known median axis. *Journal of Applied Statistics*, **31**, 575–585.

Schmidt-Koenig, K. (1963). On the role of the loft, the distance and site of release in pigeon homing (the "cross-loft experiment"). *The Biological Bulletin*, **125**, 154–164.

van der Vaart, A. (2002). The statistical work of Lucien Le Cam. *Annals of Statistics*, **30**, 631–682.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

O bootstrap para a estimación non paramétrica da función de razón de fallo

Inés Barbeito[1] e Ricardo Cao[1]

[1] Grupo de Investigación MODES, Departamento de Matemáticas, Centro de Investigación CITIC, Universidade da Coruña.

RESUMO

No contexto da estimación non paramétrica da función de razón de fallo, é de vital importancia a correcta elección do parámetro de suavizado. Neste sentido, propóñense dous novos selectores de ventá bootstrap para a estimación non paramétrica da función de razón de fallo considerando que os datos son iid, establecendo unha expresión pechada para a versión bootstrap suavizada do erro cadrático medio integrado (MISE) dunha aproximación do estimador non paramétrico da función de razón de fallo. Deste xeito, pódese prescindir da aproximación de Monte Carlo para obter remostras bootstrap. Para comprobar o bo funcionamento na práctica destes selectores, realízase un estudo de simulación, comparando os selectores propostos cos xa existentes na literatura. Ilústrase tamén o comportamento empírico dos mesmos cunha aplicación a datos de diabetes.

Palabras e frases chave: selector de ventá, erro cadrático medio integrado, razón de fallo, bootstrap suavizado

1. INTRODUCIÓN

A estimación non paramétrica da función de razón de fallo é un tema que se leva estudando ao longo dos últimos cincuenta anos dende a súa introdución por Watson e Leadbetter (1964a, 1964b). De feito, esta función é moi útil en análise de supervivencia, polo que no caso de datos censurados publicáronse unha gran cantidade de artigos ata o momento (ver Tanner e Wong (1983, 1984), Lo et al (1989) ou Patil (1993a, 1993b)). No caso de datos sen censura, que será no que nos centraremos nesta comunicación, podemos citar algunhas referencias existentes na literatura actual como Murthy (1965), Rice e Rosenblatt (1976), Singpurwalla e Wong (1980, 1983), Rao (1983) ou Sarda e Vieu (1991).

Consideremos $(X_1, X_2, \ldots, X_n)$ unha mostra aleatoria simple que provén dunha poboación con densidade f e función de distribución F. O noso obxectivo será estimar a función de razón de fallo, r, que pode expresarse como segue:

$$r(x) = \lim_{\delta \to 0^+} \frac{\mathbb{P}\left(X \leq x + \delta \,|X > x\right)}{\delta} = \frac{f(x)}{1 - F(x)}.$$

Se ademais non supoñemos ningunha hipótese paramétrica sobre a distribución subxacente, un xeito natural de proceder sería estimar a función r por medio da estimación tipo núcleo:

$$\hat{r}_h^s(x) = \frac{\hat{f}_h(x)}{1 - \hat{F}_h(x)}, \tag{1}$$

onde $\hat{f}_h(x) = \dfrac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$ é o estimador non paramétrico tipo núcleo da densidade e $\hat{F}_h(x) = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{K}\left(\dfrac{x - X_i}{h}\right)$ é o estimador non paramétrico tipo núcleo da función de distribución, K é unha

función núcleo e $\mathbb{K}(t) = \int_{-\infty}^{t} K(u)du$ é a función de distribución acumulativa asociada a K.

Un dos principais problemas cos que nos atopamos é a correcta elección do parámetro de suavizado, h, que controla o grao de suavización do estimador. Polo tanto, unha mala escolla do mesmo pode provocar que o estimador da función de razón de fallo sexa infrasuavizado ou sobresuavizado. Na literatura existente é posible atopar numerosos estudos acerca da selección do parámetro de suavizado, tanto nun contexto de censura como de non censura: métodos de validación cruzada (ver Tanner e Wong (1984), Sarda e Vieu (1991), Patil (1993a, 1993b)) ou métodos bootstrap (ver González-Manteiga et al (1996), no caso de datos con censura).

Unha das aplicacións máis interesantes da estimación non paramétrica da razón de fallo, ademais de poder coñecer a probabilidade de que a un individuo lle ocorra un certo suceso ao longo do tempo, é poder extraer conclusións para unha determinada poboación a partir dunha mostra que sexa representativa da mesma. Isto é de especial relevancia xa que en numerosas ocasións non é posible obter os datos de toda a poboación de estudo. Cobran gran importancia neste ámbito as técnicas de remostraxe, entre as que destaca o bootstrap (ver Efron and Tibishirani (1993)). Con elas, a partir dunha mostra que sexa representativa da poboación, é posible xerar tantos datos como se queira cunha distribución similar á distribución da poboación orixinal. Centrarémonos neste tipo de técnicas nesta comunicación, cuxo obxectivo será establecer o parámetro de suavizado, h, para o cal se minimice un criterio de erro a saber, o erro cadrático medio integrado (ou MISE). Abordarase a obtención do selector h dende un punto de vista que difire do habitual, de xeito que se obterá unha expresión pechada para as versións bootstrap de dito criterio de erro (MISE*) considerando aproximacións do estimador non paramétrico da función de razón de fallo dado en (1), en vez de tratar de obter unha aproximación do mesmo por Monte Carlo. Desta forma, aforrarase tempo de execución, así como se obterá un selector de ventá máis preciso, sen erros da aproximación de Monte Carlo. Esta proposta é paralela á dada por Cao (1993) no contexto de estimación da densidade con datos independentes, ou ás dadas por Barbeito e Cao (2016, 2017) baixo dependencia.

Así, nas Seccións 2 e 3 establécense os dous novos selectores de ventá bootstrap para a función de razón de fallo atopando unha expresión pechada para a versión bootstrap do MISE. Na Sección 4 considérase un caso particular: qué ocorre cando o núcleo é Gaussiano, xa que as expresións obtidas nas seccións anteriores para o MISE* simplifícanse. Finalmente, na Sección 5, levarase a cabo un estudo de simulación e unha aplicación a datos reais para comprobar o bo comportamento empírico destes novos selectores, así como para comparar o seu comportamento na práctica cos seus competidores existentes na literatura actual. As conclusións derivadas da Sección 5 expoñeranse na Sección 6.

2. PRIMEIRO SELECTOR DE VENTÁ BOOTSTRAP

Como mencionamos anteriormente, o noso obxectivo será establecer unha expresión pechada para a versión bootstrap do erro cadrático medio integrado (MISE). Consideremos o seguinte estimador 'non observable' da función de razón de fallo:

$$\tilde{r}_{h,1}(x) = \frac{\hat{f}_h(x)}{1 - F(x)}. \tag{2}$$

Como podemos ver en (2), o denominador segue dependendo dalgún xeito da función de distribución teórica subxacente, F. Logo, trátase dunha versión teórica de (1), non dun estimador real.

Coa finalidade de probar os seguintes resultados, precisamos dalgunhas condicións:

(C1) A función de peso w toma valores non negativos, está limitada e existe algún intervalo compacto $[a, b]$ tal que $soporte(w) \subset [a, b]$.

(C2) K é unha función de densidade simétrica e limitada.

(C3) A densidade f está limitada.

(C4) $F(b) < 1$.

O seguinte resultado presenta unha expresión explícita para o MISE do estimador da función de razón de fallo $\tilde{r}_{h,1}$, dado en (2).

Teorema 1 *Se a mostra $(X_1, \ldots, X_n)$ é iid e se cumpren as Condicións (C1)-(C4), entón o MISE de $\tilde{r}_{h,1}$ (2) pode expresarse como:*

$$\begin{aligned} MISE_{w,\tilde{r}_{h,1}}(h) &= \int \left(\frac{K_h * f(x) - f(x)}{1 - F(x)} \right)^2 w(x)\, dx \\ &\quad + \frac{1}{nh} \int \frac{K_h^2 * f(x)}{(1 - F(x))^2} w(x)\, dx - \frac{1}{n} \int \left(\frac{K_h * f(x)}{1 - F(x)} \right)^2 w(x)\, dx, \end{aligned}$$

sendo $*$ o operador convolución.

Versión bootstrap do MISE

Consideremos $(X_1^*, \ldots, X_n^*)$ unha remostra bootstrap suavizada (ver Silverman e Young (1987) ou Cao (1993)), que depende dunha ventá piloto, g, e a versión bootstrap do estimador $\tilde{r}_{h,1}$, dado en (2):

$$\tilde{r}_{h,1}^*(x) = \frac{\hat{f}_h^*(x)}{1 - \hat{F}_g(x)}.$$

O seguinte resultado presenta unha expresión pechada para o $\mathrm{MISE}^*_{\tilde{r}_{h,1},w}(h)$ no caso iid, moi útil xa que con ela non precisamos da aproximación de Monte Carlo para obter as remostras bootstrap.

Teorema 2 *Baixo as condicións (C1)-(C4), a versión bootstrap suavizada do MISE para $\tilde{r}_{h,1}$ admite a seguinte expresión pechada:*

$$\begin{aligned} MISE^*_{\tilde{r}_{h,1},w}(h) &= \int \left[\frac{1}{n(1 - \hat{F}_g(x))} \sum_{i=1}^n (K_h * K_g(x - X_i) - K_g(x - X_i)) \right]^2 w(x) dx \\ &\quad + \frac{1}{nh} \int \left[\frac{1}{n(1 - \hat{F}_g(x))^2} \sum_{i=1}^n K_h^2 * K_g(x - X_i) \right] w(x)\, dx \\ &\quad - \frac{1}{n} \int \left[\frac{1}{n(1 - \hat{F}_g(x))} \sum_{i=1}^n K_h * K_g(x - X_i) \right]^2 w(x)\, dx. \end{aligned}$$

Minimizando en h esta función, atoparemos o selector de ventá que estamos buscando: $h_{BOOT1} = h_{\mathrm{MISE}^*_{\tilde{r}_{h,1},w}} = \arg\min_{h>0} \mathrm{MISE}^*_{\tilde{r}_{h,1},w}(h)$.

3. SEGUNDO SELECTOR DE VENTÁ BOOTSTRAP

De xeito similar á Sección 2, o noso propósito é propoñer un novo selector de ventá bootstrap para a estimación da función de razón de fallo. A principal diferencia co estimador dado en (2) é que usaremos unha versión aproximada da expresión $\hat{r}_h(x) - r(x)$, onde $\hat{r}_h$ é o estimador non paramétrico da función de risco dado en (1). Usaremos o feito de que $\hat{F}_h$ é o estimador tipo núcleo de F. O estimador aproximado, denominado $\tilde{r}_{h,2}$, vén dado por:

$$\begin{aligned} \tilde{r}_{h,2}(x) &= (\hat{r}_h(x) - r(x)) \frac{1 - \hat{F}_h(x)}{1 - F(x)} + r(x) \qquad (3) \\ &= \frac{1}{1 - F(x)} \hat{f}_h(x) + \frac{f(x)}{(1 - F(x))^2} \hat{F}_h(x) - \frac{f(x)}{(1 - F(x))^2} + r(x). \end{aligned}$$

A continuación, establécese unha expresión explícita para o MISE de $\tilde{r}_{h,2}$, definido en (3).

Teorema 3 *Se a mostra é iid e se asumen as condicións (C1)-(C4), entón o MISE de $\tilde{r}_{h,2}$, dado en (3), pódese expresar como segue:*

$$\begin{aligned}
MISE_{\tilde{r}_{h,2},w}(h) &= \int \frac{1}{(1-F(x))^2}\left[\frac{1}{n}K_h^2 * f(x) + \frac{n-1}{n}\left(K_h * f(x)\right)^2\right] w(x)\,dx \\
&+2\int \frac{f(x)}{(1-F(x))^3}\left[\frac{1}{n}\int K_h(x-y)\mathbb{K}\left(\frac{x-y}{h}\right) f(y)dy\right. \\
&\left.+\frac{n-1}{n}K_h * f(x)\int \mathbb{K}\left(\frac{x-y}{h}\right) f(y)dy\right] w(x)\,dx \\
&-2\int \frac{f(x)}{(1-F(x))^3}K_h * f(x)w(x)\,dx + \int \frac{f^2(x)}{(1-F(x))^4}w(x)\,dx \\
&+\int \frac{f^2(x)}{(1-F(x))^4}\left[\frac{1}{n}\int \mathbb{K}\left(\frac{x-y}{h}\right)^2 f(y)dy\right. \\
&\left.+\frac{n-1}{n}\left[\int \mathbb{K}\left(\frac{x-y}{h}\right) f(y)dy\right]^2\right] w(x)\,dx \\
&-2\int \frac{f^2(x)}{(1-F(x))^4}\int \mathbb{K}\left(\frac{x-y}{h}\right) f(y)dyw(x)\,dx.
\end{aligned}$$

Versión bootstrap do MISE

Consideremos unha ventá piloto, g, unha remostra suavizada bootstrap $(X_1^*,\ldots,X_n^*)$, e a versión bootstrap da expresión dada en (3), que resulta:

$$\tilde{r}_{h,2}^*(x) = \frac{1}{1-\hat{F}_g(x)}\hat{f}_h^*(x) + \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^2}\hat{F}_h^*(x) - \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^2}.$$

O noso obxectivo, de novo, é atopar unha expresión explícita para a versión bootstrap suavizada do MISE, de xeito que poidamos evitar o uso de Monte Carlo.

A continuación, introdúcense unhas definicións previas necesarias para o Teorema 4, que presenta unha expresión explícita para $\mathrm{MISE}^*_{\tilde{r}_{h,2},w}(h)$ no contexto de datos iid.

Definición 1 *Para calquera función integrable $\psi:\mathbb{R}\longrightarrow\mathbb{R}$, a función $\mathbb{I}_\psi$ defínese como $\mathbb{I}_\psi(t) := \int_{-\infty}^{t}\psi(u)du$. Se a función integrable $\psi:\mathbb{R}^2\longrightarrow\mathbb{R}$ depende de dúas variables, entón a función $\mathbb{I}_\psi$ vén definida por $\mathbb{I}_\psi(t_1,t_2) := \int_{-\infty}^{t_2}\int_{-\infty}^{t_1}\psi(u,v)\,du\,dv$.*

Definición 2 *Dadas dúas funcións $\psi,\varphi:\mathbb{R}\longrightarrow\mathbb{R}$ a función produto tensorial, $\psi\otimes\varphi:\mathbb{R}^2\longrightarrow\mathbb{R}$, defínese como $\psi\otimes\varphi(u,v) := \psi(u)\varphi(v)$.*

Teorema 4 *Baixo as hipóteses (C1)-(C4), a versión bootstrap suavizada da función MISE de $\tilde{r}_{h,2}$ admite a seguinte expresión pechada:*

$$
\begin{aligned}
MISE^*_{\tilde{r}_{h,2},w}(h) &= \frac{1}{n^2}\int \frac{1}{(1-\hat{F}_g(x))^2}\sum_{i=1}^{n} K_h^2 * K_g(x-X_i)w(x)\,dx \\
&+\frac{n-1}{n^3}\int \frac{1}{(1-\hat{F}_g(x))^2}\sum_{i,j=1}^{n} K_h * K_g(x-X_i)K_h * K_g(x-X_j)w(x)\,dx \\
&+\frac{2}{n^2}\int \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^3}\sum_{i=1}^{n} L_h * K_g(x-X_i)w(x)\,dx - \frac{2}{n}\int \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^3} \\
&\cdot \sum_{i=1}^{n} K_h * K_g(x-X_i)w(x)\,dx + \frac{n-1}{n^3}\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4} \\
&\cdot \sum_{i,j=1}^{n} \mathbb{I}_{\{K_h * K_g\}}(x-X_i)\cdot \mathbb{I}_{\{K_h * K_g\}}(x-X_j)\,w(x)\,dx + \frac{2n-2}{n^3} \\
&\cdot \int \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^3}\sum_{i,j=1}^{n} K_h * K_g(x-X_j)\mathbb{I}_{\{K_h * K_g\}}(x-X_i)\,w(x)\,dx \\
&-\frac{2}{n}\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4}\sum_{i=1}^{n} \mathbb{I}_{\{K_h * K_g\}}(x-X_i)\,w(x)\,dx \\
&+\frac{1}{n^2}\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4}\sum_{i=1}^{n} \mathbb{I}_{\{(K_h \otimes K_h) * K_g\}}(x-X_i, x-X_i)\,w(x)\,dx \\
&+\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4} w(x)\,dx,
\end{aligned}
$$

onde $L_h(z) = \mathbb{K}\left(\frac{z}{h}\right) K_h(z)$.

Definimos o segundo selector de ventá bootstrap para o estimador non paramétrico da función de razón de fallo como: $h_{BOOT2} = h_{\mathrm{MISE}^*_{\tilde{r}_{h,2},w}} = \arg\min_{h>0} \mathrm{MISE}^*_{\tilde{r}_{h,2},w}(h)$.

4. CASO PARTICULAR: NÚCLEO GAUSSIANO

No caso particular de que o núcleo K sexa Gaussiano, isto é, a función de densidade dunha normal estándar, as expresións dadas nos Teoremas 2 e 4 pódense simplificar, o que leva a establecer o Teorema 5.

Teorema 5 *Asumamos as condicións (C1), (C2) e (C4). Se K é o núcleo Gaussiano, entón a versión bootstrap suavizada do MISE de $\tilde{r}_{h,1}$ admite a seguinte expresión:*

$$
\begin{aligned}
MISE^*_{\tilde{r}_{h,1},w}(h) &= \int \left[\frac{1}{n(1-\hat{F}_g(x))}\sum_{i=1}^{n}\left(K_{\sqrt{g^2+h^2}}(x-X_i) - K_g(x-X_i)\right)\right]^2 w(x)dx \\
&+\frac{1}{2\sqrt{\pi}nh}\int \left[\frac{1}{n(1-\hat{F}_g(x))^2}\sum_{i=1}^{n} K_{\sqrt{g^2+h^2/2}}(x-X_i)\right] w(x)\,dx \\
&-\frac{1}{n}\int \left[\frac{1}{n(1-\hat{F}_g(x))}\sum_{i=1}^{n} K_{\sqrt{g^2+h^2}}(x-X_i)\right]^2 w(x)\,dx.
\end{aligned}
$$

De xeito análogo, a versión bootstrap suavizada do MISE de $\tilde{r}_{h,2}$ pódese expresar como segue:

$$
\begin{aligned}
MISE^{*}_{\tilde{r}_{h,2},w}(h) \;=\; & \frac{1}{n^2}\int \frac{1}{2\sqrt{\pi}(1-\hat{F}_g(x))^2}\sum_{i=1}^{n} K_{\sqrt{g^2+h^2/2}}(x-X_i)w(x)\,dx \\
& +\frac{n-1}{n^3}\int \frac{1}{(1-\hat{F}_g(x))^2}\sum_{i,j=1}^{n} K_{\sqrt{h^2+g^2}}(x-X_i)K_{\sqrt{h^2+g^2}}(x-X_j)w(x)\,dx \\
& +\frac{2}{n^2}\int \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^3}\sum_{i=1}^{n} L_h * K_g(x-X_i)w(x)\,dx - \frac{2}{n}\int \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^3} \\
& \cdot\sum_{i=1}^{n} K_{\sqrt{h^2+g^2}}(x-X_i)w(x)\,dx + \int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4}w(x)\,dx \\
& +\frac{n-1}{n^3}\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4}\sum_{i,j=1}^{n}\Phi\left(\frac{x-X_i}{\sqrt{h^2+g^2}}\right)\Phi\left(\frac{x-X_j}{\sqrt{h^2+g^2}}\right)w(x)\,dx \\
& +\frac{2n-2}{n^3}\int \frac{\hat{f}_g(x)}{(1-\hat{F}_g(x))^3}\sum_{i,j=1}^{n} K_{\sqrt{h^2+g^2}}(x-X_j)\,\Phi\left(\frac{x-X_i}{\sqrt{h^2+g^2}}\right)w(x)\,dx \\
& -\frac{2}{n}\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4}\sum_{i=1}^{n}\Phi\left(\frac{x-X_i}{\sqrt{h^2+g^2}}\right)w(x)\,dx \\
& +\frac{1}{n^2}\int \frac{\hat{f}_g^2(x)}{(1-\hat{F}_g(x))^4}\sum_{i=1}^{n}\mathbf{\Phi}_{\mathbf{0},\tilde{H}}\left(x-X_i,x-X_i\right)w(x)\,dx,
\end{aligned}
$$

onde $\Phi(z)$ é a función de distribución dunha variable normal estándar avaliada en z, $\mathbf{\Phi}_{\boldsymbol{\mu},\Sigma}(z_1,z_2)$ é a función de distribución dunha variable normal $N_2(\boldsymbol{\mu},\Sigma)$ avaliada en (z_1,z_2), $\mathbf{0}=(0,0)^T$ e

$$\tilde{H}=\begin{pmatrix} g^2+h^2 & g^2 \\ g^2 & g^2+h^2 \end{pmatrix}.$$

5. SIMULACIÓNS E APLICACIÓN A DATOS REAIS

A continuación, lévase a cabo un estudo de simulación coa fin de comprobar o comportamento empírico dos novos selectores de ventá propostos en cada caso, h_{BOOT1} e h_{BOOT2}, e tamén para comparar o seu comportamento na práctica cos selectores de ventá xa existentes ata o momento: a ventá de validación cruzada de Patil (1993a) e a ventá bootstrap de González-Manteiga et al (1996) pensada orixinalmente para datos con censura (h_{CV} e h^*_{GCM}, respectivamente).

Considéranse seis poboacións diferentes para mostrar os resultados na práctica de cada selector de ventá nas situacións que se contemplan.

Modelo 1: A mostra provén dunha normal estándar.

Modelo 2: Un modelo Gumbel con densidade subxacente dada por

$$f_{\alpha,\lambda}(x)=\alpha\lambda e^{\lambda x-\alpha(e^{\lambda x}-1)}\mathbb{1}_{[0,+\infty)}(x);\lambda=0,\alpha=1.$$

Modelo 3: A mostra é xerada a partir dunha distribución Weibull con densidade

$$f_{\alpha,\lambda}(x)=\alpha\lambda(\lambda x)^{\alpha-1}e^{-(\lambda x)^{\alpha}}\mathbb{1}_{[0,+\infty)}(x);\lambda=\alpha=1.$$

Modelo 4: A distribución de intererse é un modelo χ^2_k, considerando $k=2,3$. A función de densidade subxacente vén dada por:

$$f_k(x)=\begin{cases}\dfrac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2} & \text{se } x>0\\ 0 & \text{noutro caso}\end{cases},$$

onde Γ é a función Gamma.

Modelo 5: Unha mestura de tres densidades normais (Modelo 9 de Marron e Wand (1992)) da forma:

$$\frac{9}{20}N\left(-\frac{6}{5},\frac{9}{25}\right)+\frac{9}{20}N\left(\frac{6}{5},\frac{9}{25}\right)+\frac{1}{10}N\left(0,\frac{1}{16}\right).$$

Simuláronse, a partir destas poboacións, 1000 mostras aleatorias de tamaño $n = 100$ e $n = 200$. Considérase ademais o núcleo Gaussiano para calcular o estimador de Parzen-Rosenblatt. A ventá piloto empregada, g, necesaria para o método bootstrap suavizado, é $g = \widehat{IQR}\lambda n^{-1/7}$, onde $\widehat{IQR}$ é o rango intercuartílico estimado e λ foi escollida como 0.2 e 0.3. Realizouse ademais un estudo de simulación adicional para estudar a elección de λ. Como resultado, a elección de λ afecta en gran medida ao comportamento dos tres parámetros de suavizado bootstrap considerados. Móstrase este feito para o Modelo 5 (Figura 1), considerando doce valores diferentes de λ (especificamente, aqueles do conxunto $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3,\ 0.35, 0.4, 0.45, 0.5, 0.55, 0.6\}$). Neste último caso, o tamaño mostral considerado foi $n = 100$. Ademais, a función de peso w foi escollida como $w(x) = \mathbb{1}_{(-\infty,\hat{Q}_3]}(x)$, onde $\hat{Q}_3$ é o estimador do terceiro cuartil, denominado Q_3. Os catro selectores de ventá h_{CV}, h_{BOOT1}, h_{BOOT2} e h^*_{GCM} son os que minimizan, en h, catro funcións empíricas. Como estes mínimos non teñen expresións explícitas, empregouse un algoritmo numérico para aproximalos, que procede como segue:

Paso 1: Consideramos un conxunto equiespaciado de cinco valores de h no intervalo $[0.01, 10]$.

Paso 2: Para cada método, unha ventá h é escollida entre esas cinco dadas no paso anterior, logo de minimizar a función obxectivo (CV, $\text{MISE}^*_{\tilde{r}_{h,1},w}$, $\text{MISE}^*_{\tilde{r}_{h,2},w}$ ou $\widehat{\text{AMISE}}$). Denotarémola como h_{OPT_1}.

Paso 3: Entre o conxunto de cinco parámetros de ventá definidos no Paso 1, consideramos a anterior e posterior a h_{OPT_1}. Se h_{OPT_1} é a ventá máis pequena (máis grande) da grella, entón h_{OPT_1} será usada no lugar do anterior (posterior) valor de h_{OPT_1} na grella.

Paso 4: Construímos un conxunto de cinco valores equiespaciados de h dentro do intervalo cuxos extremos son os dous valores seleccionados no Paso 3.

Paso 5: Finalmente, repítense 10 veces os Pasos 2-4, obtendo o parámetro ventá óptimo na última etapa.

En primeiro lugar, mostraranse os seguintes resultados aproximados por simulación para algún escenario representativo. Concretamente, na Figura 2 considéranse os Modelos 1, 3 e 4 (con $k = 3$) e inclúese

- $\log(\hat{h}/h_{MISE})$ (panel esquerdo de cada figura)
- $\log(MISE(\hat{h})/MISE(h_{MISE}))$ (panel dereito de cada figura),

onde $\hat{h} = h_{BOOT1}, h_{BOOT2}, h_{CV}, h^*_{GCM}$.

En segundo lugar, na Figura 3 ilústrase o comportamento na práctica de h_{BOOT1}, h_{BOOT2}, h_{CV} e h^*_{GCM}, aplicando estes selectores a un conxunto de datos reais: o número de mulleres con ascendencia india actualmente vivindo preto de Phoenix, Arizona, que realizaron as probas da diabetes. En concreto, analizarase o índice de masa corporal por medio da estimación non paramétrica da función de razón de fallo, considerando tres poboacións: as mulleres diagnosticadas con diabetes ($n = 177$), as mulleres que resultaron non padecer de diabetes ($n = 355$), e a poboación total ($n = 532$). Pola súa banda, na Figura 4 amósanse as anteriores estimacións todas xuntas.

Finalmente, as Táboas 1 e 2 mostran, respectivamente, os parámetros de ventá obtidos e os tempos de computación (en segundos) requeridos para a obtención dos mesmos.

6. CONCLUSIÓNS

Os resultados de simulación amosan que o parámetro de suavizado con mellor comportamento na práctica é h_{BOOT1}. En particular, como se pode ver na Figura 2 (paneis do lado esquerdo), a mediana de h_{BOOT1} é aproximadamente h_{MISE}. Ademais, h^*_{GCM} sitúase na segunda posición da clasificación.

Atendendo ao MISE, h_{BOOT1} manifesta mellor comportamento empírico que o seu principal competidor, h^*_{GCM}, como podemos ver na Figura 2 (paneis centrais e do lado dereito). Por outra banda, h_{CV} é, sen lugar a dúbidas, o peor, presentando maior varianza que h_{BOOT1}, h^*_{GCM} e incluso que h_{BOOT2}. En concreto, h_{CV} sempre produce estimacións sobre ou infrasuavizadas da función de razón de fallo, dependendo do modelo a estudar.

Considerando o selector de ventá h_{BOOT2}, temos que éste tamén xera estimacións sobresuavizadas da función de razón de fallo en case todos os casos. Especificamente, a Figura 2 (paneis centrais) amosa que MISE(h_{BOOT2}) é maior que MISE(h_{BOOT1}) ou MISE(h^*_{GCM}). Sen embargo, no caso de estudar o Modelo 1, o comportamento de h_{BOOT2} é mellor (ver Figura 2, primeira fila).

Para rematar coa análise dos resultados de simulación, se nos fixamos na Figura 1, resulta evidente que o mellor comportamento dos selectores de ventá h_{BOOT1}, h_{BOOT2} e h^*_{GCM} acádase cando o parámetro λ toma os valores 0.2, 0.25 ou 0.3.

Centrándonos agora na aplicación a datos reais e, en concreto, na Figura 3, vemos que h_{CV} parece sobresuavizar a estimación da función de razón de fallo. Concretamente, a función $CV(h)$ é decrecente, para todo valor de h comprendido no intervalo de búsqueda $[0.01, 10]$, polo que $CV(h)$ acada o valor máis baixo no extremo dereito do mesmo, é dicir, $h_{CV} = 10$.

Por outra banda, o comportamento na práctica de h_{BOOT2} parece ser moito mellor que o de h_{CV}. Así e todo, h_{BOOT2} semella que produce, tamén, unha sobresuavización da estimación da función de razón de fallo, aínda que dun xeito máis tenue (ver Táboa 1). Cabe destacar, ademais, que tanto h_{CV} como h_{BOOT2} son os que requiren máis tempo de computación (ver Táboa 2) con ampla diferencia.

O selector de ventá bootstrap, h^*_{GCM}, parece estar posicionado novamente no segundo lugar da clasificación, pois aínda que semella ter bo comportamento empírico a maioría das veces, se atendemos á Táboa 1 (segunda fila), vemos que h^*_{GCM} é próximo a h_{BOOT2}. En consecuencia, h^*_{GCM} parece que nos achega unha sobresuavización da estimación da función de razón de fallo neste escenario.

Finalmente, h_{BOOT1} produce unha estimación suavizada da función de razón de fallo en todos os casos, que semella ser a máis precisa. Cabe mencionar que h_{BOOT1} e h^*_{GCM} requiren un tempo de computación similar.

AGRADECEMENTOS

Os autores agradecen o financiamento do MINECO a través do proxecto MTM2014-52876-R e da Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 e Centro Singular de Investigación de Galicia ED431G/01). Todas estas axudas foron cofinanciadas polo FEDER. Parte do traballo do primeiro autor foi levado a cabo durante unha visita á University of California, San Diego, financiada por INDITEX, con referencia INDITEX-UDC 217.

REFERENCIAS

Barbeito, I. and Cao, R. (2016) Smoothed stationary bootstrap bandwidth selection for density estimation with dependent data. Comput. Statist. & Data Anal. 104, 130–147.

Barbeito, I. and Cao, R. (2017) A review and some new proposals for bandwidth selection in nonparametric density estimation for dependent data. To appear in 'From Statistics to Mathematical Finance'. Festschrift in Honour of Winfried Stute. Springer.

Barbeito, I. and Cao, R. (2017) Smoothed bootstrap bandwidth selection for nonparametric hazard rate estimation. Preprint.

Cao, R. (1993) Bootstrapping the mean integrated squared error. J. Mult. Anal. 45, 137–160.

Efron, B. and Tibishirani, R. (1993) An Introduction to the Bootstrap. Chapman and Hall, New York.

Estévez-Pérez, G., Quintela del Río, A. and Vieu, P. (2002) Convergence rate for cross-validatory bandwidth in kernel hazard estimation from dependent samples. J. Stat. Plan. Infer. 104, 1—30.

González-Manteiga, W., Cao, R. and Marron, J. (1996) Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. J. Am. Stat. Assoc. 91(435), 1130—1140.

Lo, S., Mack, Y. and Wang, J. (1989) Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. Probab. Theory Relat. Fields. 80, 461–473.

Marron, J. and Wand, M. (1992) Exact mean integrated squared error. Ann. Statist. 20(2), 712–736.

Murthy, V.K. (1965) Estimation of jumps, reliaility and hazard rate. Ann. Math. Statist. 36(3), 1032–1040.

Patil, P. (1993a) Bandwidth choice for nonparametric hazard rate estimation. J. Stat. Plan. Infer. 35, 15–30.

Patil, P. (1993b) On the least squares cross-validation bandwidth in hazard rate estimation. Ann. Math. Statist. 33, 1065–1076.

Rao, B.P. (1983) Nonparametric Functional Estimation. Academic Press, New York.

Rice, J. and Rosenblatt, M. (1976) Estimation of the log survivor function and hazard function. Sankhyā Ser. A. 38(1), 60–78.

Sarda, P. and Vieu, P. (1991) Smoothing parameter selection in hazard estimation. Statist. Probab. Lett. 11, 429–434.

Silverman, B. and Young, G. (1987) The bootstrap: To smooth or not to smooth? Biometrika. 74, 469–479.

Singpurwalla, N.D. and Young, M.Y. (1980) Estimation of the failure rate - A survey of nonparametric methods. Part I: Non-Bayesian methods. Commun. Stat. Theory Methods. 12, 5–40.

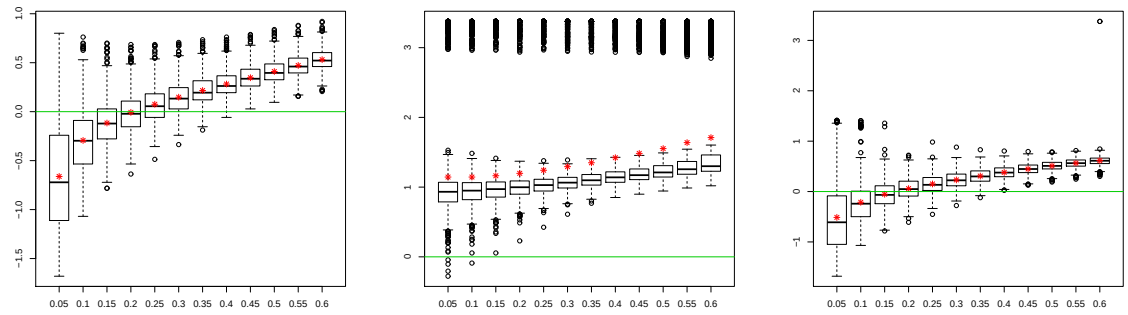
Singpurwalla, N.D. and Young, M.Y. (1983) Kernel estimators of the failure-rate function and density estimation: An analogy. J. Am. Stat. Assoc. 78(382), 478–481.

Tanner, M. and Wong, W. (1983) The estimation of the hazard function from randomly censored data by the kernel method. Ann. Statist. 11, 989–993.

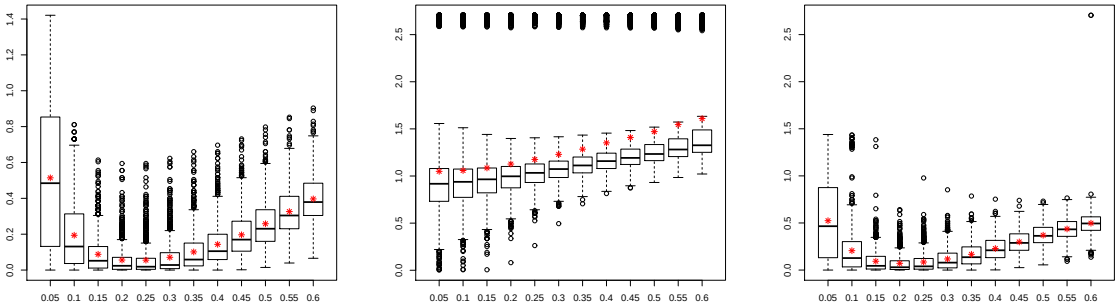
Tanner, M. and Wong, W. (1984) Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. J. Am. Stat. Assoc. 79, 174–182.

Watson, G., Leadbetter, M. (1964a) Hazard analysis i. Biometrika 51, 175–184.

Watson, G., Leadbetter, M. (1964b) Hazard analysis ii. Sankhyā. Ser. A. 26, 101–116.



(a) $\hat{h} = h_{BOOT1}$ (b) $\hat{h} = h_{BOOT2}$ (c) $\hat{h} = h^*_{GCM}$



(d) $\hat{h} = h_{BOOT1}$ (e) $\hat{h} = h_{BOOT2}$ (f) $\hat{h} = h^*_{GCM}$

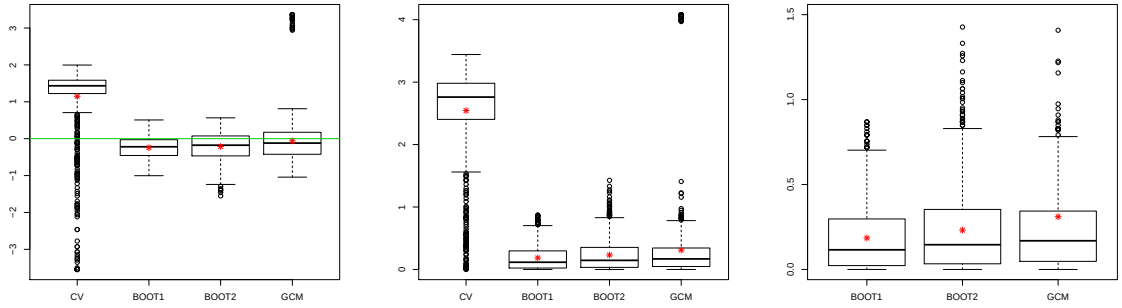
Figura 1: Boxplots de $\log\left(\hat{h}/h_{MISE}\right)$ (fila superior) e $\log\left(MISE(\hat{h})/MISE(h_{MISE})\right)$ (fila inferior) para o Modelo 5, $\lambda \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6\}$ e $n = 100$, onde $\hat{h} = h_{BOOT1}$ (esquerda), h_{BOOT2} (centro) e h^*_{GCM} (dereita).

	h_{BOOT1}	h_{BOOT2}	h_{CV}	h^*_{GCM}
Mulleres diabéticas	1.5026	2.6636	10	1.5173
Mulleres non diabéticas	1.9904	2.4246	10	2.3124
Poboación total	1.3514	1.7465	10	1.5758

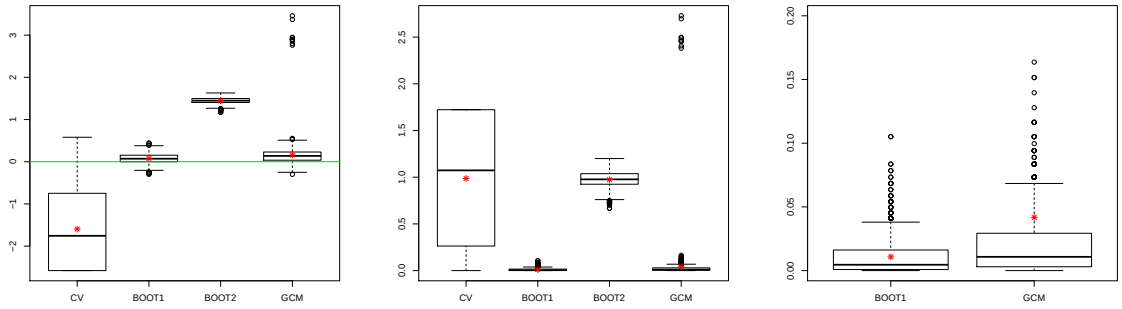
Táboa 1: Parámetros de suavizado para o Índice de Masa Corporal considerando o conxunto de datos das mulleres diabéticas, o conxunto de datos das mulleres non diabéticas e o conxunto de datos da poboación total.

	$CPU_{h_{BOOT1}}$	$CPU_{h_{BOOT2}}$	$CPU_{h_{CV}}$	$CPU_{h^*_{GCM}}$
Mulleres diabéticas	2.04	272.74	134.78	2.24
Mulleres non diabéticas	2.49	678.08	589.37	3.15
Poboación total	5.05	1475.3	1349.5	4.73

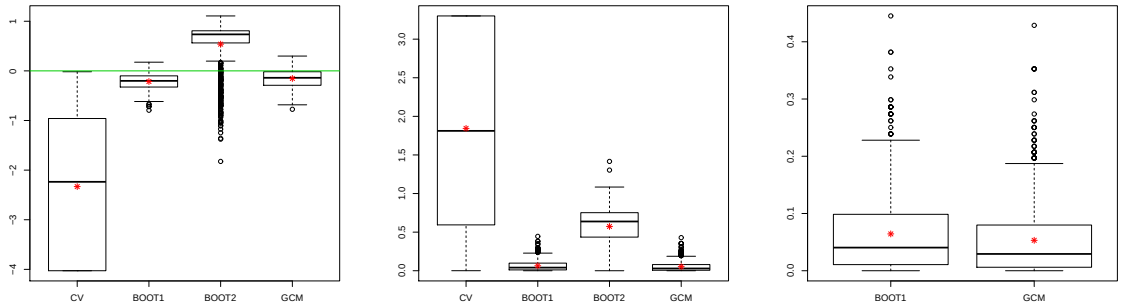
Táboa 2: Tempos de CPU (en segundos) para os catro parámetros de suavizado e empregando o Índice de Masa Corporal, considerando o conxunto de datos das mulleres diabéticas, o conxunto de datos das mulleres non diabéticas e o conxunto de datos da poboación total.



(a) Modelo 1 (b) Modelo 1 (c) Modelo 1

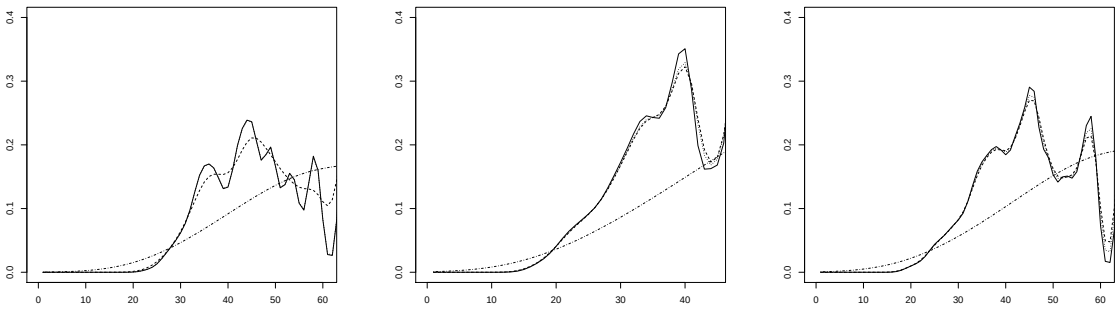


(d) Modelo 3 (e) Modelo 3 (f) Modelo 3



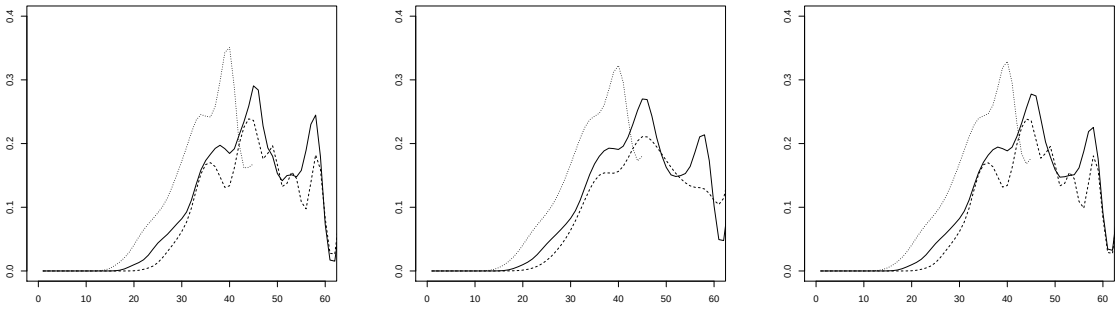
(g) Modelo 4, $k = 3$ (h) Modelo 4, $k = 3$ (i) Modelo 4, $k = 3$

Figura 2: Boxplots de $\log(\hat{h}/h_{MISE})$ (lado esquerdo) e $\log(MISE(\hat{h})/MISE(h_{MISE}))$ (centro) para os Modelo 1 (primeira fila), 3 (fila central) e 4 con $k = 3$ (última fila), $n = 200$ e $\lambda = 0.2$. As figura dos paneis dereitos son zooms das figuras dos paneis do centro.



(a) Mulleres diabéticas (b) Mulleres non diabéticas (c) Poboación total

Figura 3: Estimación non paramétriza da función de razón de fallo para o Índice de Masa Corporal considerando o conxunto de datos de mulleres diabéticas (esquerda), o conxunto de datos de mulleres non diabéticas (centro) e a poboación completa (dereita), empregando os parámetros de suavizado h_{BOOT1} (liña continua), h_{BOOT2} (liña con guións), h_{CV} (liña con guións e puntos) e h^*_{GCM} (liña de puntos).



(a) h_{BOOT1} (b) h_{BOOT2} (c) h^*_{GCM}

Figura 4: Estimación non paramétrica da función de razón de fallo para o Índice de Masa Corporal considerando o conxunto de datos de mulleres diabéticas (liñas con guións), mulleres non diabéticas (liñas punteadas) e a poboación total (liñas continuas), considerando os selectores de ventá h_{BOOT1} (panel esquerdo), h_{BOOT2} (panel central) e h^*_{GCM} (panel dereito).

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Intervalos de confianza bootstrap para a densidade condicional en procesos de Markov

Inés Barbeito[1], Ricardo Cao[1,3] e Dimitris Politis[2]

[1]Grupo de Investigación MODES, Departamento de Matemáticas, Centro de Investigación CITIC, Universidade da Coruña
[2] Departamento de Matemáticas, University of California, San Diego
[3] ITMATI

RESUMO

Establécense intervalos de confianza bootstrap para a densidade condicional en procesos de Markov. Para iso, revísanse algoritmos bootstrap recentes na literatura actual pensados para construir intervalos de predición. Estes métodos son: model-free bootstrap, model-free preditivo, limit-model-free, o bootstrap pensado para modelos autorregresivos non paramétricos con residuos axustados e preditivos, o bootstrap local e o bootstrap baseado na estimación das densidades de transición (ver Pan e Politis (2016a, 2016b), Politis (2015), Paparoditis e Politis (2001,2002) e Rajarshi (1990), respectivamente). Sen embargo, para obter unha boa probabilidade de cobertura, é moi importante usar un selector de ventá apropiado á hora de estimar a función de densidade condicional. Neste sentido, considéranse os parámetros de suavizado escollidos mediante a regra do pulgar e de xeito determinista. Para comprobar o bo funcionamento na práctica destes métodos, realízase un estudo de simulación en cada un dos contextos, comparando estas propostas coas xa existentes na literatura.

Palabras e frases chave: model-free bootstrap, método kernel, bootstrap local, bootstrap para modelos autorregresivos

1. INTRODUCIÓN

Os métodos bootstrap para series de tempo foron extensivamente estudados ao longo das últimas tres décadas. Recentemente, Kreiss e Paparoditis (2011) ofreceron unha revisión do estado da arte da literatura.

Consideremos o proceso $(X_1, \ldots, X_n)$, e en particular, os pares $(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})$, con $Y_i = X_{i+1}$ e $i = 1, 2, \ldots, n-1$. Neste contexto, o noso propósito é construir intervalos de confianza para a función de densidade de Y_i condicionada a $X_i = x$, estudado na Sección 2 e para o cal foron xa propostos dous métodos:

1. O bootstrap baseado no estimador tipo núcleo da densidade de transición dun proceso de Markov (ver Rajarshi (1990)).

2. O bootstrap local para procesos de Markov (ver Paparoditis e Politis (2001,2002)).

Ademais, en termos de establecer intervalos de predición, Pan e Politis (ver Pan e Politis (2016a) e Politis (2015)) propuxeron tres algoritmos bootstrap: o model-free bootstrap para procesos de Markov, o model-free bootstrap preditivo e o limit-model-free bootstrap. En concreto, o algoritmo model-free bootstrap foi estendido por Politis (2015) ao caso de construir intervalos de confianza para a esperanza condicional.

A idea principal destes últimos tres algoritmos é transformar o conxunto de datos dependentes noutro conxunto de datos que sexa i.i.d., e polo tanto máis fácil de manexar. O model-free bootstrap remostrexa directamente do conxunto de datos transformados, i.i.d. Para iso, Pan e Politis (2016a) denotan por $D_x(y)$ á función de distribución condicional de Y_t con $t \in \mathbb{Z}$, que vén definida por:

$$D_x(y) = \mathbb{P}\left(Y_t \leq y | X_t = x\right), x \in \mathbb{R}.$$

Logo, dadas as observacións $X_1, \ldots, X_n$, Pan e Politis (2016a) propoñen estimar $D_x(y)$ como segue:

$$\hat{D}_x(y) = \frac{\sum_{i=1}^{n-1} \mathbf{1}\{Y_i \leq y\} K\left(\frac{x - X_i}{h}\right)}{\sum_{k=1}^{n-1} K\left(\frac{x - X_k}{h}\right)},$$

onde K é unha función tipo núcleo.

Propoñen ademais unha versión suavizada, e polo tanto diferenciable, do estimador da distribución condicional, denotado por $\bar{D}_x(y)$. Sexa $\Lambda(\cdot)$ unha función de distribución cumulativa que é absolutamente continua e estrictamente crecente sobre o seu soporte, e h_0 e h parámetros ventá positivos. Defínese o estimador suavizado de $D_x(y)$ como:

$$\bar{D}_x(y) \quad = \quad \frac{\sum_{i=1}^{n-1} \Lambda\left(\frac{y - Y_i}{h_0}\right) K\left(\frac{x - X_i}{h}\right)}{\sum_{k=1}^{n-1} K\left(\frac{x - X_k}{h}\right)}.$$

En consecuencia, os datos transformados $\{\nu_t; t = 1, \ldots, n-1\}$ poden ser calculados como:

$$\nu_t = \bar{D}_{Y_t}(X_{t+1}); t = 1, 2, \ldots, n-1. \tag{1}$$

Pola súa banda, se consideramos un modelo autorregresivo non paramétrico, Pan e Politis (2016b) propuxeron dous métodos bootstrap para axustar o modelo empregando o método núcleo, asumindo que os erros son i.i.d.:

1. Bootstrap con residuos axustados.
2. Bootstrap con residuos preditivos.

Nesta comunicación, os algoritmos bootstrap mencionados son comparados empiricamente mediante a construción de intervalos de confianza para a densidade condicional no contexto de procesos de Markov. Así, na Sección 2, trátanse distintos algoritmos bootstrap para a construcción de intervalos de confianza para a densidade condicional. Por último, na Sección 3, realízase un estudo de simulación para comprobar e comparar o bo comportamento na práctica destes algoritmos. As conclusións derivadas da Sección 3 expoñeranse na Sección 4.

2. INTERVALOS DE CONFIANZA BOOTSTRAP

Consideremos unha mostra aleatoria bidimensional, $(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})$, con $Y_i = X_{i+1}$ e $i = 1, 2, \ldots, n-1$. O noso propósito é construir intervalos de confianza para a función de densidade de Y_i condicionada a $X_i = x$, é dicir, para a función de densidade de X_{i+1} condicionada a $X_i = x$, que denotaremos por $f(y|x)$, e cuxo estimador non paramétrico tipo núcleo vén dado por:

$$\hat{f}_{h_1,h_2}(y|x) = \frac{\hat{f}_{h_1,h_2}(x,y)}{\hat{f}_{h_1}(x)}, \tag{2}$$

onde $\hat{f}_{h_1,h_2}(x,y) = \dfrac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}, \frac{y-Y_i}{h_2}\right)$, K é unha densidade de probabilidade definida en $\mathbb{R}^2$, $x, y \in \mathbb{R}$, $h_1, h_2 > 0$ son parámetros ventá e $\hat{f}_{h_1}(x) = \dfrac{1}{nh_1}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right)$.

Con esta finalidade, describiranse a continuación sete algoritmos bootstrap para a obtención de intervalos de confianza para a densidade condicional.

Algoritmo 1: Bootstrap con residuos axustados

1. Construir o estimador tipo núcleo de Nadaraya-Watson da función de regresión, dado por:

$$\hat{m}_h(x) = \frac{\sum_{t=1}^{n-1} K\left(\dfrac{x-X_t}{h}\right) X_{t+1}}{\sum_{t=1}^{n-1} K\left(\dfrac{x-X_t}{h}\right)}, \text{ para todo } x \in \mathbb{R},$$

   onde $K(\cdot)$ é unha función de densidade simétrica.

2. Analogamente, construir o estimador tipo núcleo da función de densidade condicional, dado en (2).

3. Calcular os residuos do seguinte xeito: $\hat{\epsilon}_i = X_{i+1} - \hat{m}_h(X_i)$, para $i = 1, \ldots, n-1$.

4. Centrar os residuos: $\hat{r}_i = \hat{\epsilon}_i - (n-1)^{-1}\sum_{t=1}^{n-1} \hat{\epsilon}_t$, para $i = 1, \ldots, n-1$.

5. Mostrexar aleatoriamente con reemprazamento os valores $\hat{r}_1, \ldots, \hat{r}_{n-1}$ para crear 'erros' bootstrap ϵ_i^*, con $i = 1, \ldots, n$.

6. Sexa $X_1^* = X_I$, onde I é unha variable aleatoria discreta uniforme nos valores $1, \ldots, n$. Logo, xerar X_{i+1}^* como segue:

$$X_{i+1}^* = \hat{m}_h(X_i^*) + \epsilon_i^*; i = 1, \ldots, n-1.$$

7. Recalcular o estimador tipo núcleo da función de densidade condicional a partir da serie bootstrap $\{X_1^*, \ldots, X_n^*\}$.

8. Calcular a versión bootstrap do estatístico de interese, $\hat{f}_{h_1,h_2}^*(y|x) - \hat{f}_g(y|x)$.

9. Repetir os pasos 5-8 B veces, de xeito que coas B réplicas bootstrap recollidas se calcule o cuantil empírico de orde α, denotado como $q(\alpha)$.

10. A continuación, construimos o intervalo de confianza $(1-\alpha)$ para a función de densidade condicional, dado por:

$$\left[\hat{f}_{h_1,h_2}(y|x) - q(1-\alpha/2), \hat{f}_{h_1,h_2}(y|x) - q(\alpha/2)\right].$$

Algoritmo 2: Bootstrap con residuos preditivos

1. Pasos 1 e 2 do Algoritmo 1.

2. Usar a mostra eliminando a observación Y_t para calcular o estimador tipo núcleo:

$$\hat{m}_h^{(t)}(x) = \frac{\sum_{i=1, i\neq t}^{n-1} K\left(\dfrac{x-X_i}{h}\right) X_{i+1}}{\sum_{i=1, i\neq t}^{n-1} K\left(\dfrac{x-X_i}{h}\right)}, \text{ para } t = 1, \ldots, n-1.$$

3. Calcular os residuos preditivos como segue: $\hat{\epsilon}_t^{(t)} = X_{t+1} - \hat{m}_h^{(t)}(X_t)$, para $t = 1, \ldots, n-1$.

4. Reemprazar $\hat{\epsilon}_t$ por $\hat{\epsilon}_t^{(t)}$ no Algoritmo 1; o resto dos pasos son análogos.

Algoritmo 3: Bootstrap baseado no estimador tipo núcleo da densidade de transición

1. Construir o estimador tipo núcleo da densidade condicional dado en (2).

2. Xerar X_1^* a partir do estimador tipo núcleo da función de densidade. Alternativamente, fixar $X_1^* = X_I$, sendo I unha variable discreta uniforme nos valores $1, \ldots, n$, é dicir, elexir X_1^* como unha observación da serie $X_1, \ldots, X_n$.

3. Xerar $X_{t+1}^* \sim \hat{f}_{h_1,h_2}(\cdot|X_t^*)$, para $t = 1, \ldots, n-1$.

4. Construir o estimador non paramétrico da densidade condicional de xeito similar ao paso 1, pero considerando a serie bootstrap $\{X_1^*, \ldots, X_n^*\}$ en vez da serie orixinal.

5. Calcular a versión bootstrap do estatístico de interese, $\hat{f}^*_{h_1,h_2}(y|x) - \hat{f}_g(y|x)$.

6. Repetir os pasos 2-5 B veces, obtendo B réplicas bootstrap do estatístico cuxo cuantil empírico α será denotado como $q(\alpha)$.

7. A continuación, construimos o intervalo de confianza $(1-\alpha)$ para a función de densidade condicional, dado por:

$$\left[\hat{f}_{h_1,h_2}(y|x) - q(1-\alpha/2), \hat{f}_{h_1,h_2}(y|x) - q(\alpha/2)\right].$$

Algoritmo 4: Bootstrap local

1. Construir o estimador tipo núcleo da densidade condicional dado en (2).

2. Sexa $X_1^* = X_I$, onde I foi xerado como unha variable aleatoria discreta uniforme nos valores $1, \ldots, n$.

3. Escoller un núcleo W e unha ventá piloto, $g > 0$. Supoñer que foron xerados $X_1^*, \ldots, X_t^*$, para $t \geq 1$. Sexa J unha variable aleatoria discreta tomando valores no conxunto $\{1, \ldots, n-1\}$ con función de masa de probabilidade dada por:

$$\mathbb{P}(J = s) = \frac{W_g(X_t^* - X_s)}{\sum_{m=1}^{n-1} W_g(X_t^* - X_m)}.$$

   Entón xeramos $X_{t+1}^* = X_{J+1}$.

4. Repetir 3 para $t = 1, 2, \ldots, n-1$ para xerar $X_2^*, \ldots, X_n^*$.

5. Construir o estimador tipo núcleo da densidade condicional de xeito similar ao paso 1, pero considerando a serie bootstrap $\{X_1^*, \ldots, X_n^*\}$ en vez da serie orixinal.

6. Calcular a versión bootstrap do estatístico de interese, $\hat{f}^*_{h_1,h_2}(y|x) - \hat{f}_g(y|x)$.

7. Repetir os pasos 2-6 B veces, obtendo B réplicas bootstrap do estatístico cuxo cuantil empírico α será denotado como $q(\alpha)$.

8. A continuación, construimos o intervalo de confianza $(1-\alpha)$ para a función de densidade condicional, dado por:

$$\left[\hat{f}_{h_1,h_2}(y|x) - q(1-\alpha/2), \hat{f}_{h_1,h_2}(y|x) - q(\alpha/2)\right].$$

Algoritmo 5: Model-free bootstrap

1. Obter os datos transformados $\nu_2, \ldots, \nu_n$ a partir da expresión (1).
2. Construir o estimador tipo núcleo da densidade condicional dado en (2).
3. Remostrexar aleatoriamente (con reemprazamento) as variables transformadas $\nu_2, \ldots, \nu_n$ para crear $\nu^*_{-M}, \nu^*_{-M+1}, \ldots, \nu^*_0, \nu^*_1, \ldots, \nu^*_{n-1}, \nu^*_n$, para algún enteiro positivo grande M.
4. Fixar $X^*_{-M+1} = X_I$, sendo I xerada como unha variable aleatoria uniforme nos valores $1, \ldots, n$.
5. Xerar $X^*_{t+1} = \bar{D}^{-1}_{X^*_t}(\nu^*_t)$, para $t = -M+1, \ldots, n-1$.
6. Construir o estimador tipo núcleo da densidade condicional de xeito similar ao paso 1, pero considerando a serie bootstrap $\{X^*_1, \ldots, X^*_n\}$ en vez da serie orixinal.
7. Calcular a versión bootstrap do estatístico de interese, $\hat{f}^*_{h_1,h_2}(y|x) - \hat{f}_g(y|x)$.
8. Repetir os pasos 3-7 B veces, obtendo B réplicas bootstrap do estatístico cuxo cuantil empírico α será denotado como $q(\alpha)$.
9. A continuación, construimos o intervalo de confianza $(1-\alpha)$ para a función de densidade condicional, dado por:

$$\left[\hat{f}_{h_1,h_2}(y|x) - q(1-\alpha/2), \hat{f}_{h_1,h_2}(y|x) - q(\alpha/2)\right].$$

Algoritmo 6: Model-free bootstrap preditivo

1. Calcular as variables $\nu^{(2)}_2, \ldots, \nu^{(n)}_n$ a partir da seguinte expresión:

$$\nu^{(t)}_t \quad = \quad \bar{D}^{(t)}_{X_t}(X_{t+1}), \text{ para } t = 1, \ldots, n-1, \tag{3}$$

onde $\bar{D}^{(t)}_x(y)$ é o estimador de $D_x(y)$ baseado no conxunto de datos que exclúe o par (X_t, X_{t+1}), e que vén dado por:

$$\bar{D}^{(t)}_x(y) = \frac{\sum_{i=1, i\neq t}^{n-1} \Lambda\left(\frac{y - Y_i}{h_0}\right) K\left(\frac{x_t - X_i}{h}\right)}{\sum_{k=1, k\neq t}^{n-1} K\left(\frac{x_t - X_k}{h}\right)}, \text{ para } t = 1, \ldots, n-1.$$

2. O algoritmo é idéntico ao Algoritmo 5, logo de substituir as variables $\nu^{(2)}_2, \ldots, \nu^{(n)}_n$ obtidas a partir da expresión (3) en lugar de $\nu_2, \ldots, \nu_n$.

Algoritmo 7: Limit model-free bootstrap

1. Construir o estimador tipo núcleo da densidade condicional dado en (2).
2. Para algún enteiro positivo grande, M, xerar $\nu^*_{-M}, \nu^*_{-M+1}, \ldots, \nu^*_0, \nu^*_1, \ldots, \nu^*_{n-1}, \nu^*_n$ como i.i.d. a partir dunha variable aleatoria Uniforme(0,1).
3. Fixar $X^*_{-M+1} = X_I$, sendo I unha variable aleatoria discreta uniforme nos valores $1, \ldots, n$.
4. Xerar $X^*_{t+1} = \hat{D}^{-1}_{X^*_t}(\nu^*_t)$, para $t = -M+1, \ldots, n-1$.

5. Construir $\hat{f}^*_{h_1,h_2}(y|x)$ de xeito similar ao paso 1, pero considerando a serie bootstrap $\{X_1^*, \ldots, X_n^*\}$ en vez da serie orixinal.

6. Calcular a versión bootstrap do estatístico de interese, $\hat{f}^*_{h_1,h_2}(y|x) - \hat{f}_g(y|x)$.

7. Repetir os pasos 2-6 B veces, obtendo B réplicas bootstrap do estatístico cuxo cuantil empírico α será denotado como $q(\alpha)$.

8. A continuación, construimos o intervalo de confianza $(1-\alpha)$ para a función de densidade condicional, dado por:

$$\left[\hat{f}_{h_1,h_2}(y|x) - q(1-\alpha/2), \hat{f}_{h_1,h_2}(y|x) - q(\alpha/2)\right].$$

3. ESTUDO DE SIMULACIÓN

A continuación, lévase a cabo un estudo de simulación coa fin de comparar empiricamente as probabilidades de cobertura obtidas polos Algoritmos 1-7. Consideraremos 500 réplicas bootstrap e 500 mostras aleatorias de tamaño $n = 100$ que proveñen dun modelo $AR(1)$ dado por: $X_t = \phi X_{t-1} + (1-\phi^2)^{1/2} a_t$. Neste caso, a_t son variables aleatorias *iid* con distribucón común normal estándar. Considérase que a autocorrelación é $\phi = -0.3$. Pola súa banda, os puntos de avaliación empregados no estudo de simulación son $(x_p, y_p) = (0.02, 0.1)$ e $(x_p, y_p) = (-0.02, 0.1)$. Ademais, a distribución marxinal vén dada por unha variable normal estándar.

Pola súa banda, os parámetros ventá h_1, h_2 foron escollidos de dous xeitos diferentes:

1. Pola regra do pulgar, de xeito que $h_1 = h_2 = 0.9 \min(\hat{\sigma}, IQR/1.34) n^{-1/4}$, onde $\hat{\sigma}$ é a estimación da desviación típica dos datos, e IQR é o rango intercuartílico.

2. Os parámetros ventá son escollidos de xeito determinista. Así, $h_1 = h_2 = 0.9 n^{-1/4}$.

En canto á ventá piloto, $g > 0$, escolleuse de tal xeito que $g = C \cdot n^{-1/6}$, onde $C = 0.9$.

	Regra do pulgar		Ventá determinista	
Cobertura nominal	90%	95%	90%	95%
A.1	0.746	0.828	0.8264	0.9099
A.2	0.74	0.812	0.8135	0.8617
A.3	0.836	0.898	0.881	0.9292
A.4	0.864	0.922	0.9132	0.9485
A.5	0.81	0.87	0.8778	0.9325
A.6	0.848	0.906	0.7556	0.8585
A.7	0.882	0.94	0.926	0.9678

Táboa 1: Probabilidades de cobertura acadadas empregando os Algoritmos 1-7, considerando a ventá determinista e a ventá escollida pola regra do pulgar e $x_p = 0.02, y_p = 0.1$ como punto de avaliación.

	Regra do pulgar		Ventá determinista	
Cobertura nominal	90%	95%	90%	95%
A.1	0.756	0.83	0.8031	0.8571
A.2	0.742	0.816	0.7992	0.8571
A.3	0.844	0.896	0.8842	0.9266
A.4	0.87	0.922	0.9035	0.9382
A.5	0.808	0.874	0.8726	0.9112
A.6	0.846	0.912	0.7915	0.8996
A.7	0.89	0.934	0.9344	0.9652

Táboa 2: Probabilidades de cobertura acadadas empregando os Algoritmos 1-7, considerando a ventá determinista e a ventá escollida pola regra do pulgar e $x_p = -0.02, y_p = 0.1$ como punto de avaliación.

4. CONCLUSIÓNS

Os resultados de simulación amosan que os algoritmos cos que se acadan mellores probabilidades de cobertura empíricas son o limit-model-free bootstrap e o bootstrap local. Así, atendendo ás Táboas 1 e 2, vemos que ámbolos dous métodos presentan probabilidades de cobertura empíricas achegadas ás coberturas nominais consideradas (90% e 95%).

Por outra banda, considerando as Táboas 1 e 2, temos que os métodos cos que se acadan peores probabilidades de cobertura son os Algoritmos 1 e 2 (bootstrap con residuos axustados e preditivos, respectivamente), baseados en modelos non paramétricos de autorregresión.

AGRADECEMENTOS

Os dous primeiros autores agradecen o financiamento do MINECO a través do proxecto MTM2014-52876-R e da Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 e Centro Singular de Investigación de Galicia ED431G/01). Todas estas axudas foron cofinanciadas polo FEDER. O traballo da primeira autora foi levado a cabo durante unha visita á University of California, San Diego, financiada por INDITEX, con referencia INDITEX-UDC 2017.

REFERENCIAS

Kreiss J.P., Paparoditis E. (2011) Bootstrap methods for dependent data: a review. J. Korean Stat. Soc. 40(4), 357—378.

Pan, L., Politis, D.N. (2016a) Bootstrap prediction intervals for Markov processes. Comp. Statist. & Data Anal. 100, 467–494.

Pan, L., Politis, D.N. (2016b) Bootstrap prediction intervals for linear, nonlinear, and nonparametric autoregressions. J. Statist. Plan. Infer. 177, 1–27.

Paparoditis, E., Politis, D.N. (2001) The local bootstrap for Markov processes. J. Statist. Plan. Infer. 108(1-2), 301–328.

Paparoditis, E., Politis, D.N. (2002) A Markovian local resampling scheme for nonparametric estimators in time series analysis. Econometr. Theor. 17, 540–566.

Politis, D. N. (2015) Model-free prediction and regression: a transformation-based approach to inference. Springer.

Rajarshi, M.B. (1990) Bootstrap in Markov-sequences based on estimates of transition density. Ann. Inst. Stat. Math. 42(2), 253–268.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Hart and Mas-Colell consistency in PERT problems

Gustavo Bergantiños[1], Alfredo Valencia-Toledo[2], and Juan Vidal-Puga[1]

[1]Departamento de Estadística e Investigación Operativa, Universidade de Vigo, Spain
[2]Research Group in Economic Analysis, Universidade de Vigo, Spain

ABSTRACT

The Program Evaluation Review Technique (PERT) is a tool used to schedule and coordinate activities in a complex project. In assigning the cost of a potential delay, we characterize the Shapley rule as the only rule that satisfies Hart and Mas-Colell consistency and other desirable properties.

Keywords: PERT problem; consistency; delay

1. INTRODUCTION

Large-scale projects require the coordination of numerous activities, some of which can be performed sequentially, while others can take place in parallel. The Program Evaluation and Review Technique (PERT) is an operational research tool used to schedule and coordinate activities. The tool was developed by the U.S. Navy in the late 1950s to manage the Polaris submarine missile program, a project involving thousands of contractors.[1] The tool uses networks to model all the activities and can help reduce both the time and cost required to finish a project.

PERT planning involves several steps, including estimating of the time required for each activity and determining the *critical activities* (activities that have the potential to delay the entire project). In the example shown in Figure 1, each arc represents an activity and the number in brackets the corresponding time. The critical activities are a, b, c, d and g.

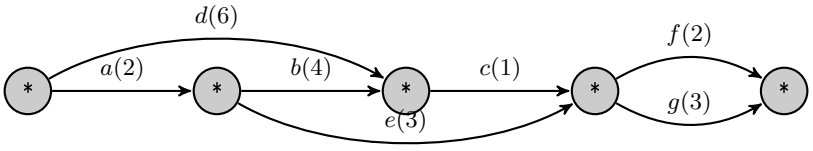


Figura 1: A PERT graph.

For projects that suffer from delays, it may be useful to identify the activities involved and the responsibility of each for causing the delay. Although costs typically arise when a project is not completed on time, it may also be the case these can be reduced by ensuring certain important

[1]Another similar technique is the Critical Path Method (CPM) which was developed in 1957 for project management in the private sector. CPM and PERT have become synonyms with a number of variations used to refer to the same technique (PERT, CPM, or PERT/CPM).

parts of the project are completed. Consider the example of Figure 1. The duration of such project is 10. Assume that we must pay a cost of 5 for each unit of time the whole project is delayed. Besides, activities a, b, c, and d have been completed in their times. Nevertheless activity e has been delayed 4 units (namely, it has been completed in 7 units, instead of 3), activity f has been delayed 3 units and activity g 1 unit. Then, the real duration of the project has been 14, which means that the whole project has been delayed 4 units. Thus, the cost of the delay is 20.

In this paper, we seek to contribute to the literature on cost allocation by studying the share of the cost of the delay between activities. The subject was first studied by Bergantiños and Sánchez (2002a) and Brânzei et al. (2002). The former present two rules, one based on serial cost sharing problems and the other on game theory. They also introduce some desirable properties and study which of these are satisfied by the rules. The second study takes a different approach, assigning two related problems to each PERT situation: a bankruptcy problem and a cooperative game. The rules and values suggested in the respective fields (e.g. the proportional rule and the Shapley value) are then studied and applied to the original PERT situation. Furthermore, Bergantiños and Vidal-Puga (2009) use an indirect approach to determine a rule that distributes the cost of the delay equally among the activities. Estévez-Fernández et al. (2007); Castro et al. (2008); Curiel (2011); Estévez-Fernández (2012) associate different cooperative games with each problem and study their respective cores[2]. Following this idea, we do not assume that the cost of delay depend solely on the final completion time.

We study cost allocation rules, which determine the way the cost of the delay is shared between the activities. In the literature on cost allocation, one of the most important issues is to find allocation rules that are efficient (the total cost of the delay must be assigned), fair (based on the some general principle of justice), and strategically stable (no coalition of activities can improve by themselves). In PERT situations, however, strategic stability has no clear definition: the graph structure is closed and, as such, no group of activities can do without the others. This means that concepts such as the core and population monotonicity are not applicable. Thus, we focus on efficiency and fairness. We only consider a property of efficiency, which says that the cost assigned should exactly cover the delay cost. We consider several fairness properties: cost monotonicity (no activity should pay more if it finishes ahead of schedule), dummy (an activity finishing on schedule pays nothing), anonymity (the allocation does not depend on the name of the activities), and symmetry (symmetric activities pay the same). In general there are many rules to satisfy these properties.

As Winter (2002) argues, one of the fundamental requirements of any legal system is that it is internally consistent. In terms of the rules we would like to use for allocating the costs of delays, the allocation will typically depend on the PERT graph, scheduled completion times, actual completion times and (as argued above) the cost function.

Let us assume some activities get their payoffs according to an original PERT situation. We now focus our attention on the remaining activities (a group of activities B). Once the project is finished, we evaluate the cost allocation to these remaining activities. This evaluation gives rise to a new cost function for B and hence a different (reduced) PERT situation. The rule is said to

[2] From a different perspective, Bergantiños and Sánchez (2002b) and Castro et al. (2007) also use a cooperative game but not study the cost of delay, but the way to share slacks. Bergantiños and Lorenzo (2016) study this problem using non-cooperative games with the idea of avoiding delays.

be *consistent* if the payoff for the activities in the reduced PERT situation are the same as in the original PERT situation.

In this paper, we define a rule for PERT problems and show that it is consistent. This rule is defined as the Shapley value of an associated cooperative game. In this game the value of a coalition S of activities is defined as the cost of the delay of the project when each activity in S has its delay and each activity not in S is not delayed.

Consider a project with two activities. A rule is said to satisfy *equal responsibility for two* if two conditions hold. Firstly, each activity pays the cost caused by its own delay. Secondly, the rest of the cost (namely the cost of the delay of the project minus the sum of the cost caused by the delay of each activity) is divided equally between both. We characterize our rule as the unique one satisfying consistency and equal responsibility for two. We also prove that the rule satisfies other fairness properties like cost monotonicity, dummy, anonymity, and symmetry.

The paper is organized as follows: in Section **??**, we give the notation and define the model; in Section we study the Shapley rule and provide a characterization based on consistency and equal responsibility for two; in Section we discuss the independence of the properties; in Section **??** we give some additional examples.

2. THE PROBLEM

We denote the set of nonnegative real numbers as $\mathbb{R}_+$ and the set of natural numbers as $\mathbb{N}$. Given a finite set N, we denote the cardinal set of N as $|N|$. Given $x, y \in \mathbb{R}^N$, we write $x \leq y$ when $x_i \leq y_i$ for all $i \in N$. Given $x \in \mathbb{R}^N$ and $S \subset N$, we denote the restriction of x to S as x_S, *i.e.* $x_S = (x_i)_{i \in S}$. Given $x \in \mathbb{R}^N$, $S \subset N$ and $y \in \mathbb{R}^S$, x/y denotes the vector $z \in \mathbb{R}^N$ with $z_i = y_i$ if $i \in S$ and $z_i = x_i$ if $i \in \overline{S}$ where $\overline{S}$ denotes the set $N \setminus S$. For simplicity when $S = \{i\}$ and $y = (y_i) \in \mathbb{R}^S$, we write x/y_i instead of x/y. 0_N denotes the vector $(0, ..., 0) \in \mathbb{R}^N_+$.

Let Π_N be the set of all permutations over the finite set $N \subset \mathbb{N}$. Given $\pi \in \Pi_N$, let $Pre(i, \pi)$ denote the set of elements of N which come before i in the order given by π, *i.e.* $Pre(i, \pi) = \{j \in N \mid \pi(j) < \pi(i)\}$. Given $S \subset N$, let π_S denote the order induced by π on S.

A game with transferable utility (TU game) is a pair (N, v) where N is a finite set and $v : 2^N \to \mathbb{R}$ is the *characteristic function* satisfying $v(\emptyset) = 0$. Given $S \subset N$, we denote the *subgame* of (N, v) to S as (S, v'), where $v'(T) = v(T)$ for all $T \subset S$. We denote by $Sh(N, v)$ the *Shapley value* (Shapley, 1953) of the TU game (N, v). It is well known that for all $i \in N$,

$$Sh_i(N, v) = \frac{1}{|N|!} \sum_{\pi \in \Pi_N} \left(v(Pre(i, \pi) \cup \{i\}) - v(Pre(i, \pi)) \right).$$

Let F be a value for TU games. Given $T \subset N$, the Hart and Mas-Colell reduced game $\left(T, v^F_T\right)$ is defined as $v^F_T(S) = v(S \cup (N \backslash T)) - \sum_{i \in N \backslash T} F_i(S \cup (N \backslash T), v)$ for all $S \subseteq T$. A value F is consistent if, for every TU game (N, v) and every $T \subset N$, $F_i\left(T, v^F_T\right) = F_i(N, v)$ for all $i \in T$. A value F is standard for two person games if for each game (N, v) with $N = \{i, j\}$ we have that $F_i(N, v) = v(i) + \frac{1}{2}[v(N) - v(i) - v(j)]$.

The next result will be used in this paper.

Theorem 1 (Hart and Mas-Colell (1989) - Theorem B) *A value is consistent and standard for two person games if and only if is the Shapley value.*

A *PERT problem with delays* over N is a tuple $P = \left(N, \prec, x^0, x, C\right)$ where

- $N \subset \mathbb{N}$ denotes a finite set of activities forming the project.
- $\prec$ is a *partial order* in N, *i.e.* $\prec$ satisfies *transitivity* ($i \prec j$ and $j \prec k$ imply $i \prec k$) and *asymmetry* ($i \prec j$ implies $j \not\prec i$). Given $i, j \in N$, $i \prec j$ means that activity j cannot begin until activity i is finished.
- $x^0 \in \mathbb{R}_+^N$ and for each $i \in N$, x_i^0 is the *expected completion time* of activity i.
- $x \in \mathbb{R}_+^N$ and for each $i \in N$, x_i is the *actual completion time* of activity i. We assume that $x \geq x^0$.
- $C : \mathbb{R}_+^N \to \mathbb{R}$ is the *cost function*. We assume that $C\left(x^0\right) = 0$. Given $x \in \mathbb{R}_+^N$, $C\left(x\right)$ denotes the cost associated with the delay in the project.

A *path* $\gamma = \{i_1, \ldots, i_p\} \subset N$ in P is a collection of activities such that $i_k \prec i_{k+1}$ for each $k \in \{1, \ldots, p-1\}$. Let Γ^P be the set of all paths in P. An *expected critical path* in P is a path $\gamma^0 \in \Gamma^P$ such that $\sum_{i \in \gamma^0} x_i^0 = \max_{\gamma \in \Gamma^P} \sum_{i \in \gamma} x_i^0$. An *actual critical path* in P is a path $\gamma^1 \in \Gamma^P$ such that $\sum_{i \in \gamma^1} x_i = \max_{\gamma \in \Gamma^P} \sum_{i \in \gamma} x_i$.

We define the *PERT delay* of the project P as $d(P) = \max_{\gamma \in \Gamma^P} \sum_{i \in \gamma} x_i - \max_{\gamma \in \Gamma^P} \sum_{i \in \gamma} x_i^0$. We define the *delay* of activity $i \in S$, d_i, as the difference between its actual completion time and its expected completion time, *i.e*, $d_i = x_i - x_i^0 \geq 0$.

Typically, the cost function is nondecreasing in each coordinate and it only depends on the PERT delay of the project i.e. there exists $C : \mathbb{R}_+ \to \mathbb{R}$ nondecreasing with $C(0) = 0$ and such that $C(y) = C\left(d\left(N, \prec, x^0, y, C\right)\right)$ for all $y \geq x^0$.

Example 1 *Assume that we have to carry out a project with three activities. The PERT situation is given by* $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$ *where* $N = \{1, 2, 3\}$, $1 \prec 2$ *and* $1 \prec 3$, $x^0 = (10, 20, 30)$, $x = (20, 40, 40)$, *and* $C(x_1, x_2, x_3) = x_1 + \max\{x_2, x_3\} - x_P^0$, *where* $x_P^0 = x_1^0 + \max\{x_2^0, x_3^0\} = 40$. *This situation is depicted in Figure 2.*

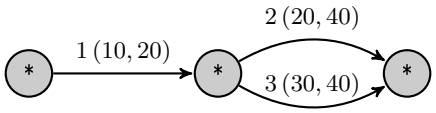


Figura 2: Graph of PERT problem with delays.

We note that there is a unique expected critical path $\gamma^0 = \{1, 3\}$ *and there are two actual critical paths* $\gamma_1^1 = \{1, 2\}$ *and* $\gamma_2^1 = \{1, 3\}$. *We obtain the PERT delay of the project as* $d(P) = 60 - 40 = 20$. *Note that there exists a function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *such that* $f(t) = t$ *for all* $t \in \mathbb{R}_+$ *which generates the cost* $C(x) = 20 + \max\{40, 40\} - 40 = 20$ *and the delay of each activity is given by* $d = (10, 20, 10)$. *Also, note that activity* 2 *can be delayed by up to* 10 *units without generating additional delay cost in the project.*

Given $S \subset N$, we denote $\prec_S$ the restriction of $\prec$ on S, *i.e.* $i \prec_S j$ iff $i \prec j$, for all $i, j \in S$.

We denote the set of PERT problems with delays over a finite set N as $\mathcal{P}\left(N\right)$. We also denote the set of all PERT problems with delays as $\mathcal{P}$, *i.e.* $\mathcal{P} = \bigcup_{N \subset \mathbb{N}, |N| < \infty} \mathcal{P}\left(N\right)$. Let $C\left(x\right)$ be the total cost produced by the delay of the project.

A *rule* is a function ϕ on $\mathcal{P}$ that assigns to each $P \in \mathcal{P}(N)$ a vector $\phi(P) \in \mathbb{R}_+^N$ satisfying:

Efficiency (EF) $\sum_{i \in N} \phi_i(P) = C(x)$ and

Null Delay (ND) $\phi_i(P) = 0$ when $x_i = x_i^0$.

Given an activity $i \in N$, $\phi_i(P)$ represents the amount paid by activity i. Efficiency says that the total cost must be distributed among the activities involved in the project. Null delay says that activity i must pay nothing when it is not delayed.

As a simple example of rule, consider the *proportional rule* (Brânzei et al., 2002), which is defined as follows: Given $P = (N, \prec, x^0, x, C) \in \mathcal{P}$,

$$PRO_i(P) = \frac{x_i - x_i^0}{\sum_{j \in N}(x_j - x_j^0)} C(x)$$

for all $i \in N$.

Example 2 *Consider a problem with two activities, say* 1 *and* 2*, which can be done simultaneously. The expected completion time of the activities are* 20 *and* 4*, respectively. The total expected completion time of the project is* 20*. We must pay* 100 *for each unit of time the project is delayed. Assume that activity* 1 *takes* 21 *units of time and activity* 2 *takes* 13 *units of time. The delay of the project is* 1*, which means that the cost is* 100*. Notice that the project is delayed only because activity* 1 *is delayed. We believe that all the cost should be paid by activity* 1*. This situation can be modeled as a parallel PERT problem with delays* $P = (N, \prec, x^0, x, C)$ *where* $N = \{1, 2\}$*,* $\prec = \emptyset$*,* $x^0 = (20, 4)$*,* $x = (21, 13)$*, and* $C(x) = 100(\max\{x_1, x_2\} - 20)$*. Thus,* $PRO(P) = (10, 90)$*, which seems rather unfair.*

We now define two extreme subclasses of PERT problems with delays. In the first all the activities are related whereas in the second activities are unrelated.

We say that P is *sequential* if activities are lined up, *i.e.* $N = \{i_k\}_{k=1}^n$ where $i_k \prec i_{k+1}$ for all $k \in \{1, \ldots, n-1\}$.

We say that P is *parallel* if activities can be done simultaneously, *i.e.* for all $i, j \in N$, $i \not\prec j$.

The previous PERT problems with delays generalize two well-known problems of the literature of cost sharing: the serial cost sharing problem and the airport problem.

Following Moulin and Shenker (1992) a *serial cost sharing problem* is a triple (N, q, c) where N is the set of agents, $q = (q_i)_{i \in N} \in \mathbb{R}_+^N$ and $c : \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing function satisfying $c(0) = 0$.

This model has several interpretations. We present the interpretation based on cost sharing. Each agent $i \in N$ demands an amount $q_i \in \mathbb{R}_+$ of a perfectly divisible good. The total amount required by the agents, $\sum_{i \in N} q_i$, is produced. The cost $c\left(\sum_{i \in N} q_i\right)$ is divided among the agents.

Consider a PERT problem with delays where activities are lined up and the cost of delaying the project depends only on the total delay of the project (the difference between the actual completion time and the expected completion time). Formally, $P = (N, \prec, x^0, x, C)$ where P is sequential and there exists a nondecreasing function $c : \mathbb{R}_+ \to \mathbb{R}_+$ such that $C(y) = c\left(\sum_{i \in N}(y_i - x_i^0)\right)$ for all $y \geq x^0$.

This PERT problem with delays can be considered as a serial cost sharing problem where the set of agents are the activities and the demands are the delays.

Conversely, given a serial cost sharing problem (N, q, c), we can associate with it a sequential PERT situation with delays $\left(N, \prec, x^0, x, C\right)$ where the set of activities is N either $i \prec j$ or $j \prec i$ for all $i, j \in N$, and $y = x^0 + q$ for some $x^0 \in \mathbb{R}^N$, and $C(y) = c\left(\sum_{i \in N}(y_i - x_i^0)\right)$ for all $y \geq x^0$.

Thus, serial cost sharing problems can be considered as a subclass of sequential PERT situations with delay. When we identify a serial cost sharing problem with a PERT problem with delays, we are thinking, in both cases of how to divide the cost in a fair way among the agents or activities.

Following Littlechild and Owen (1973), an *airport problem* is a pair (N, q, c) where N is the set of agents, $q \in \mathbb{R}_+^N$, and $c : \mathbb{R}_+ \to \mathbb{R}_+$ is a nondecreasing function satisfying that $c(0) = 0$. Even though the mathematical model is the same as in serial cost sharing problems, the situation is different. N represents several kinds of aircraft landings in some airport. Each aircraft $i \in N$ needs, for landing, a runway of length at least q_i. We have to construct a runway of length at least $\max_{i \in N}\{q_i\}$. When an aircraft lands in this airport, it must pay a tariff, which depends, among other things, of the construction cost of the runway. The problem is to determine this part of the tariff in a fair way.

Consider a PERT problem with delays where activities are parallel and the cost of delaying the project depends only on the total delay of the project. Formally, $P = \left(N, \prec, x^0, x, C\right)$ where P is parallel and there exists a nondecreasing function $c : \mathbb{R}_+ \to \mathbb{R}_+$ such that $C(y) = c\left(\max_{i \in N}\left\{y_i - x_i^0\right\}\right)$ for all $y \geq x^0$.

This PERT problem with delays can be considered as an airport problem where the agents are the activities and the demand of each agent is its delay.

Conversely, given an airport problem (N, q, c) we can associate with it a parallel PERT problem with delays $\left(N, \prec, x^0, x, C\right)$ where P is parallel, $x - x^0 = q$ for some $x^0 \in \mathbb{R}^N$, and $C(y) = c\left(\max_{i \in N}\left\{y_i - x_i^0\right\}\right)$ for all $y \geq x^0$.

Thus, airport problems can be considered as a subclass of PERT problems with delays.

3. THE SHAPLEY RULE

The Shapley value (Shapley, 1953) is defined for general cooperative games, but it has been successfully applied to many particular situations, such as voting games (Shapley and Shubik, 1954), serial cost problems (Moulin and Shenker, 1992) and airport landing problems (Littlechild and Owen, 1973), just to mention a few.

We now introduce a rule for PERT problems with delays based on the Shapley value.

Given a PERT problem with delays $P = \left(N, \prec, x^0, x, C\right)$ we associate a TU game $\left(N, v^P\right)$ where $v^P(S) = C\left(x^0/x_S\right)$ for all $S \subset N$. Notice that $v^P(S)$, represents the cost of delaying the project when activities of $\overline{S}$ are not delayed and activities of S are delayed by $x - x^0$.

Given $P \in \mathcal{P}$ we define the *Shapley rule* $\overline{Sh}$ as

$$\overline{Sh}(P) = Sh\left(N, v^P\right).$$

It is straightforward to prove that $\overline{Sh}$ is a well-defined rule, i.e. it satisfies EF and ND.

We now introduce some other properties of rules. The first five properties are quite standard in the literature. The last property is specific to PERT problems with delays. Let ϕ be a generic notation for our rules.

Given $\pi \in \Pi_N$ and $y \in \mathbb{R}_+^N$ we define $\pi(y) \in \mathbb{R}_+^N$ as the vector where $\pi(y)_i = y_{\pi^{-1}(i)}$ for all $i \in N$. Given $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$ and $\pi \in \Pi_N$, we define the PERT problem with delays

$P^\pi = \left(N, \prec_\pi, \pi\left(x^0\right), \pi\left(x\right), C^\pi\right)$ where $i \prec_\pi j$ iff $\pi\left(i\right) \prec \pi\left(j\right)$ for all $i, j \in N$, and $C^\pi\left(y\right) = C\left(\pi^{-1}\left(y\right)\right)$ for all $y \in \mathbb{R}_+^N$.

Anonymity (AN) For all $P \in \mathcal{P}$, $\pi \in \Pi_N$, and $i \in N$, $\phi_i\left(P\right) = \phi_{\pi(i)}\left(P^\pi\right)$.

Anonymity says that cost shares should not depend on the name of the activities.

Cost Additivity $(CADD)$ For all pair of cost functions C and C' with $C(x^0) = C'(x^0)$,

$$\phi\left(N, \prec, x^0, x, C + C'\right) = \phi\left(N, \prec, x^0, x, C\right) + \phi\left(N, \prec, x^0, x, C'\right)$$

where $\left(C + C'\right)\left(y\right) = C\left(y\right) + C'\left(y\right)$ for all $y \geq x^0$.

In order to illustrate this property, we consider a house building project. Since there exists a total delay cost of the project, we can distinguish between two cost functions that are generated by two different reasons. One of them is a variable cost C (for example, the rent payment in another house) and the other one is a fixed cost C' (for example, the payment of a contractual penalty). Assume that activities in N agree that ϕ is the right solution. They can proceed in two ways. First, they can apply ϕ in each of the two PERT problems with delays given by C and C'. Second, they can apply ϕ only once in the PERT problem with delays with the cost function given by $C + C'$. If both procedures give the same result, we say that ϕ is additive.

Given $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$, $S \subset N$ and $y \in \mathbb{R}^S$ we define the *reduced problem* associated with S and the rule ϕ as $P^{\phi,S} = \left(S, \prec_S, x_S^0, x_S, C^{\phi,S}\right)$ where $C^{\phi,S} : \mathbb{R}_+^S \to \mathbb{R}_+$ is defined as

$$C^{\phi,S}\left(y\right) = C\left(x/y\right) - \sum_{i \in \overline{S}} \phi_i\left(N, \prec, x^0, x/y, C\right).$$

Consistency $(CONS)$ For all $S \subset N$ we have $\phi_S\left(P\right) = \phi\left(P^{\phi,S}\right)$.

Given $S \subset N$, assume that activities in $\overline{S}$ agree that ϕ is the right solution. Activities in S can proceed in two ways. First, they can divide the cost according with ϕ. Second, they can construct a reduced problem among themselves assuming that activities in $\overline{S}$ will pay according with ϕ. If both procedures give the same result for activities in S, we say that ϕ is consistent.

Consistency is a well-known principle in the literature. See, for instance, Thomson (2011, 2012). In particular, we use the consistency notion first defined by Hart and Mas-Colell (1989).

Monotonicity (MON) For all $P = \left(N, \prec, x^0, x, C\right)$, with C nondecreasing in each coordinate, and $P' = \left(N, \prec, x^0, x/x_i', C\right)$ with $x_i \leq x_i'$ for some $i \in N$, $\phi_i\left(P\right) \leq \phi_i\left(P'\right)$.

If the delay of activity i increases and delay is always costly, the amount paid by this activity cannot decrease.

Equal Responsibility for Two (ERT) Given $P = \left(N, \prec, x^0, x, C\right)$ where N has two activities, say i and j, $\phi_i\left(P\right) = C\left(\left(x_i, x_j^0\right)\right) + \frac{1}{2}\left[C\left(\left(x_i, x_j\right)\right) - C\left(\left(x_i, x_j^0\right)\right) - C\left(\left(x_i^0, x_j\right)\right)\right]$.

Each activity pays the cost caused by its own delay (assuming that the other is not delayed). Moreover, the difference between the total cost of delaying the project and the sum of the cost caused by both activities, is divided equally among them.

ERT is inspired by the property of standard for two of *TU* games defined in Hart and Mas-Colell (1989).

Given $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$ and $\lambda \in \mathbb{R}^N_{++}$, we define the cost function $C^\lambda : \mathbb{R}^N_+ \to \mathbb{R}_+$ such that $C^\lambda(\lambda y) = C(y)$ for all $y \in \mathbb{R}^N_+$, where $\lambda y = (\lambda_i y_i)_{i \in N}$.

Scale Invariance (*SI*) For all $P = \left(N, \prec, x^0, x, C\right)$ and $\lambda \in \mathbb{R}^N_{++}$ we have that $\phi\left(N, \prec, \lambda x^0, \lambda x, C^\lambda\right) = \phi\left(N, \prec, x^0, x, C\right)$.

A change of scale does not affect the cost allocation.

We say that $x_i \in \mathbb{R}_+$ constitutes an *irrelevant completion time for activity* $i \in N$ for $\left(N, \prec, x^0, x, C\right)$ if for all $x' \in \mathbb{R}^N_+$ with $x' \geq x^0$, $C\left(x'/x_i\right) = C\left(x'/x^0_i\right)$.

This means that, independently of the delays of the other activities, if activity i has a delay $x_i - x^0_i$, the total cost of the project is the same as if activity i is not delayed. Otherwise, we say that x_i is a *relevant completion time* for activity i.

Independence of Irrelevant Delays (*IID*) Given $P = \left(N, \prec, x^0, x, C\right)$, $i \in N$, and x_i is an irrelevant completion time for activity i, $\phi_i(P) = 0$.

If the delay of activity i does not affect the total cost of the project, activity i pays nothing.

Example 3 *We have to carry out a project of two activities (*1 *and* 2*), which can be performed independently. The expected completion time of activities* 1 *and* 2 *are* 10 *and* 5*, respectively. Thus, the total expected time for finishing the project is* 10*. Moreover, we have to pay* 100 *for each unit of time that the project is delayed. In this example,* $N = \{1, 2\}$*,* $\prec$ *is empty,* $x^0 = (10, 5)$*, and* $C(x) = 100\left(\max\{x_1, x_2\} - 10\right)$*. It is not difficult to check that* x_2 *is an irrelevant completion time for activity* 2 *when* $x_2 \leq 10$*. Assume that activities* 1 *and* 2 *have a delay of* 2 *and* 4*, respectively. The cost of delaying the project is* 200*. Thus, if a rule satisfies IID, all the cost should be paid by activity* 1*.*

Some rules, such as the proportional rule *PRO*, do not satisfy *IID*. Even though *IID* has the same flavor as *ND*, they are different properties. *IID* implies *ND* but the converse is not true.

We now prove that the Shapley rule satisfies all the properties mentioned before.

Theorem 2 *The Shapley rule satisfies anonymity, cost additivity, consistency, monotonicity, scale invariance, equal responsibility for two, and independence of irrelevant delays.*

Proof. Anonymity: Let $P = \left(N, \prec, x^0, x, C\right)$ a PERT problem with delays and $\pi \in \Pi_N$. For each $S \subseteq N$, we define $\pi(S) = \{\pi(i) : i \in S\}$. For each *TU* game (N, v), the *TU* game $(N, \pi v)$ is defined as $\pi v(S) = v(\pi^{-1}(S))$ for all $S \subseteq N$. For *TU* games, the Shapley value satisfies anonymity, i.e. $Sh_{\pi(i)}(N, \pi v) = Sh_i(N, v)$ for all (N, v) and all $i \in N$ (Shapley, 1953). We will prove that $\overline{Sh}_i(P) = \overline{Sh}_{\pi(i)}(P^\pi)$ for all $i \in N$. Since $\overline{Sh}_i(P) = Sh_i\left(N, v^P\right) = Sh_{\pi(i)}\left(N, \pi v^P\right)$ and $\overline{Sh}_{\pi(i)}(P^\pi) = Sh_{\pi(i)}\left(N, v^{P^\pi}\right)$ for all $i \in N$, it is enough to check that $\pi v^P = v^{P^\pi}$. For all $S \subseteq N$, on the one hand, $\pi v^P(S) = v^P\left(\pi^{-1}(S)\right) = C\left(x^0/x_{\pi^{-1}(S)}\right)$, and, on the other hand, $v^{P^\pi}(S) = C^\pi\left(\pi(x^0)/\pi(x)_S\right) = C\left(\pi^{-1}\left(\pi(x^0)/\pi(x)_S\right)\right)$. Hence, it is enough to check that

$$x^0/x_{\pi^{-1}(S)} = \pi^{-1}\left(\pi(x^0)/\pi(x)_S\right). \tag{1}$$

Given $i \in N$, we have two cases: If $i \in \pi^{-1}(S)$, i.e. $\pi(i) \in S$, then $\left(x^0/x_{\pi^{-1}(S)}\right)_i = x_i$ and $\left(\pi^{-1}\left(\pi(x^0)/\pi(x)_S\right)\right)_i = \left(\pi(x^0)/\pi(x)_S\right)_{\pi(i)} = \pi(x)_{\pi(i)} = x_{\pi^{-1}(\pi(i))} = x_i$. If $i \notin \pi^{-1}(S)$, i.e. $\pi(i) \notin S$, then $\left(x^0/x_{\pi^{-1}(S)}\right)_i = x_i^0$ and $\left(\pi^{-1}\left(\pi(x^0)/\pi(x)_S\right)\right)_i = \left(\pi(x^0)/\pi(x)_S\right)_{\pi(i)} = \pi(x^0)_{\pi(i)} = x^0_{\pi^{-1}(\pi(i))} = x_i^0$. Then, the equality in (1) holds and $\overline{Sh}$ satisfies AN.

Cost Additivity: Let C and C' be cost functions such that $(C + C')(y) = C(y) + C'(y)$ for all $y \geq y^0$. Let $P' = \left(N, \prec, x^0, x, C'\right) \in \mathcal{P}$ and $P'' = \left(N, \prec, x^0, x, C + C'\right) \in \mathcal{P}$. We prove that $\overline{Sh}(P'') = \overline{Sh}(P) + \overline{Sh}(P')$. For $i \in N$, $\overline{Sh}_i\left(P''\right) = Sh_i\left(N, v^{P''}\right)$. Since $(C+C')(y) = C(y)+C'(y)$ for all $y \geq y^0$, $v^{P''} = v^P + v^{P'}$. Thus, $Sh_i\left(N, v^{P''}\right) = Sh_i(N, v^P) + Sh(N, v^{P'}) = \overline{Sh}_i(P) + \overline{Sh}_i(P')$.

Consistency: Given $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$ and $S \subset N$, we should prove that $\overline{Sh}_S\left(P\right) = \overline{Sh}\left(P^{\overline{Sh},S}\right)$ or, equivalently, $Sh_S\left(N, v^P\right) = Sh\left(S, v^{P^{\overline{Sh},S}}\right)$.

Hart and Mas-Colell (1989) proved that if (N, v) is a cooperative game and $S \subset N$, then $Sh_S\left(N, v\right) = Sh\left(S, v_S\right)$, where for all $T \subset S$, $v_S\left(T\right) = v\left(T \cup \overline{S}\right) - \sum_{i \in \overline{S}} Sh_i\left(T \cup \overline{S}, v\right)$.

Hence, it is enough to prove that $\left(v^P\right)_S = v^{P^{\overline{Sh},S}}$. Since Sh satisfies efficiency for TU games, for all $T \subset S$, $\left(v^P\right)_S\left(T\right) = \sum_{i \in T} Sh_i\left(T \cup \overline{S}, v^P\right)$.

Since $\overline{Sh}$ satisfies EF, ND and SI,

$$\begin{aligned} v^{P^{\overline{Sh},S}}\left(T\right) = C^{\overline{Sh},S}\left(x_S^0/x_T\right) &= \sum_{i \in S} \overline{Sh}_i\left(N, \prec, x^0, x/\left(x_S^0/x_T\right), C\right) \\ &= \sum_{i \in S} \overline{Sh}_i\left(N, \prec, x^0, x/x^0_{S\setminus T}, C\right) = \sum_{i \in T} \overline{Sh}_i\left(N, \prec, x^0, x/x^0_{S\setminus T}, C\right). \end{aligned}$$

We denote $P' = \left(N, \prec, x^0, x/x^0_{S\setminus T}, C\right)$. Thus, $v^{P^{\overline{Sh},S}}\left(T\right) = \sum_{i \in T} Sh_i\left(N, v^{P'}\right)$.

Now, it is enough to prove that $Sh_T\left(N, v^{P'}\right) = Sh_T\left(T \cup \overline{S}, v^P\right)$. We prove it in two steps.

1. We first prove that $Sh_T\left(N, v^{P'}\right) = Sh_T\left(T \cup \overline{S}, v^{P'}\right)$. By the carrier property[3] of the Shapley value (Shapley, 1953), it is enough to prove that $T \cup \overline{S}$ is a carrier for $\left(N, v^{P'}\right)$, *i.e.* $v^{P'}\left(R\right) = v^{P'}\left(R \cap \left(T \cup \overline{S}\right)\right)$ for all $R \subset N$. Given $R \subset N$,

$$v^{P'}\left(R\right) = C\left(x^0/\left(x/x^0_{S\setminus T}\right)_R\right) = C\left(x^0/x_{R \cap \overline{(S\setminus T)}}\right)$$

and $v^{P'}\left(R \cap \left(T \cup \overline{S}\right)\right) = C\left(x^0/\left(x/x^0_{S\setminus T}\right)_{R \cap \left(T \cup \overline{S}\right)}\right) = C\left(x^0/x_{R \cap \left(T \cup \overline{S}\right)}\right)$.

Since $\overline{(S \setminus T)} = T \cup \overline{S}$, the result holds.

2. We now prove that $Sh_T\left(T \cup \overline{S}, v^{P'}\right) = Sh_T\left(T \cup \overline{S}, v^P\right)$. We prove that $v^{P'} = v^P$. Given $R \subset T \cup \overline{S}$, $v^{P'}\left(R\right) = C\left(x^0/\left(x/x^0_{S\setminus T}\right)_R\right) = C\left(x^0/x_R\right) = v^P\left(R\right)$.

Monotonicity: Let $P = \left(N, \prec, x^0, x, C\right)$, with C increasing in each coordinate, $i \in N$, and $P' = \left(N, \prec, x^0, x/x'_i, C\right)$ with $x_i \leq x'_i$. Given $S \subset N$, $i \notin S$,

$$\begin{aligned} v^P\left(S \cup \{i\}\right) - v^P\left(S\right) &= C\left(x^0/x_{S\cup\{i\}}\right) - C\left(x^0/x_S\right) \\ &\leq C\left(x^0/\left(x_S, x'_i\right)\right) - C\left(x^0/x_S\right) \\ &= v^{P'}\left(S \cup \{i\}\right) - v^{P'}\left(S\right). \end{aligned}$$

Since $\overline{Sh}\left(P\right) = Sh\left(N, v^P\right)$ and $\overline{Sh}\left(P'\right) = Sh\left(N, v^{P'}\right)$, $\overline{Sh}$ satisfies MON.

[3] A carrier of v (a TU game) is any set $N \subseteq \mathbb{N}$ with $v(S) = v(N \cap S)$ for all $S \subseteq \mathbb{N}$.

Scale Invariance: It is trivial to prove that $\overline{Sh}$ satisfies SI.

Equal Responsibility for Two: Since $\overline{Sh}(P)$ is the Shapley value of v^P, it is trivial to see that $\overline{Sh}$ satisfies ERT.

Independent of Irrelevant Delays: Given $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$, let x_i be an irrelevant completion time for activity i. Assume that $S \subset N$ and $i \notin S$. Then,

$$v^P(S \cup \{i\}) = C\left(x^0/x_{S\cup\{i\}}\right) = C\left(x^0/x_S\right) = v^P(S).$$

Since $\overline{Sh}(P) = Sh\left(N, v^P\right)$, $\overline{Sh}$ satisfies IID. ■

Theorem 3 *There exists a unique rule satisfying consistency and equal responsibility for two.*

The proof consists in showing the equivalence of PERT problems and the class of TU games, together with the equivalence of our axioms for PERT problems we impose in our theorem with axioms for TU games on which an earlier characterization of the Shapley value (Hart and Mas-Colell, 1989) had been based.

Proof. Let ϕ be a rule satisfying these properties. Fix a PERT problem with delays $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$. Under ND and $CONS$, we can assume that $x_i > x_i^0$ for all $i \in N$. To see why, notice that ND implies $\phi_i(P) = 0$ for all $i \in N$ with $x_i = x_i^0$. Let $S = \left\{i \in N : x_i > x_i^0\right\}$. By $CONS$, $\phi_S(P) = \phi\left(P^{\phi,S}\right)$ and hence we can work with $P^{\phi,S}$.

We define the partial order $\prec^*$ on $\mathbb{N}$ by setting $i \prec^* j$ if and only if $i, j \in N$ and $i \prec j$. Let $\mathcal{P}'$ be the class of PERT problems with delays defined as follows: $P' = \left(N', \prec', x^{0\prime}, x', C'\right) \in \mathcal{P}'$ if and only if $\prec'=\prec^*$, $x_i^{0\prime} = x_i^0$ for all $i \in N \cap N'$, and $x_i^{0\prime} = 0$ otherwise.

Given $P' = \left(N', \prec', x^{0\prime}, x', C'\right) \in \mathcal{P}'$ we can associate the TU game $\left(N, v^{P'}\right)$ as above.

Conversely, given a TU game (N', v'), we can associate a PERT problem with delays $P' = \left(N', \prec', x^{0\prime}, x', C'\right) \in \mathcal{P}'$ as follows: $x_i' = x_i$ for all $i \in N \cap N'$ and $x_i' = 0$ otherwise. C' satisfies $C'\left(x^{0\prime}/x_S\right) = v'(S)$.

We define a value F for TU games through ϕ. For each TU game (N', v'), $F(N', v') = \phi\left(N', \prec', x^{0\prime}, x', C'\right)$.

Consider a TU game (N', v'), where $|N'| = 2$. Since ϕ satisfies ERT, $F_i(N', v') = C'(x_i', x_j^{0\prime}) + \frac{1}{2}\left[C'(x') - C'(x_i', x_j^{0\prime}) - C'(x_i^{0\prime}, x_j')\right]$ for all $i \in N'$.

Since $C'(x_i', x_j^{0\prime}) = v'(\{i\})$, $C'(x') = v'(\{i,j\})$ and $C'(x_i^{0\prime}, x_j') = v'(\{j\})$, we deduce that $F(N', v') = Sh(N', v')$, Thus F coincides with the Shapley value for TU games with two agents.

Since ϕ satisfies consistency in PERT problems with delays, it is not difficult to check that F is consistent (as defined in Hart and Mas-Colell (1989)) for TU games.

By Theorem 1, F coincides with the Shapley value. We now consider the TU game (N, v^P). It is not difficult to check that P is the PERT problem with delays $(v^P)'$ associated with (N, v^P) as above. Hence, $\phi(P) = F(N, v^P)$. Since $F(N, v^P) = Sh(N, v^P)$, we conclude $\phi(P) = Sh(N, v^P) = \overline{Sh}(P)$. ■

3. INDEPENDENCE OF THE PROPERTIES

We have characterized the Shapley rule with $CONS$ and ERT. Besides, this characterization is tight. If we remove one of the properties we have more rules. The proportional rule PRO satisfies CONS, but fails ERT. Let $R^P = \left\{i \in N : \exists x' \in \mathbb{R}_+^N, x' \geq x^0, C(x'/x_i) \neq C(x'/x_i^0)\right\}$ be the set of activities with a relevant completion time. Let δ be the rule that divides the cost applying ERT

when there are two activities, and with more activities it divides the cost proportionally to the delay among those activities in R^P. Formally, given $P = \left(N, \prec, x^0, x, C\right) \in \mathcal{P}$, we define

$$\delta_i(P) = \begin{cases} \overline{Sh}_i(P), & \text{if } |N| = 2 \\ \frac{x_i - x_i^0}{\sum_{j \in R^P}(x_j - x_j^0)} C(x), & \text{if } i \in R^P, |N| \neq 2 \\ 0, & \text{otherwise} \end{cases}$$

for all $i \in N$. This rule satisfies ERT but fails $CONS$.

Table 1 shows a summary of these results.

Property	$\overline{Sh}$	PRO	δ
Anonymity	Y	Y	Y
Cost additivity	Y	Y	Y
Consistency	Y	Y	N
Monotonicity	Y	Y	Y
Scale invariance	Y	Y	Y
Equal Responsibility for Two	Y	N	Y
Independence of Irrelevant Delays	Y	N	Y

Táboa 1: Properties of four rules in PERT problems with delays.

We prove the results in Table 1 that are not related to $\overline{Sh}$.

Example 2 shows that PRO does not satisfy ERT nor IID. Furthermore, the proofs of AN, MON and SI for PRO and δ are analogous to the proof of Theorem 2 for $\overline{Sh}$. It is also straightforward to check that δ satisfies IID. Moreover, δ does not satisfy $CONS$ because it is different from $\overline{Sh}$. For instance, in Example 1, $\delta(P) = (5, 10, 5) \neq (10, 5, 5) = \overline{Sh}(P)$.

We now prove that PRO satisfies $CONS$. Let $P = (N, \prec, x^0, x, C)$ and $S \subset N$. We will prove that $PRO_S(P) = PRO\left(P^{PRO,S}\right)$, where $P^{PRO,S} = \left(S, \prec_S, x_S^0, x_S, C^{PRO,S}\right)$ with

$$C^{PRO,S}(y) = C(x/y) - \frac{\sum_{j \in \overline{S}}\left(x_j - x_j^0\right)}{\sum_{j \in S}\left(y_j - x_j^0\right) + \sum_{j \in \overline{S}}\left(x_j - x_j^0\right)} C(x/y).$$

Fix $i \in S$. Then,

$$\begin{aligned} PRO_i\left(P^{PRO,S}\right) &= \frac{x_i - x_i^0}{\sum_{j \in S}\left(x_j - x_j^0\right)} C^{PRO,S}\left(x_S\right) \\ &= \frac{x_i - x_i^0}{\sum_{j \in S}\left(x_j - x_j^0\right)} \left[C(x) - \sum_{j \in \overline{S}} \frac{x_j - x_j^0}{\sum_{j \in N}\left(x_j - x_j^0\right)} C(x)\right] \\ &= \frac{x_i - x_i^0}{\sum_{j \in S}\left(x_j - x_j^0\right)} \left[\frac{\sum_{j \in N}\left(x_j - x_j^0\right) - \sum_{j \in \overline{S}}\left(x_j - x_j^0\right)}{\sum_{j \in N}\left(x_j - x_j^0\right)}\right] C(x) \\ &= \frac{x_i - x_i^0}{\sum_{j \in N}\left(x_j - x_j^0\right)} C(x) = PRO_i(P). \end{aligned}$$

References

Bergantiños, G. and Lorenzo, L. (2016). How to apply penalties to avoid delays in projects. Mimeo, Universidade de Vigo.

Bergantiños, G. and Sánchez, E. (2002a). How to distribute the costs associated with a delayed project. *Annals of Operations Research*, 109(1):159–174.

Bergantiños, G. and Sánchez, E. (2002b). NTU PERT games. *Operations Research Letters*, 30(2):130–140.

Bergantiños, G. and Vidal-Puga, J. (2009). A value for PERT problems. *International Game Theory Review*, 11(04):419–436.

Brânzei, R., Ferrari, G., Fragnelli, V., and Tijs, S. (2002). Two approaches to the problem of sharing delay costs in joint projects. *Annals of Operations Research*, 109(1-4):359–374.

Castro, J., Gómez, D., and Tejada, J. (2007). A project game for PERT networks. *Operations Research Letters*, 35(6):791–798.

Castro, J., Gómez, D., and Tejada, J. (2008). A polynomial rule for the problem of sharing delay costs in PERT networks. *Computers & Operations Research*, 35(7):2376 – 2387.

Curiel, I. (2011). Project management games. *International Game Theory Review*, 13(3):281–300.

Estévez-Fernández, A. (2012). A game theoretical approach to sharing penalties and rewards in projects. *European Journal of Operational Research*, 216(3):647–657.

Estévez-Fernández, A., Borm, P., and Hamers, H. (2007). Project games. *International Journal of Game Theory*, 36(2):149–176.

Friedman, E. and Moulin, H. (1999). Three methods to share joint costs or surplus. *Journal of Economic Theory*, 87(2):275 – 312.

Hart, S. and Mas-Colell, A. (1989). Potential, value, and consistency. *Econometrica*, 57(3):pp. 589–614.

Littlechild, S. and Owen, G. (1973). A simple expression for the Shapley value in a special case. *Management Science*, 20(3):370–372.

Moulin, H. and Shenker, S. (1992). Serial cost sharing. *Econometrica*, 60(5):pp. 1009–1037.

Shapley, L. (1953). 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*, pages 307–317. Princeton University Press, Princeton NJ.

Shapley, L. and Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *American Political Science Review*, 48(3):787–792.

Thomson, W. (2011). Consistency and its converse: an introduction. *Review of Economic Design*, 15(4):257–291.

Thomson, W. (2012). On the axiomatics of resource allocation: Interpreting the consistency principle. *Economics and Philosophy*, 28(3):385–421.

Winter, E. (2002). The Shapley value. In Aumann, R. and Hart, S., editors, *Handbook of game theory with economic applications*, volume III of *Handbooks in Economics*. Elsevier Science, Amsterdam.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

USING FRACTIONAL REGRESSION MODELS TO ANALYSE EFFICIENCY OF EUROPEAN MANUFACTURING FIRMS

Bhaswar Chakma [1], Joaquim J.S. Ramalho [2], and Jorge M. Azevedo Santos[3]

[1]Independent Researcher, bhaswarbd@yahoo.com
[2]Department of Economics, ISCTE-IUL Instituto Universitário de Lisboa, jjsro@iscte-iul.pt
[3]Departamento de Matemática, CIMA, ECT, Universidade de Évora, jmas@uevora.pt

ABSTRACT

Fractional regression models are effective when the variable of interest appears in the form of proportion or fraction or rate. In other words, use of fractional regression models becomes appropriate when the dependent variable is defined only on the standard unit interval. This paper surveys fractional regression models, and studies efficiency scores of manufacturing firms of Portugal, Greece, France, and Germany by using data from the years 2007 and 2013. The study consists of two stages. The first stage uses data envelopment analysis (DEA) to obtain efficiency scores. Then the second stage employs fractional regression models to analyse those scores. The main aims of the second study are to find how financial crisis, use of technology (High-Tech or Low-Tech), firm size (SME or large), and country of origin affect efficiency of firms. Results provide evidence that financial crisis does not affect efficiency of firms, use of technology does not play any role in explaining efficiency of firms, large firms more efficient than SMEs, and firms in Greece are comparatively less efficient.

Keywords Fractional regression models; Beta Regression; Financial crisis; Second-stage DEA

1. INTRODUCTION

Fractional regression is comparatively a new tool in econometrics. Econometricians employ fractional regression models when the variable of interest appears in the form of proportion or fraction or rate. More precisely, use of fractional regression models becomes germane when the dependent variable ranges between 0 and 1. Possible intervals for such dependent variable are [0,1], (0,1), (0,1], and [0,1). In economics, variable of interest often appears in such settings. Examples include Gini coefficient (inequality measures), proportion of debt in the financial mix of firms, and efficiency scores, to name a few. Before the introduction of fractional regression models, researchers primarily used linear models to study how a set of explanatory variables influences a given fractional response variable. This approach, however, is not appropriate because linear models do not guarantee that the predicted values of the dependent variable will lie within the desired interval. By contrast, fractional regression models are efficacious in restricting the predicted values of the dependent variable within its actual interval.

Jeffrey Wooldridge and Leslie Papke, professors of economics at Michigan State University, first introduced fractional regression models in a seminal paper in 1996. In that paper, they studied participation rates of employees in firms' 401(k) retired plans by employing fractional regression models. Since then, these models have been applied in numerous studies consisting various topics. For example, to evaluate an education policy, Papke and Wooldridge (1996) studied the pass rates for an exam administered to fourth grade Michigan students over time, and J. J. S. Ramalho and

Silva (2009) studied financial leverage decisions of micro, small, medium, and large firms. The fractional regression models are either based on maximum likelihood(ML) method or quasi-maximum likelihood(QML) method. In fact, ML based models can only be used when the dependent variable does not contain 0s or 1s i.e. (0,1). By contrast, QML estimators work on any possible interval of the fractional dependent variable.

This paper surveys fractional regression models, and presents an application of these models. To be specific, section 2 of this paper delineates fractional regression. In the first subsection, it explains why other regression models are inappropriate given a fractional dependent variable. Then, the second subsection presents fractional regression models. Following the discussion of fractional regression models, the next section presents an application of these models using financial data of European manufacturing firms which either use high technology or low technology from the years 2007 and 2013. To be precise, section 3 presents a study on efficiency of firms. In doing so, it first employs data envelopment analysis to obtain efficiency scores, and then uses fractional regression models to study how those scores vary given some exogenous variables.

2. ECONOMETRIC METHODOLOGY

This chapter begins with a brief discussion on the inability of standard regression models to model a fractional response variable from a methodological perspective. Then follows a detailed discussion on maximum likelihood and quasi-maximum likelihood, which are building blocks of modeling fractional response variables. Finally, the last section presents a discussion of fractional regression models.

2.1 Standard Regression Models

Before the introduction of fractional regression models by Papke and Wooldridge (1996), there were mainly three approaches to model fractional response variables. To describe these approaches suppose that the fractional variable of interest is y such that $0 \leq y \leq 1$. Let $\mathbf{x}$ be a vector of k covariates and $\boldsymbol{\theta}$ be the vector of parameters to be estimated. Furthermore, let $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ denote the conditional density of y, which may be known or unknown.

The first approach was to use linear models such as the Ordinary Least Squares(OLS). This approach simply ignores the bounded nature of y and assumes a linear conditional mean for y:

$$E(y|\mathbf{x}) = \mathbf{X}\boldsymbol{\theta}$$

Linear models estimate a constant partial effect for the explanatory variable. However, y is strictly bounded from above and below. It is unlikely that the effect of any explanatory variable is constant throughout its entire range. Hence, constant partial effects generated by linear models are incompatible with the boundness of the dependent variable. Moreover, unless necessary adjustments are made, a linear specification cannot guarantee that predicted values of y lie between 0 and 1.

Some researchers, considering the flaws of the first approach, opted for a logistic relationship:

$$E(y|\mathbf{x}) = \frac{e^{\mathbf{X}\boldsymbol{\theta}}}{1 + e^{\mathbf{X}\boldsymbol{\theta}}}.$$

This relationship is, of course, a natural choice for modeling fractional response variables since it ensures that $0 \leq E(y|x) \leq 1$. However, because it requires nonlinear techniques, most researchers, rather than estimating the equation directly, used to choose OLS to estimate the selected log-odds ratio model:

$$E(log\frac{y}{1-y}|\mathbf{x}) = \mathbf{X}\boldsymbol{\theta}.$$

This equation corresponds to the linearization of the equation that results from solving $y = \frac{e^{(\mathbf{X}\boldsymbol{\theta})}}{1+e^{(\mathbf{X}\boldsymbol{\theta})}}$ with respect to $\mathbf{X}\boldsymbol{\theta}$. Alike the first approach, this approach is not flawless. For instance, it is not

straightforward to recover $E(y|\mathbf{x})$. As a result, the interpretation of θ, i.e. the main goal of modeling, becomes difficult. Moreover, in the log-odds ratio model the transformed dependent variable is not well defined for the boundary values of 0 and 1 of y unless ad-hoc adjustments are made — such as adding an arbitrarily chosen small constant to all observations of y.

The third approach was to use Tobit models. These models are used when there are many observations at the upper and/or lower bound of the fractional response variable. A two-limit Tobit does manage to restrict the predicted values of y between 0 and 1. However, it could be only applied when observations pile up at both extremes. In practice, often observations do not exhibit such characteristic. In addition, Tobit model is appropriate for describing censored data in the interval [0,1] but its application to data defined naturally in that interval is not easy to justify (E. A. Ramalho, Ramalho, Murteira, 2011). Also, Tobit model holds strict assumption: normality and homoscedasticity of the dependent variable, prior to censoring.

2.2 Fractional Regression Models

There are mainly two approaches for modeling fractional response variables without boundary observations. The first approach uses nonlinear models that requires only the correct specification of the conditional expectation of the fractional response variable. The second approach uses the same nonlinear models and in addition assumes a particular conditional distribution, beta distribution to be specific, for the fractional response variable. E. A. Ramalho et al. (2011) point that only the first approach is applicable when there are a finite number of boundary observations. The following subsections present discussion on the two approaches mentioned above.

2.2.1 QML Estimators

To deal with fractional response variables, Papke and Wooldridge (1996) were the first to propose that only the assumption on functional form for y is required. That is, the functional form should impose desired constraints on the conditional mean of the dependent variable:

$$E(y|\mathbf{x}) = G(\mathbf{X}\boldsymbol{\theta}),$$

where $G(.)$ is a known nonlinear function satisfying $0 \leq G(.) \leq 1$. They suggested that any cumulative function, in this process, works as a possible specification for $G(.)$. In this case, logistic function is an obvious choice. However, instead of using linearization first as discussed earlier, estimation should be done directly using nonlinear techniques.

E. A. Ramalho et al. (2011) point that a model, which is defined by the equation $E(y|\mathbf{x}) = G(\mathbf{X}\boldsymbol{\theta})$, can be estimated in two ways. First, using nonlinear least squares which was used in an empirical application by Hermalin and Wallace (1994). Second, using QML as suggested by Papke and Wooldridge (1996). In the latter case, the authors used a particular QML method based on Bernoulli log-likelihood function. It can be expressed as follows

$$LL_i(\theta) = y_i \log[G(x_i\theta] + (1 - y_i)\log[1 - G(x_i\theta)].$$

Bernoulli distribution is a member of the linear exponential family(LEF), for which the quasi-MLE is consistent even when the density is partially misspecified, hence the QML estimator of θ defined by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} LL_i(\theta)$$

is consistent and asymptotically normal, regardless of the true distribution of y conditional on x, provided that $E(y|\mathbf{x})$ discussed above is correctly specified. The asymptotic distribution of the QML estimator, as discussed earlier, is given by

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \rightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{V}).$$

2.2.2 The Beta Fractional Regression Model

A beta regression model could be employed when the value of the dependent variable is greater than zero but less than one i.e. $0 < y < 1$. It uses a parameterization of the beta distribution in terms of its mean and dispersion (or precision) parameters. This is because the standard beta density function is based on shape parameters and in practice they are of little interest. Thus, for modeling purposes, a simplified parameterization of the beta density is used. E. A. Ramalho et al. (2011) note that Paolino (2001) and Ferrari and Cribari-Neto (2004) independently proposed the mean-dispersion parametrization of the beta density. The modified parameterization makes the interpretation of beta regression models significantly simpler. These simpler beta regression models are, in fact, similar to QML models discussed previously.

The standard beta density function could be expressed in the following way

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{(p-1)}(1-y)^{(q-1)},$$

where $\Gamma(.)$ denotes the gamma function, $p > 0$ and $q > 0$ are shape parameters. By contrast, letting $p = \mu\phi$ and $q = (1-\mu)\phi$, a mean-dispersion parameterization of the beta density could be expressed as

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma[(1-\mu)\phi]} y^{(\mu\phi-1)}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$ and $\phi > 0$. This implies that

$$E(y) = \frac{p}{p+q} = \mu$$

and

$$Var(y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{\phi+1}.$$

Here ϕ can be regarded as precision parameter with the notion that, for a constant μ, as the value of ϕ gets larger the smaller the variance of y becomes.

E. A. Ramalho et al. (2011) mention about two beta regression models that could be applied based on the simplified beta density function. The first model simply assumes $\mu = G(\mathbf{X}\boldsymbol{\theta})$, and consider ϕ as a nuisance parameter. In the second model, the precision parameter and the mean are estimated differently. The model assumes that $\mu = G(\mathbf{X}\boldsymbol{\theta})$ and $\phi = \exp(\mathbf{z}\boldsymbol{\alpha})$, where $\mathbf{z}$ is a set of independent variables, potentially distinct from $\mathbf{x}$, and $\boldsymbol{\alpha}$ is a vector of parameters. The interest of this model is in analyzing whether a variable contributes to the variance of y beyond its effect upon the mean. In addition, both models are estimated by maximizing the log-likelihood function of the simplified beta density function. Hence, as in the case of MLE, estimators of these regression models are consistent and efficient. Also, the resulting estimators have the exact asymptotic distribution of MLE.

3 APPLICATION: ANALYSING EFFICIENCY

This section presents an application of fractional regression models, where the aim is to model relationship between DEA efficiency scores and exogenous factors. The first subsection outlines the main hypotheses. Then, the subsequent subsections present the data and analysis in a structured manner.

3.1 Main Hypotheses

This study focuses on the efficiency scores of manufacturing firms in Portugal, Greece, France, and Germany, and tries to analyze them by taking exogenous factors such as economic condition (before and after financial crisis), use of technology (high or low), type of firm (SME or large), and country of origin into account. Given the paucity of empirical research on this particular topic,

hypotheses of this study are formulated using pragmatic approach. The four hypotheses are given below:

1. *Firms become more efficient subsequent to financial crisis*: When a financial crisis hits an economy, firms' operations and performances, in general, suffer significantly. To ameliorate such situation, firms become more cautious about their use and allocation of resources. Thus, it is plausible that firms become more efficient after the crisis.

2. *High-Tech firms are more efficient than Low-Tech firms*: High-Tech firms, by nature, tend to spend more on research and development (R&D). This implies that, if efficiency of firms varies significantly on the basis of their use of technology, High-Tech firms would be more efficient than Low-Tech firms.

3. *SMEs are less efficient than large firms*: Large firms typically have more available funds to spend on R&D than SMEs. In effect, large firms spend more on R&D and also gain the associated advantage. In other words, large firms, as they spend more R&D, tend to use their resources more efficiently. So, one could argue that SMEs are less efficient than large firms.

4. *Firms in Portugal, France, and Germany are more efficient than firms in Greece*: Political stability, macroeconomic environment, quality of labor force, tax policies, and other related factors significantly affect how firms do their businesses. Moreover, these factors differ country to country. In other words, business environment tends to vary across nations. As a result, a particular type of firm in a given country, for example, will have to operate differently in another country. In the aftermath of financial crisis 2007-2008, Greece faced sovereign debt crisis. And since then the country has been suffering economically and politically. So, it is obvious that business environment there, compared to the other countries, is less favorable. Hence, it is plausible to infer that efficiency of firms in Portugal, France, and Germany are higher than that of Greece.

3.3 The Data

The data used in this study were obtained from Amadeus, a database of comparable financial and business information on European companies. From the Amadeus database, information about balance sheets, income statements and other characteristics of manufacturing firms for the year 2007 and 2013 were drawn, and countries namely, Portugal, Greece, France, and Germany were considered. To be specific, manufacturing firms that fall into either High-Tech or Low-Tech category were taken into consideration using Eurostat definition. For details see Table 1. Moreover, the

Table 1: Eurostat Definition of High-Tech and Low-Tech

Manufacturing industries	**NACE Rev. 2 codes – 3-digit level**
High-technology	21 Manufacture of basic pharmaceutical products and pharmaceutical preparations 26 Manufacture of computer, electronic and optical products 30.3 Manufacture of air and spacecraft and related machinery
Low-technology	10 to 17 Manufacture of food products, beverages, tobacco products, textiles, wearing apparel,leather and related products, wood and of products of wood, paper and paper products 18 Printing and reproduction of recorded media excluding 18.2 Reproduction of recorded media 31 Manufacture of furniture 32 Other manufacturing excluding 32.5 Manufacture of medical and dental instruments and supplies

study deals only with companies that are active and have unconsolidated accounts, and excludes firms that have negative equity.

This study uses pooled cross sections, and to capture the effect of the financial crisis, as an obvious case, uses a dummy variable. The variable CRISIS equals 1 if data correspond to year 2013 and 0 otherwise. That is, the base year is 2007. Furthermore, this study incorporates another dummy variable, SME, to find whether capital structure varies according to firms' size. Here,

SME equals 1 if firms are small or medium sized, and equals 0 for large firms. The study uses the definitions provided by European Commission (recommendation 2003/361/EC) to categorize firms based on their sizes. By definition, SMEs are those enterprises which employ fewer than 250 persons and have either an annual turnover not exceeding 50 million euros, or an annual balance sheet total not exceeding 43 million euros. HIGHTECH is another dummy variable that equals 1 for High-Tech firms, and 0 for Low-Tech firms.

In the first stage of DEA analysis, operating revenue turnover was considered as the only output, and total assets, shareholders' funds, and number of employees were considered as inputs. Assuming variable returns to scale, efficiency scores were then obtained using an output oriented BCC model. In this process, a random sample of 1024 companies, 512 for each year were considered. Moreover, due to the use of super efficiency, some scores exceeded 1. To apply fractional regression model in the second stage, these scores were normalized by dividing them by the maximum score. Table 2 presents the summary statistics of the first stage DEA. The mean of efficiency scores, SCORES, is 0.201 with minimum being 0.028 and maximum being 1[1]. The standard deviation for SCORES is 0.115, which gives an indication that efficiency scores do not vary significantly. The inputs and outputs, however, have very high standard deviations indicating greater variability. The mean of operating revenue turnover, total assets, and shareholders' funds are respectively 33,562.080, 24,938.228, and 5801.248[2].

Table 2: Summary Statistics DEA

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
SCORES	0.201	0.115	0.028	1
Operating Revenue Turnover	33562.080	340668.468	-1027.170	8487000
Total Assets	24938.228	259213.954	6.520	6721000
Shareholders' Funds	5801.248	29381.416	-5677.880	538583.813
Number of Employees	99.948	882.770	1	20930
N		1024		

Table 3 presents the frequency distribution of countries. In the sample, roughly 86% of the firms belong to France (50.20%) and Portugal (35.94%). The rest of the firms belong to Germany (8.01%) and Greece (5.86%). Table 4 presents summary statistics for the other three exogenous variables. The mean of the variable CRISIS is 0.5, indicating there are equal number of observations for the year 2007 and 2013. The variable HIGHTECH has a very small mean, 0.063, confirming that fewer number of firms in the sample use high technology. Last, the mean of the variable SME is 0.940, indicating that most of firms in the sample belong to SME category.

Table 3: Summary Statistics-Country

Country	Freq.	Percent	Cum.
DE	60	5.86	5.86
FR	514	50.20	56.05
GR	82	8.01	64.06
PT	368	35.94	100
Total	1024	100	

Figure 1 presents four different kernal density plots of efficiency scores, which are categorized by country, use of technology, type of firms, and crisis-period. All the four plots exhibit a common pattern: plots are positively skewed. This particular pattern indicates that most firms, regardless of their categorization, have efficiency scores below 50%. Figure 1(a) reveals that firms in Greece are less efficient compared to the firms in other countries. Figure 1(b) shows that efficiency scores

[1] In the sample there are only 7 firms with SCORES=1

[2] All the units are measured in thousands of euros.

Table 4: Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.
CRISIS	0.5	0.5	0	1
HIGHTECH	0.063	0.242	0	1
SME	0.940	0.237	0	1
N		1024		

of High-Tech and Low-Tech do no vary by much as their distributions look very similar. Figure 1(c) gives an indication that efficiency of SMEs and large firms differ significantly. In fact, density plot for SMEs is more peaked between 0.00 and 0.25 than that of large firms. Moreover, the plot for large firms exhibit a fatter right tail. Overall, these facts hint that large firms are more efficient than SMEs. Last, Figure 1(d) shows that efficiency scores of firms, loosely speaking, do not change before and after financial crisis because the density plots look almost identical.

3.4 Empirical Results

3.4.1 Main Findings

To estimate the QML model, this study primarily considers the logit specification discussed in section 2. That is,

$$E(\text{SCORES}|\mathbf{x}) = G(\beta_0 + \beta_1\text{CRISIS} + \beta_2\text{HIGHTECH} + \beta_3\text{SME} + \beta_4\text{DE} + \beta_5\text{FR} + \beta_6\text{PT}), \quad (1)$$

where $G(.)$ is the logistic function. Table 5 presents the empirical results for the main model obtained from estimating the equation (1).

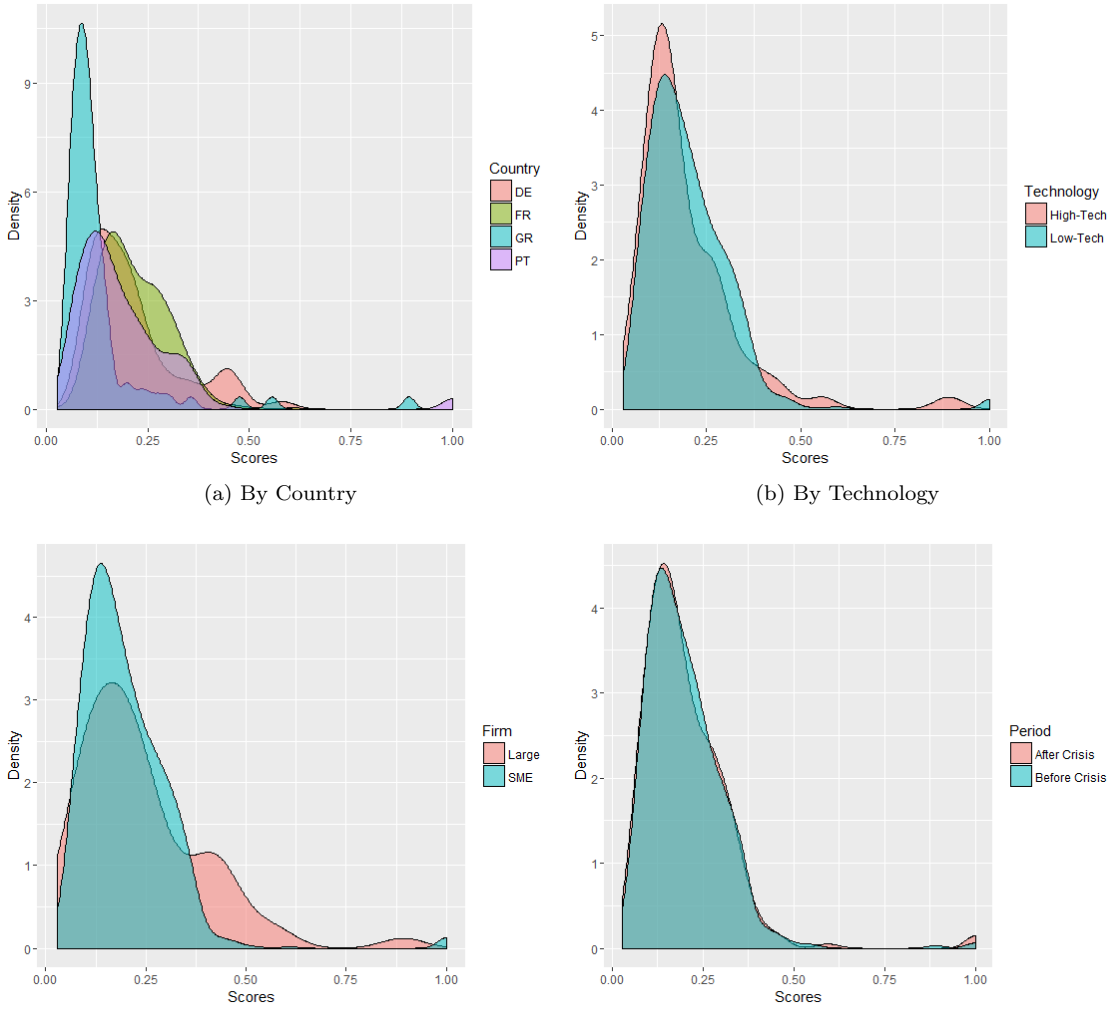
The variables CRISIS and HIGHTECH are not statistically significant. That is, financial crisis crisis and use of technology do not play any role in explaining efficiency of manufacturing firms. The variable SME, however, is statistically significant at 10% significance level and the sign of the coefficient is negative as expected. In short, SMEs are less efficient than large firms. Last, the country dummies DE, FR, and PT are statistically significant at 5% significance level, and their coefficients have positive sign. This corroborate the hypothesis that efficiency of firms vary across countries. Moreover, manufacturing firms in Germany, France, and Portugal are more efficient than manufacturing firms in Greece.

Table 5: Estimation Results : Main Model (QML Logit)

Variable	Coefficient	(Std. Err.)
CRISIS	0.007	(0.044)
HIGHTECH	-0.109	(0.109)
SME	-0.269*	(0.116)
DE	0.563**	(0.168)
FR	0.647**	(0.116)
PT	0.489**	(0.120)
Constant	-1.667**	(0.195)

3.4.2 Robustness Tests

This subsection has three aims. One, to examine whether estimates of a model change significantly when specification for G(·) changes. Two, to find out if the small number of fully-efficient firms i.e. firms with SCORES=1 affect the regression results. Three, to find out whether alternative estimation methods i.e ML and QML produce significantly different results. In doing so, this subsection first makes the use of two additional models: QML probit and QML cloglog, and compares their results with that of the main model. See Table 6 for details of the models' specifications. Furthermore, we employ QML logit model to a new sample which excludes fully-efficient firms,



(a) By Country

(b) By Technology

(c) By Type of Firm

(d) By Period

Figure 1: Distribution of Scores

Table 6: Specifications

Model Designation	Distribution Function	Cumulative Distribution $G(\mathbf{X}\boldsymbol{\theta})$
Logit	Logistic	$\frac{e^{(\mathbf{X}\boldsymbol{\theta})}}{1+e^{(\mathbf{X}\boldsymbol{\theta})}}$
Probit	Standard normal	$\boldsymbol{\Phi}(\mathbf{X}\boldsymbol{\theta})$
Cloglog	Extreme minimum	$1-\frac{e^{(\mathbf{X}\boldsymbol{\theta})}}{1+e^{(\mathbf{X}\boldsymbol{\theta})}}$

and compare the estimates with that of the main model. We then apply Beta logit model to the restricted sample, and compare the results with the new QML logit model. Estimates of these models are presented in Table 7. To avoid confusion, in the discussion of the regression results which follows next, the QML logit(1) for the full sample is referred to as the main model, and the QML logit(4) for the restricted sample is simply referred to as QML logit.

Table 7 shows that the estimates of the QML models are similar. All the variables that are significant in the main model remain significant with the same level of significance. Moreover, signs of the coefficients of the explanatory variables remain the same. In short, QML probit, QML

Table 7: Robustness Tests

	FULL SAMPLE			RESTRICTED SAMPLE	
VARIABLES	QML logit (1)	QML probit (2)	QML cloglog (3)	QML logit (4)	Beta logit (5)
CRISIS	0.007	0.003	0.007	-0.024	-0.018
	(0.044)	(0.025)	(0.039)	(0.036)	(0.033)
HIGHTECH	-0.109	-0.060	-0.100	-0.091	-0.141**
	(0.109)	(0.062)	(0.097)	(0.108)	(0.071)
SME	-0.269**	-0.162**	-0.231**	-0.283**	-0.176**
	(0.116)	(0.069)	(0.099)	(0.114)	(0.072)
DE	0.563***	0.311***	0.517***	0.556***	0.581***
	(0.168)	(0.090)	(0.155)	(0.167)	(0.099)
FR	0.647***	0.362***	0.588***	0.647***	0.659***
	(0.116)	(0.060)	(0.110)	(0.115)	(0.071)
PT	0.489***	0.272***	0.447***	0.386***	0.366***
	(0.120)	(0.062)	(0.114)	(0.114)	(0.073)
Constant	-1.667***	-0.986***	-1.770***	-1.639***	-1.734***
	(0.195)	(0.107)	(0.177)	(0.192)	(0.098)
Observations	1,024	1,024	1,024	1,017	1,017

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

cloglog, and the main model provide similar estimates.

The estimates of the QML logit model and the main model do not contradict. In fact, the explanatory variables which are significant in the main model, also remain significant in the QML logit model at the same level of significance. Moreover, the coefficients of the variables maintain the same sign across these two models. In short, exclusion of the fully-efficient firms does not bring any change to the regression results.

Last, considering the results of the two different estimation methods applied to the restricted sample, Table 7 shows that Beta logit and the main model provide similar estimates except for the variable HIGHTECH. In the Beta logit model, HIGHTECH is statistically significant at 5% significance level. By contrast, in the QML logit model the variable becomes insignificant. In any case, however, the Beta logit and QML logit do not yield any contradictory estimates in terms of signs.

3.5 Conclusion

This study uses fractional regression models and presents an analysis of efficiency scores of manufacturing firms that are based in Portugal, Greece, France, and Germany. Regression models provide evidence that financial crisis does not affect efficiency of firms. Also, use of technology does not play any role in explaining efficiency of firms. By contrast, efficiency scores of large firms and SMEs differ significantly. In fact, large firms are, on average, more efficient than SMEs. Last, there is sufficient evidence that efficiency of firms varies across countries. Fractional regression models, to be specific, indicate that firms in Portugal, France, and Germany are more efficient than firms in Greece.

REFERENCES

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics, 31(7), 799-815.

Hermalin, B. E., & Wallace, N. E. (1994). The determinants of efficiency and solvency in savings and loans. The RAND Journal of Economics, 361-381.
Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. Political Analysis, 9(4), 325–346.
Papke, L. E., & Wooldridge, J. (1996).Econometric methods for fractional response variables with an application to 401 (k) plan participation rates.National Bureau of Economic Research Cambridge, Mass., USA.
Ramalho, E. A., Ramalho, J. J. S., & Murteira, J. M. R. (2011). Alternative estimating andtesting empirical strategies for fractional regression models.Journal of Economic Surveys, 25(1), 19–68.
Ramalho, J. J. S., & Silva, J. V. (2009). A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms.Quantitative Finance, 9(5), 621–636.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Estimación non paramétrica da media para datos nesgados de gran volume

Laura Borrajo[1] e Ricardo Cao[2]

[1]Grupo de investigación MODES, Departamento de Matemáticas, CITIC, Campus de Elviña, Universidade da Coruña, 15071 A Coruña, España, laura.borrajo@udc.es

[2]Grupo de investigación MODES, Departamento de Matemáticas, CITIC e ITMATI, Campus de Elviña, Universidade da Coruña, 15071 A Coruña, España, ricardo.cao@udc.es

RESUMO

Crawford (2013) advertiu sobre os riscos que conleva afirmar que *con suficientes datos, os números falan por si mesmos*. Algúns dos problemas que xorden na análise estatística de datos de gran volume, derivados de ignorar o nesgo no proceso de mostraxe, foron analizados por Cao (2015). Neste traballo considérase o problema de inferencia estatística non paramétrica en datos de gran volume baixo a presenza de nesgo na mostraxe. Estúdase o problema da estimación da media neste contexto, nun marco non paramétrico, cando se coñece a función peso (non realista) e cando esta é descoñecida (realista). Considéranse dous escenarios diferentes para remediar o problema da función peso descoñecida: (i) ter unha mostra aleatoria simple de tamaño pequeno da poboación real e (ii) observar unha mostra dunha distribución dobremente nesgada. En ambos casos o problema está relacionado coa estimación non paramétrica da densidade. Considéranse expresións asintóticas para o erro cadrático medio dos estimadores propostos, obtendo fórmulas asintóticas para o parámetro de suavizado óptimo. O comportamento dos métodos non paramétricos propostos neste traballo ilústrase mediante resultados de simulación.

Palabras e frases chave: Big Data, corrección do nesgo, estimación da media, inferencia non paramétrica, método núcleo

1. INTRODUCIÓN

Co auxe dos datos de gran volume (Big Data, BD) xorde, a súa vez, a falsa suposición de que *con suficientes datos, os números falan por si sós*. Non obstante, con frecuencia estes datos non son totalmente representativos da poboación, senón que están nesgados: datos de gran volume mais nesgados (Big-but-biased Data, BBBD=B3D). Este nesgo provén, a miúdo, do procedemento de obtención dos datos, xa que os métodos de mostraxe nos que os individuos se autoseleccionan conlevan, en certas ocasións, nesgos na mostraxe.

Un bo exemplo ao que Crawford (2013) fai mención é o dos datos recollidos na cidade de Boston por medio do Street Bump. Trátase dunha aplicación para teléfonos intelixentes que detecta os solavancos nas estradas por medio dos acelerómetros do teléfono e os datos do GPS durante a condución dun automóbil. Estes datos son enviados ao Departamento de Tráfico da Cidade de Boston, dende onde se planifica a súa reparación e a xestión de recursos da forma máis eficiente posible. Non obstante, un dos problemas observados foi que algúns segmentos da poboación, como as clases menos favorecidas e os grupos de idade avanzada, teñen unha baixa taxa de acceso a teléfonos intelixentes. Deste xeito, estes datos proporcionan unha mostra grande en número pero moi nesgada da poboación de solavancos na cidade, o que repercute na infraestimación do número de solavancos en certos barrios e nunha xestión deficiente dos recursos.

Outro exemplo citado por Crawford (2013) é a base de datos de máis de 20 millóns de tweets orixinados polo furacán Sandy entre o 27 de outubro e o 1 de novembro de 2012. Estes datos proveñen dunha mostra nesgada da poboación, xa que a maior parte de tweets sobre Sandy procederon

de Manhattan, feito relacionado co elevado número de propietarios de teléfonos intelixentes en Nova York, mentres que poucos tweets tiveron a súa orixe nas zonas máis afectadas pola catástrofe, nas que ademais os cortes eléctricos ocasionaron problemas de acceso a internet e que moitos dispositivos se quedasen sen batería.

Noutros exemplos, como os citados en Hargittai (2015), obsérvase como os individuos non seleccionan de xeito aleatorio os sitios web, senón que o seu uso está nesgado, dando lugar a mostras que limitan a xeneralización de resultados.

Neste contexto, considérase unha poboación con función de distribución F (densidade f) e unha mostra aleatoria simple

$$\mathbf{X} = (X_1, \dots, X_n)$$

de tamaño n desa poboación. Supóñase que non se pode observar esta mostra aleatoria simple $\mathbf{X}$, pero si se observa unha mostra

$$\mathbf{Y} = (Y_1, \dots, Y_N)$$

de tamaño moito maior ($N >> n$), procedente dunha distribución nesgada G (densidade g), tal que

$$g(x) = w(x)f(x)$$

para algunha función peso w, con $w(x) \geq 0$.

Na sección 2 abórdase o problema da estimación non paramétrica da media no contexto de nesgo proposto, cando a función peso é coñecida (caso non realista) e descoñecida (caso realista). Considéranse dous escenarios diferentes para solventar o problema da función peso descoñecida: (i) ter unha mostra aleatoria simple de tamaño pequeno da poboación real e (ii) observar unha mostra dunha distribución dobremente nesgada. En ambos casos, o problema está relacionado coa estimación non paramétrica da densidade. Na sección 3 amósanse expresións asintóticas para o erro cadrático medio dos estimadores propostos, obtendo fórmulas asintóticas para o parámetro de suavizado óptimo. Na sección 4 ilústrase o comportamento dos métodos non paramétricos propostos por medio dalgunhas simulacións.

2. ESTIMACIÓN DA MEDIA EN B3D

Para abordar a situación de nesgo en Big Data, comézase cun problema sinxelo: a estimación da media $\mu = \int x dF(x)$ nun contexto non paramétrico. Se se observa a mostra $\mathbf{X}$, entón bastaría con empregar a súa media mostral, $\overline{X}$, que constitúe un estimador innesgado de μ, coincidindo o seu erro cadrático medio coa súa varianza $Var(\overline{X}) = \frac{\sigma^2}{n}$. Nese caso, o Teorema Central do Límite permitiría facer inferencia sobre μ.

Non obstante, no contexto B3D proposto, pode xurdir a dúbida de se é factible estimar a media ao non ser posible observar $\mathbf{X}$. Efectivamente, supoñendo que a distribución F é continua, con densidade f, a media da mostra Big Data nesgada, $\overline{Y}$, non é un estimador consistente de $\mu = \int x f(x)\, dx$, pero si o é de

$$\mu_g = \int x g(x)\, dx = \int x w(x)\, f(x)\, dx.$$

Por outra banda, ao ser $w(x) = \frac{g(x)}{f(x)}$, é sinxelo comprobar que o cociente $\frac{Y}{w(Y)}$ é un estimador innesgado da media:

$$\begin{aligned} E\left(\frac{Y}{w(Y)}\right) &= \int \frac{y}{w(y)} g(y)\, dy = \int \frac{y}{g(y)/f(y)} g(y)\, dy \\ &= \int y f(y)\, dy = \mu. \end{aligned}$$

Polo tanto, en caso de coñecer a función peso w (caso non realista) poderíase empregar a esperanza anterior para definir un estimador de μ como segue:

$$\hat{\mu}^{BBBS,w} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{w(Y_i)}, \tag{1}$$

onde *BBBS*, polas súas iniciais en inglés, fai referencia a mostra Big Data nesgada. Este estimador consiste na media mostral da poboación $Z = Y/w(Y)$, sendo o seu nesgo, varianza e distribución asintótica:

$$\begin{aligned} E\left(\hat{\mu}^{BBBS,w}\right) &= \mu, \\ Var\left(\hat{\mu}^{BBBS,w}\right) &= \frac{\sigma_Z^2}{N}, \\ \frac{\sqrt{N}\left(\hat{\mu}^{BBBS,w} - \mu\right)}{\sigma_Z} &\to N(0,1), \end{aligned}$$

onde $\sigma_Z^2 = \int y^2 f(y)^2 g(y)^{-1} dy - \mu^2 = \int x^2 f(x) w(x)^{-1} dx - \mu^2$.

Non obstante, a situación máis realista é que a función peso w sexa descoñecida, nese caso débese estimar. Como a mostra Big Data non é suficiente, propóñense diferentes escenarios nos que certa información adicional permita facelo:

1. Dispor dunha mostra aleatoria simple de tamaño pequeno, n, $(X_1, \ldots, X_n)$, da verdadeira poboación
2. Replicar o mecanismo de nesgo mediante o cal a mostra nesgada foi obtida para obter unha mostra $(Z_1, \ldots, Z_n)$ procedente dunha distribución dúas veces nesgada M (densidade m)
3. Enriquecer a mostra nesgada con información das frecuencias (ν_i, para $i = 1, \ldots, N$) de cada dato da mostra Y observada

Ao longo deste traballo considéranse os escenarios 1 e 2.

Lloyd e Jones (2000) estudaron o estimador da densidade tipo núcleo non paramétrico para datos nesgados con función peso descoñecida, pero nun contexto diferente: próximo ao escenario 2, utilizando dúas mostras nesgadas e sendo $w(x)$ a probabilidade de aceptación.

Escenario 1. Considérase unha poboación con función de distribución F (densidade f). Ademais da mostra Big Data

$$\mathbf{Y} = (Y_1, \ldots, Y_N)$$

de tamaño N procedente da distribución nesgada G (densidade g), tamén se observa unha mostra aleatoria simple

$$\mathbf{X} = (X_1, \ldots, X_n)$$

de menor tamaño mostral ($n << N$) desa poboación. Supóñase que

$$g(x) = w(x) f(x)$$

para algunha función peso w, tal que $w(x) \geq 0$, $\forall x$.

Emprégase o estimador de Parzen-Rosenblatt para estimar $f(x)$ e $g(x)$:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

$$\hat{g}_b(x) = \frac{1}{N} \sum_{i=1}^{N} K_b(x - Y_i)$$

onde $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$, sendo K a función núcleo e h e b os parámetros de suavizado. Isto permite estimar de forma sinxela a función peso w como o cociente de ámbalas dúas densidades estimadas:

$$\hat{w}_{h,b}(x) = \frac{\hat{g}_b(x)}{\hat{f}_h(x)}$$

De xeito análogo ao caso coñecido, substituíndo en (1) a función peso w pola sua versión estimada $\hat{w}_{h,b}$, obtense o seguinte estimador da media para o escenario 1:

$$\hat{\mu}^{BBBS,\hat{w}_{h,b}} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{\hat{w}_{h,b}(Y_i)} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i \hat{f}_h(Y_i)}{\hat{g}_b(Y_i)}, \tag{2}$$

o cal depende de ámbolos dous parámetros de suavizado h e b.

Escenario 2. Considérase unha poboación con función de distribución F (densidade f). Supóñase por un intre que non é posible observar unha mostra aleatoria simple desa poboación pero, en vez diso, obsérvase outra mostra

$$\mathbf{Y} = (Y_1, \ldots, Y_N)$$

de tamaño N procedente dunha distribución nesgada G (densidade g) e tamén se observa unha mostra

$$\mathbf{Z} = (Z_1, \ldots, Z_n)$$

de tamaño $n << N$ procedente dunha distribución dúas veces nesgada M (densidade m).

Na Táboa 1 descríbese o proceso a seguir neste escenario, onde cabe destacar que a función peso w_0 debe ser a mesma en ambos casos.

paso	densidade	función nesgo	densidade nesgada	función nesgo normalizada
1	$f(x)$	$w_0(x)$	$g(x) \propto w_0(x)f(x)$	$w(x) = \dfrac{w_0(x)}{\int w_0(y)f(y)dy}$
2	$g(x)$	$w_0(x)$	$m(x) \propto w_0(x)g(x)$	$w_2(x) = \dfrac{w_0(x)}{\int w_0(y)g(y)dy}$

Táboa 1: Proceso de obtención da mostra dúas veces nesgada no escenario 2.

Suponse, entón, que a densidade g se corresponde coa densidade f ponderada por unha función peso w, mentres que a densidade m se corresponde co produto da densidade g por unha función peso w_2:

$$g(x) = f(x)w(x)$$
$$m(x) = g(x)w_2(x)$$

sendo $w(x), w_2(x) \geq 0$, $\forall x$.

A relación entre ambas funcións peso ven dada por $w_2(x) = \rho w(x)$, onde

$$\rho = \frac{\int f(y)w_0(y)dy}{\int g(y)w_0(y)dy}.$$

No tocante á estimación da media nesta situación, dado que $w_2(x) = \frac{m(x)}{g(x)}$ e $\frac{m(x)}{g(x)} = \rho\frac{g(x)}{f(x)}$, é sinxelo comprobar que a esperanza de $\frac{Y}{w_2(Y)}$ é a media pola inversa da constante ρ:

$$\begin{aligned} E\left(\frac{Y}{w_2(Y)}\right) &= \int \frac{y}{w_2(y)} g(y)\, dy = \int \frac{y}{m(y)/g(y)} g(y)\, dy \\ &= \int \frac{y}{\rho g(y)/f(y)} g(y)\, dy = \frac{1}{\rho}\int y f(y)\, dy = \frac{1}{\rho}\mu. \end{aligned}$$

Por outra parte, a esperanza da inversa da función peso w_2 é a inversa da constante ρ:

$$E\left(\frac{1}{w_2(Y)}\right) = \int \frac{1}{w_2(y)} g(y)\, dy = \frac{1}{\rho}\int f(y)\, dy = \frac{1}{\rho}.$$

Consecuentemente, o cociente das dúas esperanzas anteriores é exactamente a media:

$$\frac{E\left(\dfrac{Y}{w_2(Y)}\right)}{E\left(\dfrac{1}{w_2(Y)}\right)} = \mu.$$

Polo tanto, se ocorre que w_2 é coñecida, pódense empregar as esperanzas anteriores para motivar un estimador para μ, simplemente reemprazando as esperanzas teóricas polas medias empíricas:

$$\hat{\mu}^{BBBS_2,w_2} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{w_2(Y_i)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{w_2(Y_i)}} = \frac{\frac{1}{N}\sum_{i=1}^{N}Y_i\frac{g(Y_i)}{m(Y_i)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{g(Y_i)}{m(Y_i)}} \tag{3}$$

Seguindo un procedemento análogo ao do escenario 1, recórrese ao estimador tipo núcleo de Parzen-Rosenblatt para estimar $g(x)$ e $m(x)$:

$$\hat{g}_h(x) = \frac{1}{N}\sum_{i=1}^{N}K_h(x - Y_i)$$

$$\hat{m}_b(x) = \frac{1}{n}\sum_{i=1}^{n}K_b(x - Z_i)$$

o que permite estimar de xeito sinxelo a función peso w_2 como o cociente das densidades estimadas:

$$\hat{w_{2h,b}}(x) = \frac{\hat{m}_b(x)}{\hat{g}_h(x)}.$$

Introducindo $\hat{w_{2h,b}}$ en (3), defínese un novo estimador para a media no escenario 2:

$$\hat{\mu}^{BBBS_2,\hat{w_{2h,b}}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{\hat{w_{2h,b}}(Y_i)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{w_{2h,b}}(Y_i)}} = \frac{\frac{1}{N}\sum_{i=1}^{N}Y_i\frac{\hat{g}_h(Y_i)}{\hat{m}_b(Y_i)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{g}_h(Y_i)}{\hat{m}_b(Y_i)}}, \tag{4}$$

o cal depende dos parámetros de suavizado h e b.

3. VERSIÓNS APROXIMADAS DOS ESTIMADORES E RESULTADOS ASINTÓTICOS

Co obxectivo de estudar as propiedades asintóticas dos estimadores propostos, considéranse versións aproximadas dos mesmos, consistentes en supoñer que a densidade g é coñecida, unha situación moi natural no contexto B3D no que se está a traballar, posto que $N >> n$ (é dicir, $\lim_{n\to\infty} N/n = \infty$).

Escenario 1. Co fin de analizar o comportamento asintótico do estimador (2) proposto no primeiro escenario analizado, considérase unha versión aproximada deste supoñendo g coñecida:

$$\hat{\mu}^{BBBS,\hat{w}_h} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i\hat{f}_h(Y_i)}{g(Y_i)}$$

obtendo un novo "estimador" da media que tan só depende do parámetro de suavizado h.

Resulta sinxelo comprobar que $\hat{\mu}^{BBBS,\hat{w}_h}$ é un estimador innesgado:

$$\begin{aligned} E\left(\hat{\mu}^{BBBS,\hat{w}_h}\right) &= E\left(\frac{Y_1\hat{f}_h(Y_1)}{g(Y_1)}\right) = E\left(\frac{Y_1 \cdot K_h * f(Y_1)}{g(Y_1)}\right) \\ &= \int y \cdot K_h * f(y)\,dy = \int y \cdot f(y)\,dy = \mu, \end{aligned}$$

sendo a súa varianza asintótica e, consecuentemente, o seu erro cadrático medio ($h \to 0$, $nh \to \infty$ e $N/n \to \infty$) o seguinte:

$$\begin{aligned} MSE\left(\hat{\mu}^{BBBS,\hat{w}_h}\right) &= Var\left(\hat{\mu}^{BBBS,\hat{w}_h}\right) = \frac{C_1}{n} + \frac{C_2 h^2}{n} + \frac{C_3}{Nnh} + \frac{C_4}{N} \\ &+ O\left(\frac{h^4}{n}\right) + O\left(\frac{h^2}{N}\right) + O\left(\frac{1}{Nn}\right), \end{aligned}$$

onde

$$\begin{aligned} C_1 &= \int x^2 f(x)dx - \mu^2 \geq 0, \\ C_2 &= \mu_2(K)\left[\frac{1}{2}\int x^2 f''(x)dx + \int x f'(x)dx\right], \\ C_3 &= R(K^2)\int x^2 \frac{f(x)}{g(x)}dx \geq 0, \\ C_4 &= \int x^2 \frac{f(x)^2}{g(x)}dx - \mu^2 \geq 0, \end{aligned}$$

$$R(K^2) = \int K(u)^2 du, \qquad \mu_2(K) = \int u^2 K(u)du.$$

O erro cadrático medio asintótico, en función do parámetro de suavizado, h, ven dado por

$$AMSE\left(h\right) = \frac{C_1}{n} + \frac{C_2 h^2}{n} + \frac{C_3}{Nnh} + \frac{C_4}{N}.$$

Dependendo do signo que teña a constante C_2 cómpre diferenciar entre dúas situacións:

- No suposto de que $C_2 > 0$, a expresión anterior pode ser sinxelamente minimizada en h, obtendo a seguinte expresión para a ventá óptima:

  $$h_{AMSE} = \left(\frac{C_3}{2C_2}\right)^{1/3} N^{-1/3}.$$

  Pero de ser ese o caso, todos os termos na expresión do $AMSE(h)$ serían positivos, e polo tanto

  $$AMSE(h_{AMSE}) > \frac{C_1}{n} = MSE(\overline{X});$$

  é dicir, empregar a media da mostra aleatoria simple como estimador sería unha opción preferible ao estimador proposto.

- No caso de que $C_2 < 0$, a mellor opción sería escoller un h moi grande, xa que o $AMSE(h)$ decrece a medida que este aumenta. Non obstante, espérase que algúns dos termos desprezables do $MSE(h)$ incrementen o seu valor con h. Nese caso, o h óptimo non debería ser infinitamente grande e polo tanto, teríase que

  $$MSE(h_{MSE}) < MSE(\overline{X}).$$

Escenario 2. Co obxectivo de estudar as propiedades asintóticas do estimador da media proposto neste escenario, considérase unha versión aproximada de (4), supoñendo que a densidade g é coñecida:

$$\hat{\mu}^{BBBS_2,\hat{w}_{2b}} = \frac{\frac{1}{N}\sum_{i=1}^{N} Y_i \frac{g\left(Y_i\right)}{\hat{m}_b\left(Y_i\right)}}{\frac{1}{N}\sum_{i=1}^{N} \frac{g\left(Y_i\right)}{\hat{m}_b\left(Y_i\right)}} \tag{5}$$

obtendo así un novo "estimador" que tan só depende do parámetro de suavizado b.

O estimador $\hat{\mu}^{BBBS_{2},\hat{w}_{2b}}$ non é un estimador innesgado:

$$E\left(\hat{\mu}^{BBBS_{2},\hat{w}_{2b}}-\mu\right)=\frac{C_5}{nb}-C_6b^2-\frac{C_7}{n}+\frac{C_8b}{n}-\frac{2C_6b^2}{n}+O(b^4)$$

e a súa varianza asintótica ($b\to 0$, $nb\to\infty$ e $N/n\to\infty$) ten a seguinte expresión:

$$\begin{aligned}Var\left(\hat{\mu}^{BBBS_{2},\hat{w}_{2b}}\right) &= \frac{C_9}{n}+\frac{C_{10}}{N}+\frac{C_{11}b^2}{n}+\frac{C_{12}}{n^2b}+\frac{C_{13}}{Nnb}+\frac{C_{14}b^2}{N}\\ &+ O\left(\frac{b}{n^2}\right)+O\left(\frac{1}{n^3b}\right),\end{aligned}$$

onde

$$\begin{aligned}C_5 &= \omega R(K^2)\int (y-\mu)\frac{g(y)^2}{m(y)^2}dy,\\ C_6 &= \frac{\omega\mu_2(K)}{2}\int (y-\mu)\frac{g(y)^2m''(y)}{m(y)^2}dy,\\ C_7 &= \omega\int (y-\mu)\frac{g(y)^2}{m(y)}dy,\\ C_8 &= \frac{\omega\mu_2(K^2)}{2}\int (y-\mu)\frac{g(y)^2m''(y)}{m(y)^3}dy,\\ C_9 &= \omega^2Var\left((Z-\mu)\frac{g(Z)^2}{m(Z)^2}\right),\\ C_{10} &= \omega^2Var\left((Y-\mu)\frac{g(Y)}{m(Y)}\right),\end{aligned}$$

$$\omega=\left(\int\frac{g(y)^2}{m(y)}dy\right)^{-1},\qquad \mu_2(K^t)=\int u^2K^t(u)du.$$

Omítense o resto de constantes por ter expresións máis complexas e ser menos relevantes na obtención da ventá óptima.

En consecuencia, o erro cadrático medio é

$$\begin{aligned}MSE\left(\hat{\mu}^{BBBS_{2},\hat{w}_{2b}}\right) &= \frac{{C_5}^2}{n^2b^2}+{C_6}^2b^4-\frac{2C_5C_6b}{n}+\frac{C_9}{n}+\frac{C_{10}}{N}+\frac{(C_{12}-2C_5C_7)}{n^2b}\\ &+ \frac{(C_{11}+2C_6C_7)b^2}{n}+\frac{({C_7}^2+2C_5C_8)}{n^2}+\frac{C_{13}}{Nnb}+\frac{C_{14}b^2}{N}\\ &+ O\left(\frac{b}{n^2}\right)+O\left(\frac{1}{n^3b}\right).\end{aligned}$$

Polo tanto, o erro cadrático medio asintótico ven dado por

$$AMSE\left(b\right) = \frac{C_9}{n}+\frac{C_{10}}{N}+\frac{({C_7}^2+2C_5C_8)}{n^2}+\Psi(b),$$

sendo $\Psi(b)$ o termo dependente do parámetro de suavizado b,

$$\begin{aligned}\Psi(b) &= \frac{{C_5}^2}{n^2b^2}+{C_6}^2b^4-\frac{2C_5C_6b}{n}+\frac{(C_{12}-2C_5C_7)}{n^2b}\\ &+ \frac{(C_{11}+2C_6C_7)b^2}{n}+\frac{C_{13}}{Nnb}+\frac{C_{14}b^2}{N},\end{aligned}$$

e polo tanto, o relevante na obtención da ventá óptima que minimice o erro cadrático medio asintótico.

O óptimo acádase no suposto de que se cumpra a condición $nb^3\to c$, onde c é unha constante, obtendo a seguinte expresión para a ventá óptima:

$$b_{AMSE} = \left(\frac{C_5}{C_6}\right)^{1/3}n^{-1/3},$$

xa que nese caso, tense que

$$\begin{aligned}\Psi(c\cdot n^{-1/3}) &= ({C_5}^2+{C_6}^2-2C_5C_6)n^{-4/3}+O(n^{-5/3})+O(N^{-1}n^{-2/3})\\ &= (C_5-C_6)^2n^{-4/3}+o(n^{-4/3}).\end{aligned}$$

Pode comprobarse como, de supoñer o contrario, se se considera b_n tal que ${nb_n}^3\to 0$ ou ${nb_n}^3\to\infty$ e se supón que b_n é o óptimo,

$$AMSE(b_n) \leq AMSE(c\cdot n^{-1/3}),$$

se chegaría a unha contradición, resultante de ter en conta certas condicións asintóticas e de aplicar límites.

4. ESTUDO DE SIMULACIÓN

Nesta sección amósase o comportamento dos estimadores propostos mediante resultados de simulación.

Escenario 1. No primeiro escenario analizado, considérase unha poboación con distribución uniforme e densidade

$$f(x)=\begin{cases}\frac{1}{2} & \text{se} \quad x\in[0,2]\\ 0 & \text{se} \quad x\notin[0,2]\end{cases}$$

da que se observa unha mostra de pequeno tamaño.

Considérase a seguinte clase de funcións peso

$$w(x)=\begin{cases}\varepsilon^k & \text{se} \quad x\in[0,\varepsilon]\\ x^k & \text{se} \quad x\in[\varepsilon,2]\\ 0 & \text{se} \quad x\notin[0,2]\end{cases}$$

mediante a cal se obtén a seguinte expresión para a densidade nesgada

$$g(x)=\begin{cases}\dfrac{k+1}{k\varepsilon^{k+1}+2^{k+1}}\varepsilon^k & \text{se} \quad x\in[0,\varepsilon]\\ \dfrac{k+1}{k\varepsilon^{k+1}+2^{k+1}}x^k & \text{se} \quad x\in[\varepsilon,2]\\ 0 & \text{se} \quad x\notin[0,2]\end{cases}$$

Ademais da mostra aleatoria simple da poboación, tamén se observa unha mostra Big Data procedente da distribución nesgada G.

Na Figura 1 amósase o gráfico das densidades involucradas nesta simulación: a densidade uniforme, f, en cor negra e a densidade nesgada, g, en cor vermella, para a particular elección de $k=1$ e ε =1.5. Tamén se amosa o comportamento da función peso en cor azul.

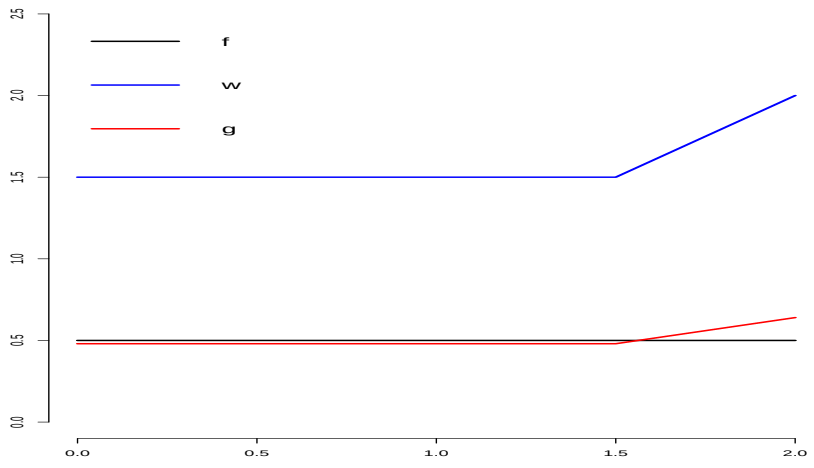


Figura 1: Densidades involucradas na simulación para o escenario 1 con $k=1$ e ε =1.5.

Simulando unha posible situación do escenario 1, considérase $n = 10^3$ e $N = 10^6$. Na Figura 2 represéntase o erro cadrático medio, transformado por logaritmo, do estimador (2) en función de diferentes valores do logaritmo dos parámetros de suavizado h e b. Neste caso, o logaritmo do erro cadrático medio cometido ao empregar a media da mostra aleatoria simple, $\overline{X}$, como estimador é de -3.46; mentres que coa media da mostra Big Data, $\overline{Y}$, aumenta ata -2.95. Observando a Figura 2, pódese comprobar como para certas combinacións das ventás b e h, o logaritmo do erro cadrático medio cometido co estimador proposto (2) é menor que o obtido mediante a estimación por medio de ámbalas dúas mostras.

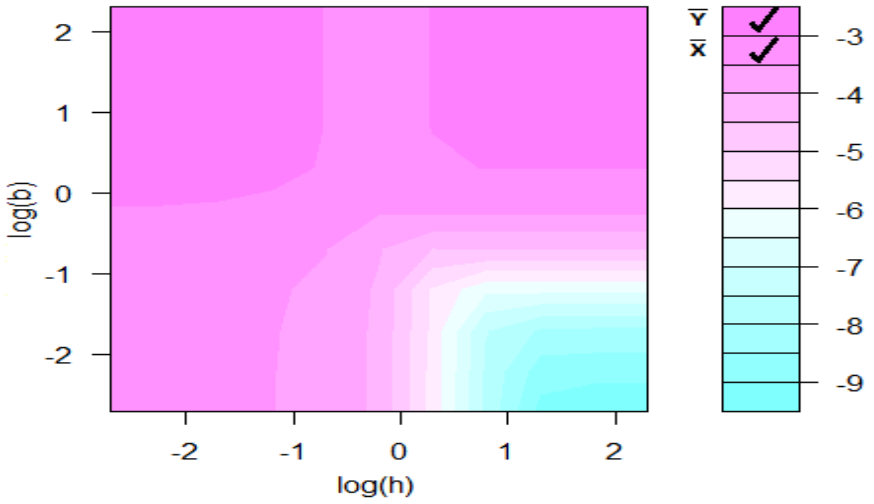


Figura 2: Logaritmo do erro cadrático medio do estimador (2) en función de $\log(h)$ e $\log(b)$.

Escenario 2. Considérase de novo o modelo proposto para o escenario 1. Non obstante, supóñase que non é posible observar unha mostra aleatoria simple da poboación, senón que se observa unha mostra de gran tamaño procedente da distribución nesgada G (densidade g) e outra mostra, de menor tamaño, da distribución dúas veces nesgada M, con densidade

$$m(x) = \begin{cases} \dfrac{2k+1}{2k\varepsilon^{2k+1} + 2^{2k+1}}\varepsilon^{2k} & \text{se} \quad x \in [0, \varepsilon] \\ \dfrac{2k+1}{2k\varepsilon^{2k+1} + 2^{2k+1}}x^{2k} & \text{se} \quad x \in [\varepsilon, 2] \\ 0 & \text{se} \quad x \notin [0, 2] \end{cases}$$

Na Figura 3 amósase o gráfico das densidades involucradas neste escenario: a densidade uniforme f en cor negra, a densidade nesgada g en cor vermella e a densidade dúas veces nesgada m en cor verde, para a selección $k = 1$ e ε =0.2. Tamén se amosa o comportamento da función peso en cor azul.

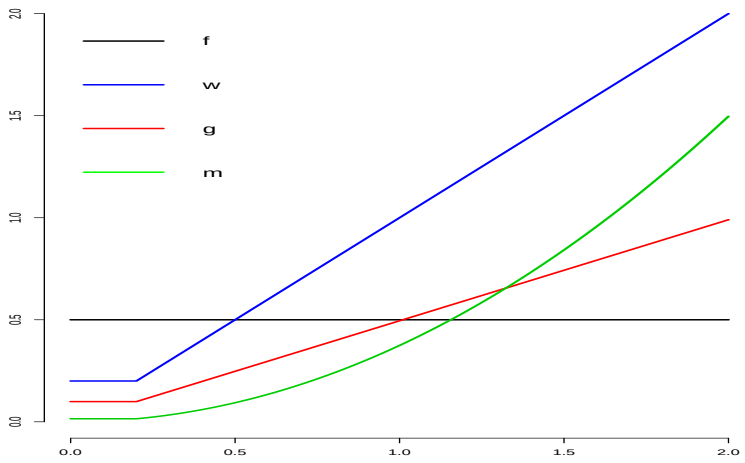


Figura 3: Densidades involucradas na simulación para o escenario 2 con $k = 1$ e ε =0.2.

A simulación deste modelo realízase con tamaño mostral $n = 10^3$ para a mostra dobremente nesgada e $N = 10^6$ para a mostra Big Data. Na Figura 4 represéntase, en cor verde, o erro cadrático medio cometido ao empregar o estimador (5), en función de diferentes valores do parámetro de suavizado b; en cor vermella, amósase o erro cometido no suposto caso de que a función peso w fose coñecida; en cor morada, o erro ao empregar a media da mostra Big Data nesgada e en cor laranxa, o correspondente a utilizar a media da mostra dobremente nesgada Z como estimador da media. Os resultados obtidos confirman o bo funcionamento do estimador proposto, pois non só mellora a estimación realizada pola media das mostras involucradas senón que tamén se aproxima ao caso en que a función peso se supón coñecida.

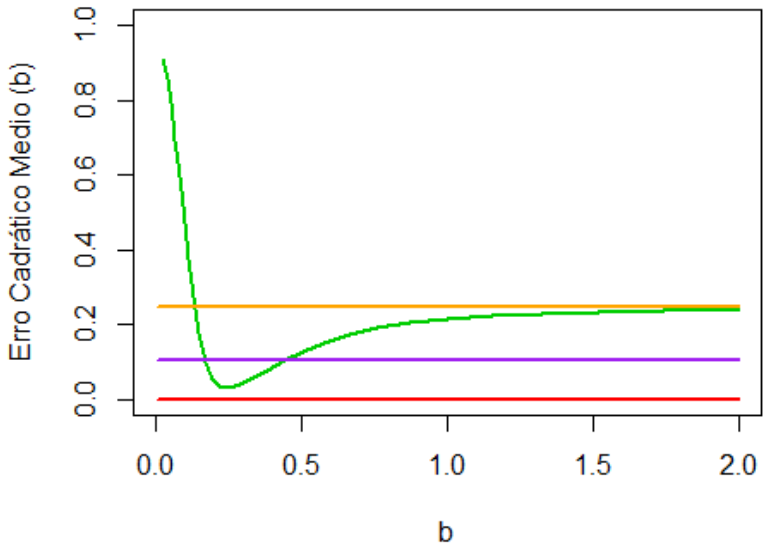


Figura 4: Erro cadrático medio do estimador (5) en función de b (verde), do estimador (1) (vermello), de $\overline{Y}$ (morado) e de $\overline{Z}$ (laranxa) .

5. CONCLUSIÓNS

Para concluír, é importante salientar os novos retos estatísticos que se presentan co Big Data, posto que neste contexto o nesgo ten unha maior tendencia a producirse. Este feito ocasiona que contrastar se estamos ante unha situación de nesgo sexa un problema relevante e de moito interese en Big Data.

Neste traballo foron propostos diferentes estimadores cos que abordar diversas situacións de nesgo en datos de gran volume. Os resultados obtidos permiten concluír que o nesgo pode correxirse co prezo dalgunha información adicional necesaria: posuír unha mostra aleatoria simple de pequeno tamaño da poboación ou unha mostra dúas veces nesgada, obtida mediante o mesmo proceso de nesgo polo que foi obtida a mostra Big Data. Ademais, obsérvase como certas ideas sobre datos nesgados por lonxitude ou datos nesgados en xeral poden xogar un papel importante no contexto B3D no que se está a traballar; así como as técnicas de suavizado non paramétricas, que teñen un papel fundamental na corrección do nesgo.

Como traballo futuro preténdese realizar contrastes de detección de nesgo en Big Data e xorde a posibilidade de estender este estudo á estimación doutros parámetros, como a varianza, e a resolución doutros problemas, como modelos de regresión e clasificación.

Agradecementos Esta investigación foi financiada polo MINECO a través do proxecto MTM-2014-52876-R e pola Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 e Centro Singular de Investigación de Galicia ED431G/01). Todas estas axudas foron cofinanciadas polo FEDER. A investigación da primeira autora foi financiada polas axudas de apoio á etapa predoutoral nas universidades do Sistema Universitario Galego, nos organismos públicos de investigación de Galicia e noutras entidades do Sistema galego de I+D+i, cuxo financiamento procede da Unión Europea a través do Fondo Social Europeo nun 80 % e no 20 % restante da Secretaría Xeral de Universidades, pertencente á Consellería de Cultura, Educación e Ordenación Universitaria da Xunta de Galicia.

REFERENCIAS

Cao, R. (2015) Inferencia estadística con datos de gran volumen. La Gaceta de la RSME, 18, 393–417

Crawford, K. (2013) The hidden biases in big data. Harvard Business Review, April 1st. Dispoñible en https://hbr.org/2013/04/the-hidden-biases-in-big-data

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. The Annals of the American Academy of Political and Social Science, 659, 63–76.

Lloyd, C.J., Jones, M.C. (2000). Nonparametric Density Estimation from Biased Data with Unknown Biasing Function. Journal of the American Statistical Association, 95, 865–876.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

CHARACTERISATION OF POINT PROCESSES WITH COVARIATES THROUGH THE FIRST-ORDER INTENSITY FUNCTION

M.I. Borrajo[1], W. González-Manteiga [1] e M.D. Martínez-Miranda[2]

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

[2] Departamento de Estadística e Investigación Operativa. Universidad de Granada.

ABSTRACT

Point processes are a branch of spatial statistics, whose main aim is to study the geometrical structure of patterns formed by objects that are distributed randomly in number and space. This type of data arise in many different fields such as ecology, J. B. Illian, Møller, and Waagepetersen (2009); epidemiology, P. J. Diggle (1990); astronomy, Babu and Feigelson (1996); forestry, Stoyan and Penttinen (2000) and seismology, Ogata and Zhuang (2006) and Schoenberg (2011). The first-order intensity function is one of the characteristic functions of a point processes and has generated a great interest since the eighties. Covariates are a way of gathering extra information, allowing more precise models for point processes without assuming parametric restrictions.

This work provides two main contributions: first a new kernel intensity estimator is proposed and a theoretical framework to guarantee its consistency is developed. A new bootstrap resampling procedure as well as two new data-driven bandwidth selection methods are suggested. The second contribution is a test o check whether the assumed model with covariates is or not appropriate, for which a L^2-distance based statistic is proposed and is proved to be asymptotically normal. The good behaviour of all the proposed techniques is shown through different extensive simulation studies and they are also applied to a real data set consisting of locations of gold deposits in the Murchison area of Western Australia.

Keywords: Point processes; Covariates; First-order intensity; Kernel estimation; Testing significance.

1. INTRODUCTION TO POINT PROCESSES

Point process theory has its roots in the introduction of the Poisson distribution in 1837 by Siméon-Denis Poisson, which derived it as a passage to the limit from the binomial distribution, It allowed to predict the pattern in which random events of low probability occurs in the course of a large number of trials. The main advances in the field of point processes have been done during the last decades of the 20th century, see Daley and Vere-Jones (1988), Moller and Waagepetersen (2003), J. Illian, Penttinen, Stoyan, and Stoyan (2008), P. J. Diggle (2013) and Baddeley, Rubak, and Turner (2015) for an overview of the theoretical background of point processes.

In this work we have focused on spatial point processes, i.e, point processes in $\mathbb{R}^2$. First of all we need to introduce some basic concepts that are fundamental in the understanding of the theory.

Definition 1 *Let X be a point process in $W \subset \mathbb{R}^2$, let $\mathcal{P}(W)$ denote the parts of W, i e., the family of all possible subsets of W; then $N : \mathcal{P}(W) \to \mathbb{Z}^+$, where $N(A)$ denote the number of occurrences of the process in the set $A \in \mathcal{P}(W)$, is defined as the* ***counting measure*** *associated with the process X.*

Definition 2 *Let X be a point process in $W \subset \mathbb{R}^d$ and N its associated counting measure, then a realisation (or a sample) of the process, $X_1, \ldots, X_N$, is called a* ***point pattern****.*

Figure 1 provides examples of point patterns. These graphics gather the existing possibilities for a pattern in terms of "clusterisation": the one on the left is the most clustered, i.e., we can clearly distinguish groups of points that lie together; the one on the right is regular, the points tend to avoid each other; and the one in the middle, shows points randomly placed and do not following any specific pattern.

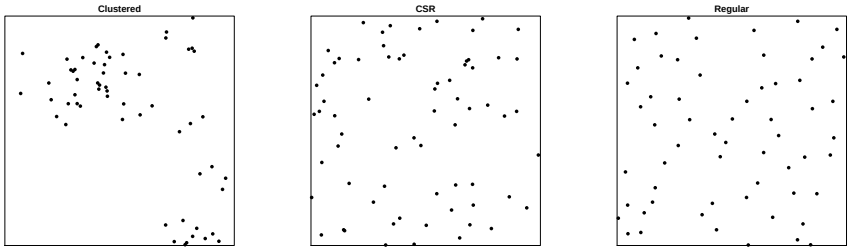


Figure 1: Example of clustered (left), random (middle) and regular (right) point patterns.

Let focus on the random case. In this situation the points have no preference for any spatial location (homogeneity) and the information about the outcome in one region of space has no influence on the outcome in other regions (independence). This situation is known as complete spatial randomness (CSR) and is formally characterise by the following properties:

CSR1 The number of events in a planar region $A \subset W$ follows a Poisson distribution with mean $\lambda|A|$, with $\lambda \in \mathbb{R}^+$.

CSR2 Conditional to $N = n$ events in W, the points $X_1, \ldots, X_n$ are an independent identically distributed (i.i.d.) sample from the uniform distribution in W.

The interest on CSR lies on the basis that it represents an idealised standard which should be tested as a first step in any practical situation. In case we accept this hypothesis the modelisation of the process became almost trivial, see P. J. Diggle (2013) for more explanation on this.

The Poisson point process is one of the most used and studied due to its particularly convenient properties. There are two different types: the homogeneous and the nonhomogeneous. The first one is the simplest existing model and it corresponds exactly to CSR, it is usually denoted by $Pois(\lambda)$ where λ is known as rate or intensity. It satisfies some interesting properties, such as stationarity (the distribution of the point process is invariant under translations) and isotropy (the distribution is invariant under rotations around the origin). The nonhomogeneous Poisson point process is an extension of this one, where the rate value λ is now a function of the location, so $\lambda(x)$ with $x \in W$, and it is used in many practical situations. Formally, the nonhomogeneous Poisson point process is defined as follows:

Definition 3 *Let X be a point process in $W \subset \mathbb{R}^2$ with N its associated counting measure and a spatially varying function $\lambda(x)$. Then we say that X is an nonhomogeneous Poisson point process with intensity $\lambda(x)$ if it verifies:*

NHP1 $N(A)$ has a Poisson distribution with mean $\int_A \lambda(x)dx$ for any $A \subset W$.

NHP2 Conditional to $N = n$, the n events in W form an independent random sample from the distribution in W with density proportional to $\lambda(x)$.

There is another relevant characteristic that usually arise in point process theory, the ***edge effect***. As we have seen, point process are stochastic mechanism that are observed in a bounded region, even though they may be defined in a bigger one. Hence, when studying them we have to take care about it. We will see later in this manuscript some assumptions that allow us to directly avoid this effect, and when this does not happen, how edge correction terms are introduced.

Marks and covariates are two different ways of including some extra information in a point process model, (J. Illian et al., 2008, Chap. 5). The main difference among them is that marks are directly linked to the events, while covariates include information about the whole observation region. This second scenario is where we will work on.

The rest of this paper is structured as follows: in Section 2 we make a brief overview on the existing methods in kernel intensity estimation. In Section 3 we set up a new framework for kernel intensity estimation with covariates and develop asymptotic theory for it; we propose a smooth bootstrap procedure and two new data-driven bandwidth selection methods; a simulation study over all these new procedures and the existing competitors is carried out. Section 4 is devoted to test the goodness-of-fit of the intensity model previously assumed through an L^2 test statistic, which asymptotic normality is proved and it behaviour is analysed in a simulation study. An application to a real data set is detailed in Section 5 and finally we draw some conclusions in Section 6.

2. THE FIRST-ORDER INTENSITY FUNCTION

Modelling the first-order intensity function is one of the main aims in point process theory. This function computes the mean number of events per (length, area or volume) unit, and it is one of the functions that can completely characterizes a point process.

Let X be a point process defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area and let $X_1, \ldots, X_N$ be a realisation of the process. The first-order intensity, from now on referred as intensity, is defined following P. J. Diggle (2013) as:

$$\lambda(x) = \lim_{|dx| \to 0} \frac{\mathbb{E}[N(dx)]}{|dx|},$$

where $|dx|$ denotes the area of an infinitesimal region containing the point $x \in W$.

There is extensive literature on parametric point processes models and intensity estimation under this assumption, see Schoenberg (2005). Assuming a parametric model for the intensity function may be a way of estimating it, using for instance a likelihood score such as the Akaike Information Criteria (AIC), see Van Lieshout (2000), Moller and Waagepetersen (2003) and P. J. Diggle (2013) or pseudolikelihood procedures, see Waagepetersen (2007). In the Bayesian context J. B. Illian, Sørbye, and Rue (2012) proposed some models based on log-gaussian Cox processes. However it is well known that we can obtain unreliable estimates when the assumed parametrization deviates from the true intensity. This is the main reason that supports the use of nonparametric techniques, in particular kernel methodology. P. J. Diggle (1985) proposed the first kernel intensity estimator for one dimensional point process, which have been easily extended to the plane $\mathbb{R}^2$:

$$\hat{\lambda}_H^D(x) = \frac{\sum_{i=1}^N K_H(x - X_i)}{p_H(x)}, \quad x \in W, \tag{1}$$

where H is a bandwidth matrix, $K_H(x) = |H|^{-1/2} K\left(H^{-1/2}x\right)$ and $p_H = \int_W |H|^{-1/2} K(H^{-1/2}(x - y))dy$ is an edge correction term where K denotes a kernel function.

This estimator has been widely used during decades for exploratory analysis, but the inference performed with it has been limited due to its lack of consistency. To overcome this problem, Cucala (2006) defined the "density of events locations" as $\lambda_0(x) = \lambda(x)/m$, where $m = \int_W \lambda(x)dx$ is the expected number of events lying on W. He proposes a kernel estimator:

$$\hat{\lambda}_{0,h}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) 1_{\{N \neq 0\}}, x \in \mathbb{R}$$

with $1_{\{\}}$ denoting the indicator function and h a one-dimensional bandwidth parameter. He proves its consistency under an increasing domain asymptotic framework. Fuentes-Santos, González-Manteiga, and Mateu (2015) extended these ideas to the two-dimensional situation using bandwidth matrices, as it has been done in the context of multivariate density estimation.

Now, let $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ be a spatial continuous covariate that is exactly known in every point of the region of interest W and $Z_1, \ldots, Z_N$ the realisation of the transformed process, i.e, $Z_i = Z(X_i)$. In practice this covariate will commonly be know in an enough amount of points spread over the region, so the values for the rest of the points can be interpolated and it can be assumed that these values are indeed the real ones.

In some cases it can be assumed that spatial point process intensity depends on the covariate, see for instance Baddeley, Chang, Song, and Turner (2012), so it can be written

$$\lambda(u) = \rho(Z(u)),\ u \in W \subset \mathbb{R}^2, \tag{2}$$

where ρ is an unknown function. As Z is known, only ρ needs to be estimated in order to obtain an estimation of λ, which is the target.

To this purpose it is necessary to deal with the transformed univariate point process, $Z(X)$, and establish the theoretical relationship between this one and the original spatial point process X. If X is a Poisson point process in $W \subset \mathbb{R}^2$ with intensity function (2), then $Z(X)$ is a Poisson point process in $\mathbb{R}$ with intensity $\rho g^\star$ and with the same expected number of events, where $g^\star$ is the unnormalized version of the derivative of the spatial cumulative distribution function.

The proposals in Guan (2008) and Baddeley et al. (2012), following the assumption (2), are similar kernel intensity estimators. Guan (2008) develop a kernel estimator based on the definition of the distance between two points by the distance trough their covariates values:

$$\hat{\lambda}_h^G(u) = \frac{\sum_{i=1}^N K_h(||Z(u) - Z(X_i)||)}{q_h(u)},$$

with $q_h(u) = \int_W K_h(||Z(u) - Z(s)||)ds$ the edge correction term.

Considering the increasing domain asymptotic framework and adding also some suitable assumptions, the consistency of his proposal is proved. A bandwidth selection criterion using cross-validation techniques is defined, as well as a dimension reduction tool that allows to handle with high-dimensional covariates.

Baddeley et al. (2012) propose two types of nonparametric intensity estimators based on two nonparametric density estimators, one relies on local likelihood and the other on kernels. We will focus on the last one, in particular on a kernel intensity estimator for the ρ function with a one-dimensional covariate:

$$\hat{\rho}_W(z) = \sum_{i=1}^N \frac{1}{g^\star(Z_i)} K_h(z - Z_i). \tag{3}$$

Note that to obtain the bandwidth parameter h in practice, Baddeley et al. (2012) use the common Silverman's rule-of-thumb for density estimation applied to the transformed data and they do not have any ad-hoc procedure.

3. A CONSISTENT THEORETICAL FRAMEWORK FOR KERNEL INTENSITY ESTIMATION

3.1. New intensity estimation procedure and its asymptotics

In this section we work under the transformed space assuming (2), and the point process obtained from the original one, X, through the covariate, $Z(X)$, defined in the previous section.

Now, following the idea of Cucala (2006) we use the relationship between the intensity and the density function, and we define the following "artificial" density function:

$$f(\cdot) = \frac{\rho(\cdot)g^\star(\cdot)}{m}. \tag{4}$$

What we propose is to take profit of this relationship: firstly estimating the density, f, and then going back to our target problem, that is the estimation of the intensity $\lambda(u) = \rho(Z(u))$.

Following the above notation, we define the estimator of the relative density as follows:

$$\hat{f}_h(z) = g^\star(z)\frac{1}{N}\sum_{i=1}^N \frac{1}{g^\star(Z_i)} K_h\left(z - Z_i\right) 1_{\{N \neq 0\}}, \tag{5}$$

where K is a univariate kernel function and $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. This is an estimator of f, and once we have it, we can go back to the intensity function by plug-in and letting $\hat{\lambda}(u) = \hat{\rho}_h(Z(u))$, where $\hat{\rho}_h$ can be replaced by (3).

In the following statement we obtain the value of the pointwise mean and variance of $\hat{f}_h$ with the corresponding error rates, as well as its mean squared error (MSE), which is defined as $MSE(h,z) = E\left[\left(\hat{f}_h(z) - f(z)\right)^2\right]$.

Hereafter we will establish that our point process X is an nonhomogeneous Poisson point process in $W \subset \mathbb{R}^2$. Although the intensity estimator we propose, as well as the bandwidth selectors, can be applied to non-Poisson processes, the previous assumption is required to prove the consistency of the estimator. We also need to introduce some regularity conditions:

(A.1) $\int_{\mathbb{R}} K(z)dz = 1$; $\int_{\mathbb{R}} zK(z)dz = 0$ and $\mu_2(K) := \int_{\mathbb{R}} z^2K(z)dz < \infty$.

(A.2) $\lim_{m\to\infty} h = 0$ and $\lim_{m\to\infty} \frac{A(m)}{h} = 0$, where $A(m) := \mathbb{E}\left[\frac{1}{N}1_{\{N\neq 0\}}\right]$.

(A.3) G is three times differentiable.

(A.4) z is a continuity point of ρ.

(A.5) ρ is three times differentiable.

Notice that we use an infill structure asymptotic framework, which means that the observation region remains fixed while the sample size increases. In this scenario the bandwidth h is considered as a function of the sample size, this is, $h \equiv h(m)$ and hence it is a sequence of values when $m \to \infty$.

Theorem 1 *Under conditions (A.1) to (A.4) we have that:*

$$E\left[\hat{f}_h(z)\right] = \frac{g^\star(z)(K_h \circ \rho)(z)}{m}\left(1 - e^{-m}\right) \qquad \textit{and}$$
$$Var\left[\hat{f}_h(z)\right] = A(m)\frac{(g^\star(z))^2}{n}\left(K_h^2 \circ \frac{\rho}{g^\star}\right)(z) - (A(m) + e^{-2m} - e^{-m})(g^\star(z))^2(K_h \circ \rho)^2(z),$$

where $\circ$ denotes the convolution between two functions. Moreover, adding condition (A.5) we have:

$$\begin{aligned} MSE(h,z) = e^{-2m}f^2(z) + (1-e^{-m})^2\frac{h^4}{4}\left(\frac{\rho^{''}(z)g^\star(z)}{m}\right)^2\mu_2^2(K) - \\ - e^{-m}(1-e^{-m})h^2\mu_2(K)\frac{(g^\star(z))^2\rho(z)\rho^{''}(z)}{m^2} + \frac{A(m)}{h}f(z)R(K) + \\ + o(h^2(1-e^{-m})e^{-m}) + o(h^4(1-e^{-m})^2) + o\left(\frac{A(m)}{mh}\right), \end{aligned}$$

where $R(K) = \int_{\mathbb{R}} K^2(z)dz$.

Now, defining the mean integrated square error (MISE) as $MISE(h) = E\int\left(\hat{f}_h(z) - f(z)\right)^2 dz$, and denoting by $AMISE$ its asymptotic version, the next result is a consequence of Theorem 1:

Corolary 1 *Under conditions (A.1) to (A.3) and (A.5),*

$$\begin{aligned} MISE(h) = e^{-2m}R(f) + (1-e^{-m})^2\frac{h^4}{4}R\left(\frac{\rho^{''}g^\star}{m}\right)\mu_2^2(K) - \\ - e^{-m}(1-e^{-m})h^2\mu_2(K)\int_{\mathbb{R}}\frac{g^\star(z)\rho^{''}(z)f(z)}{m}dz + \frac{A(m)}{h}R(K) + \\ + o(h^2(1-e^{-m})e^{-m}) + o(h^4(1-e^{-m})^2) + o\left(\frac{A(m)}{mh}\right) \qquad \textit{and} \end{aligned}$$

$$AMISE(h) = (1-e^{-m})^2\frac{h^4}{4}R\left(\frac{\rho^{''}g^\star}{m}\right)\mu_2^2(K) + \frac{A(m)}{h}R(K).$$

As a consequence, the optimal bandwidth value which minimises AMISE is:

$$h_{AMISE} = \left(\frac{A(m)R(K)}{\mu_2^2(K)(1-e^{-m})^2 R\left(\frac{\rho'' g^\star}{m}\right)} \right)^{1/5} = \left(\frac{R(K)}{\mu_2^2(K)R(\rho'' g^\star)} \frac{A(m)}{(1-e^{-m})^2} \right)^{1/5}. \quad (6)$$

The proofs of the results included in this Section 3 can be seen in Borrajo, González-Manteiga, and Martínez-Miranda (2017b).

3.2. Resampling bootstrap method

Another important point in statistical inference are the resampling procedures, which have been widely used in different contexts to perform inference and calibrate the distribution of statistics in goodness-of-fit tests. The smooth bootstrap procedure for point processes with covariates we propose in this section is based on the work of Cao (1993) for kernel density estimation, and Cowling, Hall, and Phillips (1996), for the intensity estimation of a Poisson point process.

Let $X_1, \ldots, X_n$ be a realisation of the spatial point process X, construct $Z_1, \ldots, Z_n$ the associated realisation of the transformed univariate process, let $\hat{f}_g$ be the density estimator in (5) and $\hat{\rho}_g$ the estimator defined in the previous section, where g is a pilot bandwidth.

Now, conditional on $Z_1, \ldots, Z_n$, let $N^* \sim Poiss\,(\hat{m})$ with $\hat{m} := \int_{\mathbb{R}} \hat{\rho}_g(z) g^\star(z) dz$, generate n^* a realisation of this random variable N^* and then draw $Z_1^*, \ldots, Z_{n^*}^*$ by sampling randomly with replacement n^* times from the distribution with density proportional to $g^\star \hat{\rho}_g$, i.e. $\tilde{f}_g = \frac{\hat{\rho}_g g^\star}{\hat{m}}$.

Denote by Z^* the random variable generated by the bootstrap method presented above. From the bootstrap sample define the density estimator as:

$$\hat{f}_h^*(z) = g^\star(z) \frac{1}{N^*} \sum_{i=1}^{N^*} \frac{1}{g^\star(Z_i^*)} K_h\left(z - Z_i^*\right) 1_{\{N^* \neq 0\}}, \quad (7)$$

hence, using equation (4) we get the associated estimator of ρ:

$$\hat{\rho}_h^*(z) = \sum_{i=1}^{N^*} \frac{1}{g^\star(Z_i^*)} K_h\left(z - Z_i^*\right),$$

and then we plug-in it in (2) to obtain an estimator of λ.

The following result provides the expression of the mean, variance and mean squared error of $\hat{f}_h^*$ under the bootstrap distribution; hereafter we use E^*, Var^* and MSE^* to refer to them.

Theorem 2 *Under hypothesis (A.1) to (A.4) we get:*

$$E^*\left[\hat{f}_h^*(z)\right] = \frac{g^\star(z)}{\hat{m}} (K_h \circ \hat{\rho}_g)(z)(1-e^{-\hat{m}}) \qquad \textit{and}$$

$$Var^*\left[\hat{f}_h^*(z)\right] = \frac{(g^\star(z))^2}{\hat{m}} \left(K_h^2 \circ \frac{\hat{\rho}_g}{g^\star} \right)(z) A(\hat{m}) - \frac{(g^\star(z))^2}{\hat{m}^2} (K_h \circ \hat{\rho}_g)^2(z)(A(\hat{m}) + e^{-2\hat{m}} - e^{-\hat{m}}),$$

where $A(\hat{m}) := E^*\left[\frac{1}{N^*} 1_{\{N^* \neq 0\}}\right]$. *Moreover, adding condition (A.5) we have*

$$\begin{aligned} MSE^*(h,z) = {} & e^{-2\hat{m}} (\tilde{f}_g(z))^2 + \frac{h^4}{4} (\hat{\rho}_h''(z))^2 \frac{(g^\star(z))^2}{\hat{m}^2} \mu_2^2(L)(1-e^{-\hat{m}})^2 - \\ & - e^{-\hat{m}}(1-e^{-\hat{m}}) h^2 \tilde{f}_g(z) \frac{\hat{\rho}_g''(z) g^\star(z)}{\hat{m}} \mu_2(L) + \frac{A(\hat{m})}{h} R(L) + o_P(h^4(1-e^{-\hat{m}})^2) + \\ & + o_P(h^2(1-e^{-\hat{m}})e^{-\hat{m}}) + o_P\left(\frac{A(\hat{m})}{\hat{m}h}\right); \end{aligned} \quad (8)$$

remind that we have defined $\tilde{f}_g(\cdot) = \frac{\hat{\rho}_g(\cdot) g^\star(\cdot)}{\hat{m}}$.

In the same way as we have done for Corollary 1, the integrated and asymptotic version of the MSE^* can be easily deduced from the previous result:

Corolary 2 *Under conditions (A.1) to (A.3) and (A.5),*

$$MISE^*(h) = e^{-2\hat{m}} R(\tilde{f}_g) + \frac{h^4}{4} R\left(\frac{\hat{\rho}_g^{''} g^\star}{\hat{m}}\right) \mu_2^2(L)(1-e^{-\hat{m}})^2 -$$
$$- e^{-\hat{m}}(1-e^{-\hat{m}}) h^2 \mu_2(L) \int \frac{\tilde{f}_g(z)\hat{\rho}_g^{''}(z) g^\star(z)}{\hat{m}} dz + \frac{A(\hat{m})}{h} R(L) +$$
$$+ o(h^4(1-e^{-\hat{m}})^2) + o_P(h^2(1-e^{-\hat{m}})e^{-\hat{m}}) + o_P\left(\frac{A(\hat{m})}{\hat{m}h}\right) \qquad \textit{and}$$

$$AMISE^*(h) = \frac{h^4}{4} R\left(\frac{\hat{\rho}_g^{''} g^\star}{\hat{m}}\right) \mu_2^2(L)(1-e^{-\hat{m}})^2 + \frac{A(\hat{m})}{h} R(L).$$

Therefore the asymptotic expression of the optimal bootstrap bandwidth is:

$$h_{AMISE^*} = \left(\frac{A(\hat{m})R(L)}{\mu_2^2(L)(1-e^{-\hat{m}})^2 R\left(\frac{\hat{\rho}_g^{''} g^\star}{\hat{m}}\right)}\right)^{1/5}, \tag{9}$$

which is a plug-in version of (6).

All the results above lead to the following Corollary.

Corolary 3 *Under assumptions (A.1) to (A.4) $MISE^*$ and $AMISE^*$ are consistent estimators of $MISE$ and $AMISE$, respectively.*

3.3. Data-driven bandwidth selection

We have proposed in this work a kernel intensity estimation based on (5), and as it is well known this implies to choose a bandwidth parameter which determines the degree of smoothness to be considered in the estimation. The choice of the bandwidth parameter is crucial and it has motivated several papers in the literature in the recent decades, see for example Marron (1988), Scott (1992) and Silverman (1986) for an earlier full description of the problem. There is a lot of theory developed on this issue in areas of statistics such as density estimation and regression, meanwhile in the context of point processes it has received less attention. P. J. Diggle (1985) proposed a bandwidth selector based on the minimisation of the mean squared error (MSE) of his estimator. Later, P. Diggle and Marron (1988) showed the equivalence, for Cox processes in $\mathbb{R}$, between that procedure and the standard least-squares cross-validation method used in kernel density estimation. This is an example of the strong connection between this two problems, density and intensity estimation. Brooks and Marron (1991) proved the optimality of the least-squares cross-validation bandwidth for one-dimensional nonhomogeneous Poisson point processes. Fuentes-Santos et al. (2015) develop an extension of Cucala's theory to the two-dimensional case, and propose a two-dimensional bandwidth selection procedure based on a bootstrap method.

We describe now two new bandwidth selection methods for the intensity estimation based on (5) that we propose. These methods consist of adaptations of common selectors in the field of density estimation that have not yet been defined nor implemented in the point process framework. We propose a Normal scale rule (rule-of-thumb) and a bootstrap selector derived from the consistent resampling bootstrap procedure detailed in the previous section. All these proposals are based on estimating the infeasible optimal expression (6) where m, $A(m)$ and $\rho^{''}$ are unknown elements.

Rule-of-thumb for bandwidth selection

The basis idea of this method is the same as in Silverman (1986): we assume that the underlying density (4) is Normal, $N(\mu, \sigma)$, with the parameters being estimated from the data, and in this way we replace the unknown values in (6).

In the point processes framework the computation is slightly different from the one used in the context of density estimation, because here the density is only a feature to get the intensity. To begin with, we have to remark that in our context, (6) has some other unknown elements apart

from f, such as m and $A(m)$. The first one is the expected number of points, that in practice can be estimated by the sample size n, and the second one by $1/n$.

The only unknown element left is $\rho^{''}$. However, assuming that $f = \frac{\rho g^\star}{m}$ is Normal, we can derive that

$$\rho''(z) = m\left(\frac{f^{''}(z)}{g^\star(z)} - \frac{2f^{'}(z)(g^\star(z))^{'}}{(g^\star(z))^2} - \frac{f(z)(g^\star(z))^{''}}{(g^\star(z))^2} + \frac{2f(z)(g^\star(z)^{'})^2}{(g^\star(z))^3}\right),$$

and then compute $R\left(\frac{\rho^{''} g^\star}{m}\right)$ using numerical integration methods. Replacing all those estimations in (6) we have the rule-of-thumb bandwidth selector that we will denote by $\hat{h}_{\mathrm{RT}}$.

Bootstrap for bandwidth selection

The asymptotic expression of the optimal bootstrap bandwidth can be considered to derive a consistent bandwidth estimate. Cao (1993) suggested such approach for kernel density estimation with complete data and in Borrajo, González-Manteiga, and Martínez-Miranda (2017a) detail the result for length-biased data, as well as some remarks to extend it to general weighted data.

The expression (9) that we use to build this selector, has a few quantities that need to be computed: $\hat{m}$, $A(\hat{m})$ and $R(\frac{\hat{\rho}_g^{''} g^\star}{\hat{m}})$. The first two can be easily calculated through numerical integration methods such as Simpson's rule, while the last one requires some more development.

The main challenge in the estimation of $R(\frac{\hat{\rho}_g^{''} g^\star}{\hat{m}})$ is to obtain an appropriate value for the pilot bandwidth g. Regarding Cao (1993) and Borrajo et al. (2017a) we can assume that the order of that bandwidth in our context is $m^{-1/7}$, and that the constant has a slight influence on the final result. Hence we propose to use as pilot bandwidth a re-scaled version of the rule-of-thumb previously defined $\hat{g} = (m^{-1/5}/m^{-1/7})\hat{h}_{RT}$; obviously in practice we do not know the value of m, so we use the best approximation we can have which is the sample size of the corresponding realization of the point process.

Then, the bootstrap bandwidth we propose is:

$$\hat{h}_{\mathrm{Boot}} = \left(\frac{A(\hat{m})R(L)}{\mu_2^2(L)(1-e^{-\hat{m}})^2 R\left(\frac{\hat{\rho}_{\hat{g}}^{''} g^\star}{\hat{m}}\right)}\right)^{-1/5}.$$

We perform a simulation study to analyse the behaviour of the methods proposed in this work. Firstly we analyse the performance of the intensity estimator defined in (3), using our two bandwidths selection proposals: the rule of thumb and the bootstrap method. We compare these with the only bandwidth selector that has been previously proposed by Baddeley et al. (2012), which is the common Silverman's rule-of-thumb for density estimation.

An overview of the results of this study indicates that in general, the bootstrap bandwidth seems to perform better than the others in most of the cases, and when this does not occur, our procedure is still competitive. Any of the other methods are not far away from it, even though the rule-of-thumb specifically designed for spatial point processes shows a slightly better behaviour than the Silverman's rule-of-thumb, specially for small sample sizes. In terms of variability, the three compared methods are similar, even though the bootstrap estimates shows in general smaller values.

Looking at the bias of bandwidth estimates, i.e., the relative difference between the selected bandwidth and the benchmark $\hat{h}_{\mathrm{MISE}}$ (the minimiser of the Monte Carlo approximation based on 500 samples of the MISE criterion), the bootstrap bandwidth selector outperforms far more better than the others in most of the studied models. It is also shown that the rule-of-thumb and Silverman's procedures show the bias in the same direction, to be more specific all of them tend to choose smaller bandwidths than the optimal one, while in general the bootstrap selector does the opposite.

To complete our analysis, we have carried out a parallel simulation study to compare our proposals to the competitor described in Guan (2008). For Guan's intensity estimator we have considered two bandwidth choices: the (Monte Carlo approximated) optimal MISE bandwidth, which is considered as a benchmark for this estimator, and the practical least-squares cross-validation

bandwidth proposed in his paper. For our proposal we have considered the intensity estimator with the benchmark $\hat{h}_{\text{MISE}}$ defined above, and our bootstrap bandwidth selector. Our proposal performs considerably better than Guan's for all the simulated models except one. In the best case for the intensity estimators, this is calculating the intensity estimators with the infeasible benchmarks, our estimator achieves smaller relative errors with slightly lower variability. Considering the practical bandwidth choices for each estimator our bootstrap approach clearly beats the cross-validation method. On the other hand, there is one model that seems to be a good scenario for Guan's estimator and his practical cross-validation method, but even in this case our bootstrap proposal is still competitive. The detailed results of this simulation study can be seen in Borrajo et al. (2017b).

4. TESTING THE COVARIATE DEPENDENCE OF THE INTENSITY

Until now we have assumed model (2) being true and we have developed all the methodology under this assumption. Hence, it would be relevant to check whether this model is appropriate; so we will now focus on testing the significance of the covariate.

We formally want to test a null hypothesis $H_0 : \lambda(x) = \rho(Z(x))$, $x \in W$ versus a general alternative in which the intensity function is not explained completely through the covariate. The idea is to define a test statistic based on a L^2-distance between the classical kernel intensity estimator (1) and our proposal in Borrajo et al. (2017b). Due to the lack of consistency of Diggle's proposal we have decide to do a equivalent comparison using the concept of "density of events location" of Cucala (2006) instead of using the intensities; i.e., the null hypothesis can be equivalently rewritten as $H_0 : \lambda_0(x) = \rho(Z(x))/m$, with $\lambda_0(x) = \lambda(x)/m$ and $m = \int_W \lambda(x)dx$. Hence, the test statistic is defined as:

$$T = \int_W \left(\hat{\lambda}_{0,H}(x) - \hat{\rho}_{0,b}(Z(x))\right)^2 dx, \tag{10}$$

where $\hat{\lambda}_{0,H}(x) = \frac{1}{N p_H(x)} \sum_{i=1}^{N} K_H\left(x - X_i\right) 1_{\{N \neq 0\}}$ is the bivariate estimation for the density of events location proposed by Fuentes-Santos et al. (2015), H is a bandwidth matrix, $p_H(x)$ is the edge correction term, $\hat{\rho}_{0,b}(Z(x)) = \frac{\rho_b(x)}{N} 1_{\{N \neq 0\}}$ with $\hat{\rho}_b(x) = \sum_{i=1}^{N} \frac{1}{g^\star(Z(X_i))} L_b\left(Z(x) - Z(X_i)\right)$, b is a bandwidth parameter, K and L are kernel functions, $K_H(u) = |H|^{-1/2} K(H^{-1/2}u)$ and $L_b(u) = \frac{1}{b} L\left(\frac{u}{b}\right)$.

Asymptotic properties and calibration

Hall (1984) proposed a central limit theorem for the integrated square error of multivariate kernel density estimators, which have a similar structure to our test statistic.

Hereafter we will assume that $W = \mathbb{R}^2$ to avoid the edge effects in the theoretical developments, and we introduce some regularity conditions:

(A.6) $\int_{\mathbb{R}} L(z)dz = 1$; $\int_{\mathbb{R}} zL(z)dz = 0$ and $\mu_2(L) := \int_{\mathbb{R}} z^2 L(z)dz < \infty$.

(A.7) $\lim_{m\to\infty} b = 0$ and $\lim_{m\to\infty} \frac{A(m)}{b} = 0$, where $A(m) := \mathbb{E}\left[\frac{1}{N} 1_{\{N \neq 0\}}\right]$.

(A.8) The bandwidth matrix H is symmetric and positive-definite, and such that all entries of H tends to zero, and $m^{-1}|H|^{-1/2} \to 0$ as m increases.

(A.9) K is a continuous, symmetric, square integrable bivariate density function such that $\int_{\mathbb{R}^2} uu^T K(u)du = \mu_2(K) Id_2$ with $\mu_2(K) < \infty$.

(A.10) $Z(x)$ is a continuity point of ρ.

Theorem 3 *Under conditions (A.6) to (A.10) and assuming the null hypothesis $H_0 : \lambda(x)/m = \rho(Z(x))/m \, \forall x \in W$ holds*

$$\frac{T - \mu_T}{\sigma_T} \longrightarrow N(0,1), \text{ where}$$

$$\mu_T = A(m)|H|^{-1/2} R(K) + \frac{1}{2}\mu_2(K) \int \lambda_0(x) tr(HD^2\lambda_0(x))dx + \frac{1}{4}\mu_2^2(K) \int tr^2(HD^2\lambda_0(x))dx,$$

$$\sigma_T^2 = A(m)|H|^{-1/2} \int\int \lambda_0^2(x)\lambda_0(y)(K \circ K)(H^{-1/2}(x-y))dxdy + 2A(m)|H|^{-1/2} R(\lambda_0)R(K),$$

with $tr(\cdot)$ denoting the trace of a matrix.

The proof of this result has been detailed in Borrajo, González-Manteiga, and Martínez-Miranda (2017c).

However, this asymptotic distribution may not be the best way to calibrate our test because it requires some extra estimations, and as the convergence rate may be slow it is not suitable for small patterns. Our proposal to deal with this inaccuracy is to use our bootstrap procedure previously presented in Section 3.

We have used the smooth bootstrap procedure inspired in Cao (1993) and Cowling et al. (1996) to resample under the null hypothesis, hence, let $\hat{\lambda} = \hat{\rho}_t(Z(x))$ with t a pilot bandwidth. Now, conditional on $X_1, \ldots, X_N$ let $N^* \sim Pois\left(\int_W \hat{\rho}_t(Z(x))dx\right)$, generate n^* a realisation of this random variable N^* and then draw $X_1^*, \ldots, X_{n^*}$ by sampling randomly n^* times from the distribution with density proportional to $\hat{\lambda} = \hat{\rho}_t(Z(x))$. Note that Cowling et al. (1996) remarked that the kernel and the bandwidth matrix used in the smooth bootstrap do not need to be the same as those to perform the statistic.

We have conducted a simulation study to analyse the performance of the nonparametric test introduced above. We have checked the normality of T under the null hypothesis and compared the probabilities of rejecting H_0 provided by the asymptotic and bootstrap calibrations. We have also analysed the power of the test for different ways and degrees of departure from the null hypothesis.

We have chosen two different models constructed using real data sets to study the test under the null hypothesis; and the results show that, as expected, the asymptotic normality is not performing well for small sample sizes. Meanwhile, the bootstrap calibration, with the appropriate bandwidth (a reescaled version of the bootstrap bandwidth selector previously presented), shows a good behaviour for the two models and the different sample sizes.

The alternative hypothesis are generated based on the two models under the null, adding a parameter that allow us to control the distance to the null hypothesis and study the behaviour of the test in different situations. Here the values obtained for the power are also high. All the numbers and details of this simulations can be seen in Borrajo et al. (2017c).

5. APPLICATION TO A REAL DATA SET

Here we describe a data analysis to illustrate our proposals in this work. We use the same data set as Baddeley et al. (2012), which is available in Baddeley, Turner, et al. (2005). In this data set the points are the locations of gold deposits observed in a geological survey of the Murchison region of Western Australia, and the covariate is the distance from any point of the observation region to the nearest fault; the observation region is contained in a 330×400 km rectangle. In Figure 2 we can see the 255 locations (left) and the covariate information (right).

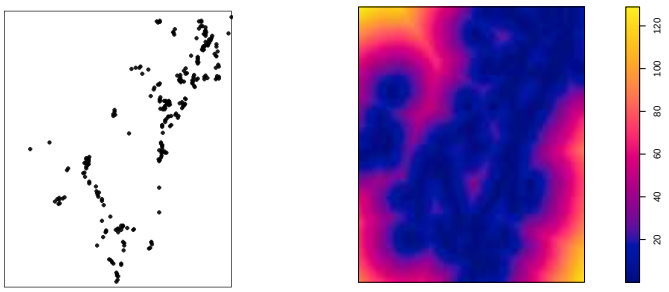
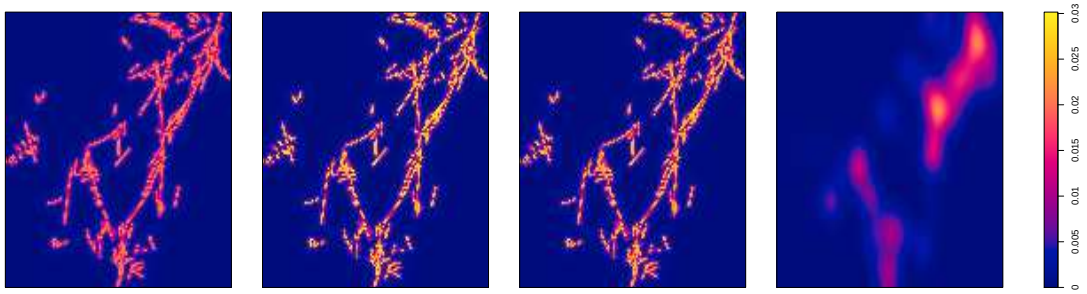


Figure 2: Location of the gold deposits (left) and distance to the faults (right).

First of all we need to compute our test statistic in order to know whether the model (2) is appropriate to describe the intensity. Actually this data set and the covariate has been already used without checking the assumption using a formal statistical test. The reason was that there were no specific methodology for it; this gap is now filled with our proposal. After computing the test we obtain a value of the statistic that allow us to accept the null hypothesis and then all the techniques presented above are applicable (p-values between 0.602 and 0.804 for a suitable range of bandwidth values).

In Figure 3 we perform intensity estimation using the covariate information and the three bandwidth selectors presented in Section 3, as well as Diggle's estimator that ignores the covariate information and uses only the information coming from the locations. We can see that the covariate is indeed useful and with it, we produce better estimations. Moreover if we compare the three bandwidth selectors, our proposals give higher intensity values than Silverman's rule used in Baddeley et al. (2012).



(a) Silverman (b) Rule-of-thumb (c) Bootstrap (d) Diggle

Figure 3: Kernel intensity estimators for the Murchison data set.

We have briefly shown here the possible applications of our methodology and the interest that it might generate in different fields, such as in forestry, where they have several meteorological and geological covariates that may be of interest in analysing point patterns of for example wildfires or species distribution. Another application to wildifires in Canada can be seen in Borrajo et al. (2017b).

6. CONCLUSIONS

In this manuscript we propose new methodology on the field of non-parametric inference for point processes with covariates. We have considered kernel intensity estimation and we have set up a theoretical framework that has allowed us to develop in detail the asymptotic expansions of the MSE, MISE and AMISE of our intensity estimator. Furthermore we have proposed a consistent smooth bootstrap procedure, and two new data-driven bandwidth selection methods. We also propose a new procedure to test if a given covariate is or not of interest for a point process, using the L^2-distance and the estimator previously developed. We detail the asymptotic normality of the test statistic and we use our bootstrap procedure to improve its calibration in practice. All the methodology is supported by extensive simulation studies that show the good behaviour of our proposals in many different situations. Finally we include an application to real data set showing the potential and the strength of the methods presented above.

REFERENCES

Babu, G. J., & Feigelson, E. D. (1996). *Astrostatistics* (Vol. 3). CRC Press.

Baddeley, A., Chang, Y. M., Song, Y., & Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, *5*, 221-236. doi: 10.4310/SII.2012.v5.n2.a7

Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial point patterns: methodology and applications with r*. CRC Press.

Baddeley, A., Turner, R., et al. (2005). Spatstat: an r package for analyzing spatial point patterns. *Journal of statistical software*, *12*(6), 1–42.

Borrajo, M. I., González-Manteiga, W., & Martínez-Miranda, M. D. (2017a). Bandwidth selection for kernel density estimation for length-biased data. *Journal of Nonparametric Statistics (To appear)*.

Borrajo, M. I., González-Manteiga, W., & Martínez-Miranda, M. D. (2017b). Bootrapping kernel intensity estimation for nonhomogeneous point processes depending on spatial covariates. *TEST (2nd review)*. Retrieved from `https://arxiv.org/abs/1703.03213`

Borrajo, M. I., González-Manteiga, W., & Martínez-Miranda, M. D. (2017c). *Testing covariate significante in spatial point processes first-order intensity (unpublished).*

Brooks, M. M., & Marron, J. S. (1991). Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. *Stochastic Processes and their Applications*, *38*(1), 157-165. doi: 10.1016/0304-4149(91)90076-O

Cao, R. (1993). Bootstrapping the mean integrated squared error. *Journal of Multivariate Analysis*, *45*(1), 137–160. doi: 10.1006/jmva.1993.1030

Cowling, A., Hall, P., & Phillips, M. J. (1996). Bootstrap confidence regions for the intensity of a poisson point process. *Journal of the American Statistical Association*, *91*(436), 1516–1524.

Cucala, L. (2006). *Espacements bidimensionnels et données entachés d'erreurs dans l'analyse des procesus ponctuels spatiaux* (Doctoral dissertation, Université des Sciences de Toulouse I). Retrieved from `https://tel.archives-ouvertes.fr/tel-00135890/document`

Daley, D. J., & Vere-Jones, D. (1988). *An introduction to the theory of point processes.* New York: Springer Verlag.

Diggle, P., & Marron, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association*, *83*, 793–800.

Diggle, P. J. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *34*(2), 138–147. doi: 10.2307/2347366

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *153*(3), 349–362. doi: 10.2307/2982977

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns.* CRC Press.

Fuentes-Santos, I., González-Manteiga, W., & Mateu, J. (2015). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. *Scandinavian Journal of Statistics*, *43*(2), 416-435. doi: 10.1111/sjos.12183

Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, *103*(483), 1238–1247. doi: 10.1198/016214508000000526

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, *14*(1), 1–16. doi: 10.1016/0047-259X(84)90044-7

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns.* John Wiley & Sons. doi: 10.1002/9780470725160

Illian, J. B., Møller, J., & Waagepetersen, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, *16*(3), 389–405. doi: 10.1007/s10651-007-0070-8

Illian, J. B., Sørbye, S. H., & Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). *The Annals of Applied Statistics*, 1499–1530.

Marron, J. S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Economics*, *13*(3-4), 187–208.

Moller, J., & Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes.* CRC Press. doi: 10.1002/sim.1896

Ogata, Y., & Zhuang, J. (2006). Space–time etas models and an improved extension. *Tectonophysics*, *413*(1), 13–23. doi: 10.1016/j.tecto.2005.10.016

Schoenberg, F. P. (2005). Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference*, *128*(1), 79–93. doi: 10.1016/j.jspi.2003.09.027

Schoenberg, F. P. (2011). Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, *98*, 789-795.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualisation.* Wiley.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* CRC press.

Stoyan, D., & Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, *15*(1), 61–78. doi: 10.1214/ss/1009212674

Van Lieshout, M. (2000). *Markov point processes and their applications.* World Scientific.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous neyman–scott processes. *Biometrics*, *63*(1), 252–258. doi: 10.1111/j.1541-0420.2006.00667.x

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

ESTUDIO DE LOS INCENDIOS FORESTALES EN GALICIA MEDIANTE UN MODELO MIXTO DE POISSON ESPACIO-TEMPORAL

Miguel Boubeta[1], María José Lombardía[1], Manuel Marey-Pérez[2] y Domingo Morales[3]

[1] MODES, Grupo de Modelización, Optimización e Inferencia Estadística, Universidade da Coruña.
[2] CIO, Centro de Investigación Operativa, Universidad Miguel Hernández de Elche.
[3] GI-1716 PROEPLA, Proxectos e Planificación. Universidade de Santiago de Compostela.

RESUMEN

Los modelos mixtos de Poisson con efectos espacio-temporales constituyen una buena alternativa para modelar variables respuesta de conteo, como por ejemplo el número de incendios forestales, que suponen una importante amenaza para la Comunidad Autónoma de Galicia. Estos modelos incorporan dos conjuntos de efectos aleatorios. El primero, ligado a los dominios, asume una estructura de correlación espacial SAR(1), mientras que el segundo, ligado a la interacción dominio-tiempo asume una estructura de independencia temporal en cada dominio. Bajo este modelo se otiene el estimador del método de los momentos (MM) del vector de parámetros del modelo y el predictor óptimo empírico (EBP) del parámetro objetivo. El rendimiento del estimador propuesto se analiza empíricamente frente a otros estimadores existentes en la literatura tales como el estimador sintético o el plug-in. Finalmente la metodoloía desarrollada se aplica al análisis y modelización del número de incendios forestales en Galicia por área forestal para el período 2007-2008.

Palabras e frases chave: Bootstrap, error cuadrático medio, incendios forestales, predictor óptimo empírico.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

STATISTICS IN "PERCEPTUAL DIALECTOLOGY": FROM THE BEGINNING TO THE FUTURE

Laura Calaza Díaz[1], Rosa M. Crujeiras[2] and Soraya Suárez Quintas[3]

[1] Modelos de Optimización, Decisión, Estatística e Aplicacións. Universidade de Santiago de Compostela.

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

[3] Departamento de Filoloxía Galega. Universidade de Santiago de Compostela.

ABSTRACT

The trajectory of an interdisciplinary project on dialectal perceptions, within the research network TecAnDaLi is reviewed in this paper. Advances from technical, linguistic and statistical perspectives are described as a consequence of the interaction between different disciplines, from data collection to modelling approaches.

Keywords: Spatial data, dialectometry, perceptual dialectology.

1. INTRODUCTION

The goal of this work is to present the trajectory of an ongoing research project conducted within TecAnDaLi (`http://ilg.usc.es/tecandali/`). TecAnDaLi is an interdisciplinary network created in 2012 that involves linguists, statisticians and computer and telecommunication engineers with a common objective: boost the development of technologies and methodologies for statistical analysis in the liguistic field.

Perceptual Dialectology is one of the areas of study where the network is currently focusing on progress. This work presents one of the main lines of research that is being developed which attempts to analyze the perception of Galician dialects by Galician speakers. In order to achieve this goal, the following research groups have been involved: FILGA (Filoloxía e Lingüística Galega, USC), MODESTYA (Modelos de Optimización, Decisión, Estatística e Aplicacións, USC) and COGRADE (Gráficos por Computador e Enxeñaría de Datos, USC).

The outline of this work is as follows. In the first section, a brief summary of the research project is presented, with the following structure: the topic, *what is Perceptual Dialectology?*, the way, *how can a representative sample be choosed?*, and the tools, *how can the data be collected and treated?*. In this sense, the joint work has enabled the design of an adequate framework of study. To highlight the importance of teamwork, in the second section the work schedule followed is presented. Finally, the statistical methodology used along this project is summarized together with the main results obtained so far.

2. PERCEPTUAL DIALECTOLOGY PROJECT

Perceptual Dialectology (PD) is a linguistic discipline that focus on investigating how non-linguists understand dialectal variation. Among other aspects, such as exploring dialectal beliefs and attitudes, this area is interested in studying where non-linguists think dialects come from. For this purpose, this PD project tries to answer the following questions in the Galician context:

- If you listen someone speaking in galician language... Will you be able to identify where his/her speech comes from? What factors could help you to identify its origin?

- Are the perceptual regions similar to the traditional geo-linguistic areas?

The need to know these answers for the overall Galician population as well as to have a tool that could facilitate the data collection, transforms these requirements into a collaborative work. The involvement of the statistical group was essential to define the study design and to take advantage of the output of the application developed. Hereafter, a brief summary is presented. More details can be seen in Calaza Díaz *et al.* (2015).

2.1. Project design

In order to perform this study:

- Six audition tracks from different locations which are representative of the linguistic areas traditionally established were selected. Traditional linguistic areas are represented on the left of Figure 1 with a gradient colour and a dividing line (from left to right: western, central and eastern regions) (Sousa Fernández, 2006). Moreover, a grid dividing the Galician area in 45 cells was ellaborated. In each cell, a total of 4 respondents (born in this region), 2 men and 2 women in different age ranges, between 20-35 and between 50-65 years, were selected (Figure 1, left).
- Different sociodemographic data about respondents were also collected (gender, age, educational level, birthplace and survey place).

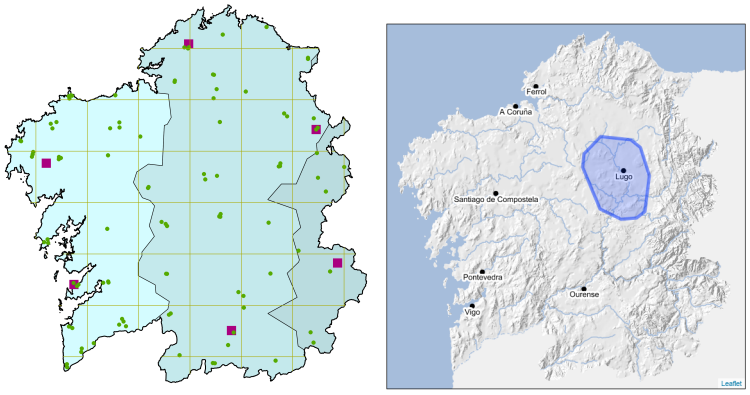


Figure 1: (Left) Representation of sample selection (green points) and audition locations (violet squares). (Right) Example of a geo-referenced region assigned to an audition by a respondent.

2.2. The application

The development of a tool that allows to automatically gather the information of Galician speakers perceptions about the different Galician dialects was crucial to elaborate this study. In particular, in relation with the technical support that offers to perform the perceptual test in which the respondents were submitted to oral stimuli and then they had to identify their geographic origin. In this way, the application transforms these perceptual regions into geo-referenced polygons (Figure 1, right). This makes easy to manage the PD database and facilitates the later statistical analysis. This application was presented in the ExAPP (Suárez Quintas & Calaza Díaz, 2016), an international congress that call together researchers who work on experimental methods to study the perception and production of language variation.

3. RESEARCH STAGES

It was back in the year 2014 when a linguistic dissertation introduced a new challenge for the network: *How can perceptual regions be studied and processed?* In Figure 2, the chronological map that briefly defines the different phases of this project is presented.

The first matter that raised at the beginning was how to describe the characteristics that will define the study. At this point there were two main obstacles:

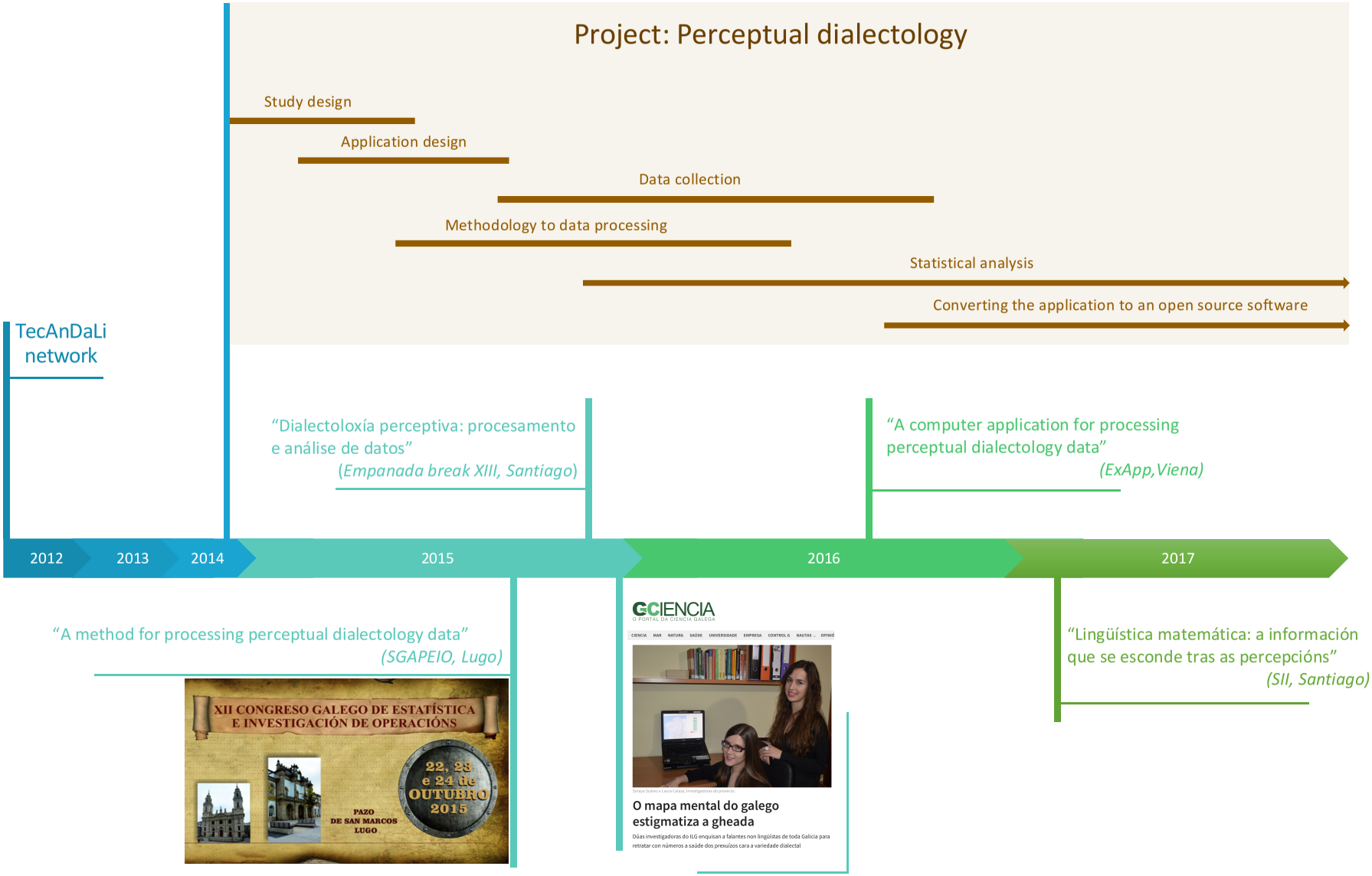


Figure 2: Chronological map of the Perceptual Dialectology project.

- *Must regions have a shape structure?* The way in which the regions were defined will determine the later statistical analysis.
- *How can implement this perceptual test in an application?*

The linguistic answer sounded clear: "*Respondents must have complete freedom to point out their perceptions*". This entailed two parallel working lines, one to implement the application and another one to adapt a methodology to study this kind of situations. This turned up in good future prospects. That was how in Calaza Díaz *et al.* (2015), the first attempts were presented by using fictitious data.

Once the structure was established, the data collection started. This enables to continuously test the application and the methodology and to correct shortcomings in both sides. In this way, in Suárez Quintas and Calaza Díaz (2015) the first results were presented by using the information of about 30 respondents.

In 2016, nearing the end of data collection, the application was presented in the ExAPP, as mentioned before. Currently, the group is working on make this tool flexible to adapt it for other PD studies in different language contexts and regions, and to distribute it as an open source code. At this moment an article describing the functionality and the data structure, along with the link to obtain the base code, is in progress (Suárez Quintas *et al.*, 2017).

After a long and arduous survey task, the database was completed at the end of that year. Therefore, the methodology adapted until this moment was applied to the final database. The results were presented in Calaza Díaz (2017) by transferring the general notions and results. Nowadays, another paper to describe the whole statistical procedure, including the variables definition, the models used and the interpretation of the final results, is in progress.

4. METHODOLOGY AND RESULTS

The spatial character associated with PD data requires the introduction of statistical methodologies that allow to quantify the geographic perception and to analyze the possible influence of different georeferenced information. Along this project, the main objectives are to study how to introduce this kind of spatial information (perceptual regions, birthplace, survey place,...) into regression models to study which factors influence on the identification of Galician speechs' origin, and also to have a method to compare the perceptual regions with the traditional ones.

4.1. Rewound to SGAPEIO 2015

Throughout this section, the main issues developed under each stage are summarized:

SGAPEIO 2015....

As mentioned before, at this point the objective was to introduce the methodology. Then, the main aspects covered were:

- *Perceptual heat maps*
  This technique allows to visualize in a colour gradient map the different perceptual areas defined for each audition. Two programs were considered to perform this representation: QGIS and R project.
- *Assess diferences between sets: Baddeley's distance*
  This lead to measure the discrepances between patterns that are observed in the same region. For instance, measuring the discrepances among perceptual regions and traditional areas, as well as, evaluating the distance between the auditions locations and the corresponding perceptual regions. Thus, let be two sets (regions or locations) $A, B \subseteq X$, being X a certain spatial domain (in our case, Galicia), the Baddeley's distance, Δ_b, is defined as

$$\Delta_b(A,B) = \left[\frac{1}{n(X)} \sum_{x \in X} |d^*(x,A) - d^*(x,B)|^p \right]^{1/p},$$

where $n(X)$ the number of pixels that define X, $d^*(x, A) = \min\{d(x, A), c\} = \min\{\inf[d(x, a), a \in A], c\}$, where $d(x, x')$ is a the distance between two points $x, x' \in X$ and the parameters c and p determine a tradeoff between misclassification error and localisation error respectively. As $c \to 0$, the value of $\frac{1}{c}\Delta_b(A, B)$ converges towards the proportion of different pixels between A and B. The index p controls the relative weight of errors of different magnitudes. We fixed for our analysis $p = 2$, to consider Δ_b as the root-mean-squared error.

- *Adjusted linear regression models*
  The regression models will involve the following variables:
  - Dependent variable, Y: Baddeley's distance between each audition location and the corresponding perceptual regions.
  - Independent variables, X: gender, age, educational level, distance between birthplace and audition place (distab) and distance between birthplace and survey place (distbs).

  Let the index j denote the different auditions ($j = 1, \ldots, 6$) and n the number of respondents. Then, two different proposals were considered:
  - A linear regression model for each audition.
  - A single regression model involving the overall information. Define $Y = \sum_{j=1}^{6} w_j Y_{ji}$, $w_j = \frac{1}{n}, \quad \forall j = 1, \ldots, 6, \quad i = 1, \ldots, n$. Note that here the variable distbs (survey place) was removed.
- *Comparing with traditional areas*
  The discrepancies between perceptual regions and traditional divisions will be measured by using Baddeley's distance. Therefore, the procedure consists of grouping the distances measures of auditions that belong to the same traditional areas, and then use the density estimates to represent which is the best recognized traditional area.

Empanada Break: a seminar at the Instituto da Lingua Galega...

In this step, real data was incorporated to the above procedure. Appart from serving as a starting point for testing the methodology, this communication discusses the first results with linguistic expertises, and, consequently takes into account their opinion for the next steps.

SII: Seminario de Iniciación á Investigación, Institute of Mathematics (USC)...

Currently, the results achieved are presented below:

- *Final heat maps*
  These representations provide a general idea of the behaviour of the respondents' perceptions. The most recognize audition was the Bueu's one, while the most confusing, in terms of the traditional linguistic areas established, was the O Barco's audition. However, a remarkable fact was the case of Mazaricos' audition, which perceptions were divided in two groups, those that selected the Portugal border and those that selected the real area.
- *Directional component*
  Appart from considering the geographical information as distances, the direction in which they are defined must be accounted for. Following the ideas exposed in Pewsey *et al.* (2013), the directional component was incorporated into the regression models. For instance, given the direction of the distance between birthplace and audition location (distab), the angle, α, is defined as the angle of this direction with respect to the horizontal (Figure 4). Then, new variables were defined, $\cos(\alpha)$ and $\sin(\alpha)$, which, together with distab, partially helped to describe the geographical location.
- *Fitted mixed regression models*
  One of the disadvantages of considering the models proposed in Calaza Díaz *et al.* (2015) was that the hierarchical structure associated to the database was not considered. In other words, the relation between the answers given by the same respondent, as well as, the relation between the answers for each audtion were not considered. Therefore, the

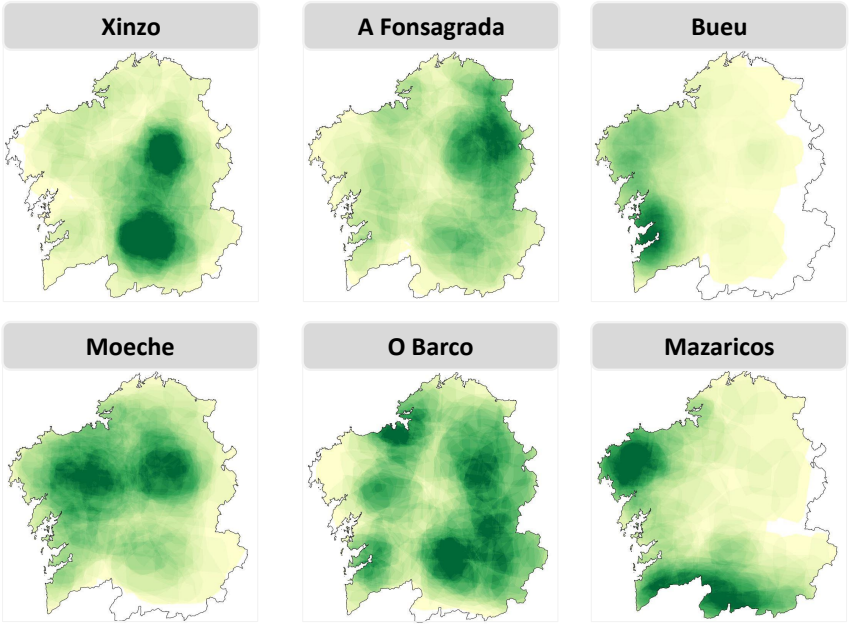


Figure 3: Final heat maps.

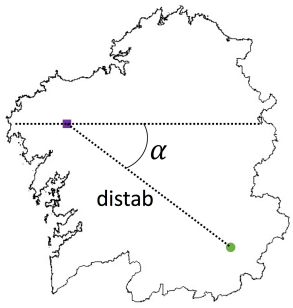


Figure 4: Directional component.

mixed models constitute an adequate tool to model this kind of situations. Given that, the following models were considered:

– one to take into account the variability by respondent (relation among the answers of the same respondent),

$$\begin{cases} Y_{ij} = \beta_0 + \beta X_{ij} + u_i + \varepsilon_{ij}, \\ u_i \in N(0, \sigma_u^2), \varepsilon_{ij} \in N(0, \sigma^2), \text{ independent} \end{cases} \tag{1}$$

– another one to capture the variability by audition (considering the own effects of each audition),

$$\begin{cases} Y_{ij} = \beta_0 + (\beta + b_j) X_{ij} + v_j + \varepsilon_{ij}, \\ v_j \in N(0, \sigma_v^2), \varepsilon_{ij} \in N(0, \sigma^2), \text{ independent.} \end{cases} \tag{2}$$

where Y denotes the response variable; X the fixed effects; β the vector of fixed effects parameters; u the vector of random intercepts related to respondents; v the vector of random intercepts associated with auditions; b the vector of random slopes related to auditions, i the index to denote respondents ($i = 1, \ldots, n$) and j the index to denote auditions ($j = 1, \ldots, 6$). Moreover, the criteria employed to models' selection and fitting were the following: the Restricted Maximum Likelihood method for parameters

estimation and a backward stepwise method for variables selection.

Fitted model (1): The response variable Y, the distance between audition locations and the related perceptual regions; and the explanatory variables, distance between birthplace and audition location (distab) and the corresponding directional component were considered in the resulting fitted model:

$$\widehat{Y}_{ij} = 45696 + \widehat{u}_i + 0.09\,\text{distab}_{ij} - 0.05\,\text{distab}_{ij} \cdot \cos(\alpha_{ij}).$$

Focusing on the significant coefficients associated with the fixed effects it could be seen that the coefficient associated with distab is positive, which means that as the birthplace moves away from the location of the audition it is more difficult to identify. However, the coefficient associated with the interaction between distab and the cosine of the angle α is negative, which means that in the first and fourth quadrants there is a decrease in the response variable (better recognition of dialects), while in the second and third quadrants it results in an increase in the response (worse recognition of dialects). These results can be interpreted as the mobility (in a horizontal sense) of Galician people, from the interior to the west coast, or as evidence that western dialectal varieties are more recognizable. Note that in this case $\widehat{u}_i$ is the parameter that gives the variability in the intercept by informant.

Fitted model (2): In this case, the random slope and the random intercept by audition were considered. The educational level and the distance between birthplace and audition location (distab) were included in the final fitted model. Therefore, given that the educational level was an ordinal variable, the model was defined by different intercepts depending on its value (primary, secondary or higher education). As a result, the resulting fitted model was the following:

Primary $\widehat{Y}_{ij} = 48602 + \widehat{v}_j + (0.09 + \widehat{b}_j)\text{distab}_{ij}$
Secondary $\widehat{Y}_{ij} = 48602 - 1743 + \widehat{v}_j + (0.09 + \widehat{b}_j)\text{distab}_{ij}$
Higher education $\widehat{Y}_{ij} = 48602 - 3863 + \widehat{v}_j + (0.09 + \widehat{b}_j)\text{distab}_{ij}$

Note that considering distab fixed, as educational level increases the response decreases (respondents recognise better dialects) and, as far as the birthplaces are the higher the response is (that means, more diffficult is to recognised dialects). In order to better visualize the way in which the random intercepts $\widehat{v}_j$ and the random slopes $\widehat{b}_j$ relax the model, the Figure 5 is shown. In this figure the model associated with each audition is plotted separately, as a result a total of six columns were represented. Inside of each one, the model is, at the same time, divided into the three levels of the educational level variable. Looking carefully, it can be clearly seen that the Bueu's audition is the best recognized (small intercept and small slope), whereas the Moeche's audition is the worst recognized (the biggest intercept) independently of the birthplace (slight slope). The rest of the auditions (Xinzo, A Fonsagrada, O Barco and Mazaricos) highlight the importance of the closeness of the birthplace in dialect recognition.

4.2. Future lines

In the near future, the objective is to focalize the efforts in deeper insight into model validation and model assessment. Moreover, one of the lines of research to be developed is to study the (spatial) correlation between sets to avoid the current simplification of the (spatial) database. Also, it will be interesting to study the possible correlation between informants. In particular, one of the main points of interest is to study the behaviour of the respondents' perceptions of the Mazaricos' audition.

REFERENCES

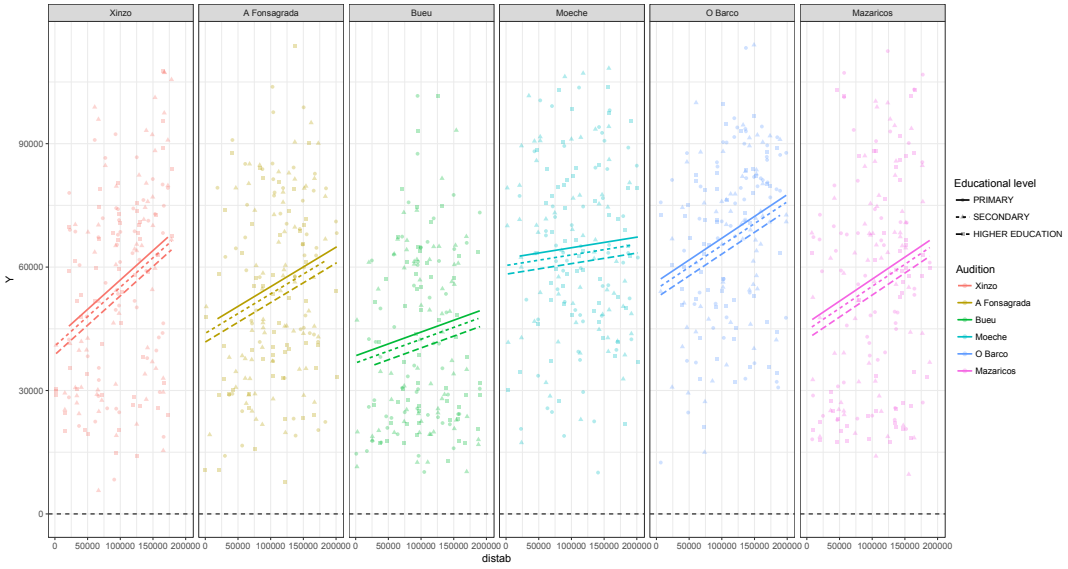


Figure 5: Fitted mixed model considering random intercept and slope by audition.

Calaza Díaz, L.; Suárez Quintas, S.; Crujeiras R.M.; Rodríguez Casal, A.; Sousa, X. and Ríos Viqueira, J.R. (2015) A method for processing perceptual dialectology data, in ACTAS XII Congreso Galego de Estatística e Investigación de Operacións. Lugo, 22-24 October 2015.

Calaza Díaz, L. (2017) Lingüística matemática: a información que se agocha tras as percepcións. Seminario de Iniciación á investigación. Santiago de Compostela, 2017.

Suárez Quintas, S. and Calaza Díaz, L. (2015) Dialectoloxía perceptiva: procesamento e análise de datos. Empanada break XIII, Santiago.

Suárez Quintas, S. and Calaza Díaz, L. (2016) A computer application for processing perceptual dialectology data. Experimental Approaches to Perception and Production of Language Variation, Viena, 2016.

Suárez Quintas, S.; Sousa, X.; Calaza Díaz, L. and Crujeiras R.M. (2017) PDMapping: documenting and analyzing perceptions about dialects with a GIS-based application. Article in preparation.

Pewsey, A., Neuhuser, M. and Ruxton G. D. (2013) Circular Statistics in R. Oxford University Press, Oxford.

Sousa Fernández, X. S. (2006) Análise dialectométrica das variedades xeolingüísticas galegas. Encontro de estudos dialectolóxicos. Actas, M C. Rola Bernardo / H. Mateus Montenegro, Ponta Delgada: Instituto Cultural de Ponta Delgada, 345–362 - Capítulo de libro.

XII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 y 28 de octubre de 2017

UN PROBLEMA DE FLOTA DE VEHÍCULOS CON REPLANIFICACIONES PERIÓDICAS

Luisa Carpente[1], Julián Costa[1], Manuel Fontenla[2], Ignacio García-Jurado[1], Silvia Lorenzo-Freire[1] y Guido Novoa[1]

[1] Grupo MODES, Departamento de Matemáticas, Universidade da Coruña.
[2] Departamento de Logística. INDITEX.

RESUMEN

En este trabajo nos centramos en el problema de gestión de la flota de camiones de la empresa GESUGA (Gestora de Subproductos de la comunidad autónoma de Galicia). Uno de los problemas con el que se encuentra diariamente esta empresa es el de diseñar de manera eficaz las rutas de recogida del día siguiente, teniendo en cuenta las solicitudes que han quedado sin atender durante esta jornada. El otro problema, es el de replanificar adecuadamente esas rutas durante la jornada posterior (ya que se van incorporando un número muy importante de nuevas solicitudes), intentando incorporar algunas de estas nuevas demandas. Para la resolución del problema planteado se han combinado modelos de programación lineal y técnicas de simulación con heurísticas asociadas al problema de VRPTW (Vehicle Routing Problem with Time Windows) adaptadas de manera conveniente a este contexto.

Palabras y frases clave: Programación lineal, algoritmos heurísticos, VRPTW.

1. INTRODUCCIÓN

GESUGA es una empresa fundada en el año 2005 que centra su área de negocio en la gestión integral de los subproductos cárnicos no destinados al consumo humano. Su actividad principal consiste en la recogida y el transporte de los distintos subproductos, generalmente cadáveres de animales, desde las explotaciones ganaderas hasta un punto de tratamiento, para su valoración y posterior destrucción o transformación, generando pieles y grasa animal para su comercialización, según el posible destino de cada producto o residuo.

El Reglamento CE 999/2001 prohíbe el enterramiento y la incineración de subproductos o residuos derivados de animales muertos en las explotaciones ganaderas. Esta medida fue tomada para evitar la propagación de posibles enfermedades contraídas por los animales, sobre todo propiciadas por la encefalopatía espongiforme bovina (conocida comúnmente como la enfermedad de las vacas locas). Desde que este Reglamento entró en vigor, la actividad de GESUGA ha experimentado un gran crecimiento.

Agroseguro, Sistema Español de Seguros Agrarios, se fundó con el objetivo de establecer una cobertura técnica y financiera que permita al sector agrario hacer frente a posibles daños o riesgos imprevistos. Agroseguro gestiona las pólizas anuales de sus asegurados montando así un servicio público de recogida en las explotaciones. De este modo, a través de Agroseguro, los propios ganaderos realizan la solicitud de recogida cuando se encuentran con un animal muerto en su explotación o alrededores y, es Agroseguro el responsable de trasladar la petición a GESUGA para que se encargue de realizar la recogida, transporte y posterior tratamiento. En la actualidad GESUGA para hacer frente a su actividad dispone de un total de 70 trabajadores y unas instalaciones compuestas por las oficinas centrales, dos plantas intermedias y una planta principal de transformación.

La planta principal se encuentra en Cerceda (A Coruña), junto a las instalaciones centrales. En esta planta se recibe diariamente toda la mercancía y se valora realizando una clasificación antes de su tratamiento siguiendo la orden PRE/2493/2011 de 16 de septiembre, por la que se establece y

regula el Programa Integral coordinado de vigilancia y control de las encefalopatías espongiformes transmisibles de los animales. Las otras dos plantas se encuentran en Outeiro de Rei (Lugo) y Vilamarín (Ourense) y sirven de base para las recogidas que se realizan en las provincias de Lugo, Ourense y Pontevedra. Estas plantas además de ayudar a mejorar la logística de recogida también ayudan a garantizar la bioseguridad de este proceso, puesto que cada camión lleva su mercancía a la planta intermedia más cercana, reduciendo así el número de camiones que se trasladan a la planta de transformación de Cerceda, lo que implica también una disminución del riesgo de contaminación. Estas tres plantas están ubicadas estratégicamente para intentar cubrir la mayor parte de terreno de Galicia de una forma equitativa lo que le permite gestionar más eficientemente su actividad a lo largo de toda la Comunidad.

Su flota consta de 36 vehículos que gestionan y realizan alrededor de unas 500 recogidas por día. Además cuenta con 3 cabezas tractoras, y 9 remolques bañera, que son los encargados de transportar diariamente el material desde las plantas intermedias hasta la planta de transformación ya que, una vez que las recogidas están en la planta intermedia, todos los camiones son pesados y su mercancía es trasladada a un camión tráiler con bañera, que es el encargado de llevarla a la planta principal para su procesado. El ciclo se cierra con el lavado y desinfección de cada vehículo para que esté listo para realizar de nuevo, el próximo día, el mismo proceso.

En la actualidad, la eficiencia de esta recogida es muy dependiente del conocimiento que cada conductor tiene de la zona. Esa es una restricción que deseamos eliminar. Para abordar el diseño de un sistema de toma decisiones automático que planifique las recogidas y diseñe las rutas de acuerdo con las solicitudes, hemos estudiado el procedimiento real que se sigue a día de hoy y se realiza de manera totalmente manual.

A continuación, en la Sección 2, describimos brevemente el problema y las técnicas de solución abordadas.

2. DESCRIPCIÓN Y SOLUCIÓN DEL PROBLEMA

Para entender el problema, a continuación describiremos brevemente el funcionamiento diario de la empresa en cuanto a la logística del diseño de rutas:

1. Se reciben avisos durante el día, principalmente a través de Agroseguro (algunos por otros medios, pero con la misma información). El aviso contiene datos sobre la explotación, incluyendo coordenadas, provincia y ayuntamiento, así como la especie de animal a recoger (sin peso especificado).

2. Con los avisos que tienen a las 19:00 de la tarde, y que no se han recogido durante esa jornada, dos empleados hacen unas pre-rutas para el día siguiente. La planificación inicial es de la manera siguiente: tienen determinadas unas rutas por ayuntamientos que han ido elaborando en base a su histórico y los avisos que vengan de un determinado ayuntamiento se asignan directamente a su ruta. Cada ruta se corresponde con un camión de los que tienen disponibles en cada planta: 10 en Lugo, 10 en Orense y 12 en A Coruña. Esto es lo que vamos a tratar como avisos previos a un día de trabajo.

3. A las 20:00, ahora ya con todos los avisos que han llegado hasta esa hora, se revisan las rutas manualmente con el jefe de logística y se hacen los cambios que consideren oportunos. La clave parece estar en tener rutas más o menos equilibradas y cumplir con el tiempo máximo de ruta del que disponen, teniendo en cuenta que cada camión y cada conductor sólo hacen una ruta por día. De esta forma, el tiempo máximo viene determinado por la jornada laboral de los conductores (unas 9 horas) en la que se deben incluir los descansos reglamentarios y el tiempo de limpieza del vehículo. Para estimar cuanto tiempo necesita cada ruta en horas, se suman los tiempos de viaje, los de parada en cada punto de recogida, los descansos y el tiempo de limpieza del camión al volver a la correspondiente base. A las 21:00 cierran todas las rutas y se las envían a los conductores.

4. A cada conductor le llega una serie de puntos que deben de recoger. Así, cada conductor organiza la ruta por su cuenta, al margen de los cálculos que hayan hecho los de logística de GESUGA.

5. Los avisos que llegan durante la jornada los van asignando para librar trabajo para el día siguiente. La asignación es totalmente manual por parte de los de logística. Partiendo de la ruta que les asigna el programa en función del ayuntamiento, revisan el mapa para ver por dónde está cada camión y llaman a los conductores para ver si alguno puede acercarse a recogerlo. Si no consiguen colocarlo dejan el aviso para el día siguiente.

El problema base que presenta GESUGA pertenece a la categoría de MDVRPTW (Multiple Depot Vehicle Routing Problem with Time Windows) ya que la flota de camiones está distribuida en tres plantas (Orense, Lugo y A Coruña) con algunas particularidades adicionales. Dado el gran número de avisos a tratar diariamente, resolver un problema exacto en todas las fases se hace inviable.

A continuación se presenta el esquema de la solución que hemos planteado para este problema.

- Para planificar las rutas del día siguiente, se consideran todos los avisos previos y se les añaden otros simulados (en función de la probabilidad de que aparezcan al día siguiente). Esto genera rutas con holgura que facilitan la incorporación de nuevos puntos en la jornada siguiente. Véanse la Figura 1 y la Figura 2.

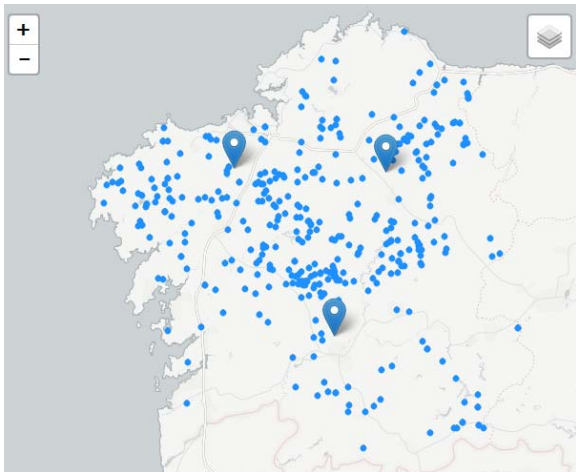


Figura 1: Avisos previos.

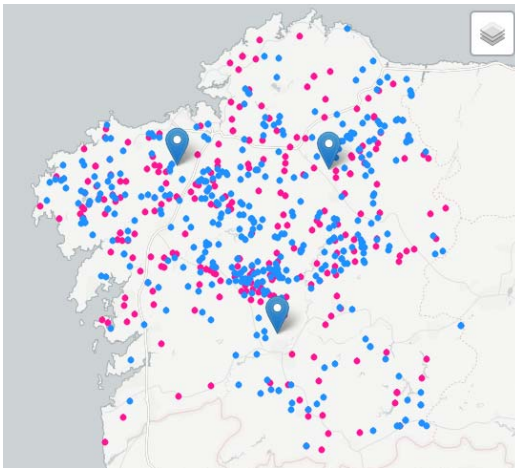


Figura 2: Avisos totales: previos y simulados.

- Teniendo en cuenta los avisos totales, se resuelve un problema de asignación generalizada (GAP). De tal forma que, a partir de aquí, se consideran tres problemas VRPTW, uno por cada planta. Véase la Figura 3.

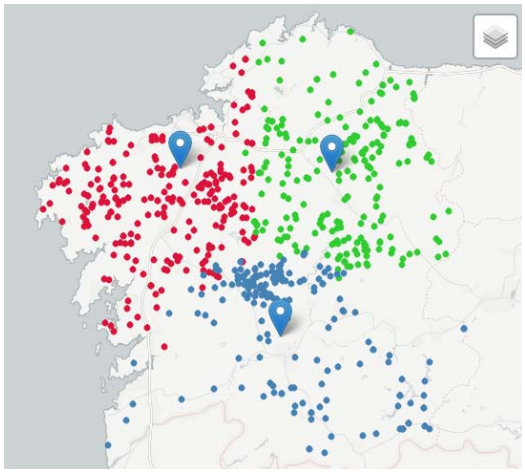


Figura 3: Avisos asignados a cada planta.

- Se aplican varios algoritmos heurísticos para obtener las rutas del día siguiente. Véase la Figura 4.

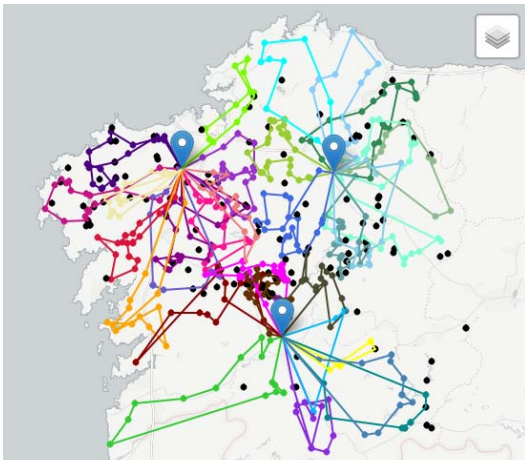


Figura 4: Rutas resultante

- Al día siguiente, dado que la información de lo que ha hecho cada camión y su posición se conocen en todo momento, se aplican algoritmos heurísticos que replanifican las rutas que quedan por hacer, teniendo en cuenta nuevas solicitudes.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad a través del proyecto MTM2014-53395-C3-1-P, por el Centro para el Desarrollo Tecnológico Industrial a través del proyecto ITC-20151247, y por la Xunta de Galicia mediante las ayudas para Grupos de Referencia Competitiva ED431C-2016-015 y Centro Singular de Investigación de Galicia ED431G/01.

REFERENCIAS

Irnich, S., Toth, P. y Vigo, D.(2014) Vehicle Routing: Problems, Methods and Applications. P. Toth and D. Vigo, eds., MOS-SIAM.

Laporte, G.(1992) The vehicle routing problem: An overview of exact and approximate algorithms. European Journal of Operational Research, vol. 59, 3, 345–358.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

A PLUG-IN BANDWIDTH SELECTOR FOR LOCAL LINEAL QUANTILE REGRESSION

Mercedes Conde-Amboage[1] and César Sánchez-Sellero[2]

[1] Research Centre for Operations Research and Business Statistics (ORSTAT)
University of Leuven

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

ABSTRACT

In the framework of quantile regression, local linear smoothing techniques have been studied by several authors. The problem of bandwidth selection was addressed in the literature by the usual approaches, such as cross-validation or plug-in methods. Most of the plug-in methods rely on restrictive assumptions on the quantile regression model in relation to the mean regression, or on parametric assumptions. Here we present a plug-in bandwidth selector for nonparametric quantile regression, that is defined from a completely nonparametric approach. To this end, the curvature of the quantile regression function and the integrated sparsity (inverse of the conditional density) are both nonparametrically estimated. The new bandwidth selector is shown to work well in different scenarios and in real data applications.

Keywords: Quantile regression; Bandwidth selection; Nonparametric regression.

1. INTRODUCTION

Although mean regression is still a traditional benchmark in regression studies, the quantile approach is receiving increasing attention, because it allows a more complete description of the conditional distribution of the response given the covariate, and it is more robust to deviations from error normality.

The quantile regression model could be stated as $Y = q_\tau(X) + \varepsilon$ where Y is the response variable of interest, X is the covariate, q_τ is the quantile regression function of order τ and ε represents the error. Then, the conditional τ-quantile of ε given X will be zero. Estimation of the quantile regression model exploits the fact that the conditional quantile, $q_\tau(x)$, is the value a that minimizes the expectation $\mathbb{E}[\rho_\tau(Y-a)|X=x]$, where $\rho_\tau(u) = u(\tau - \mathbb{I}(u<0))$ and $\mathbb{I}(\cdot)$ is the indicator function of an event. Koenker and Bassett (1978) can be considered a seminal work in estimating conditional quantiles in a parametric setup following this idea.

Along this work, we will focus on the univariate regression context, that is, $(X,Y) \in \mathbb{R}^2$. In this line, Yu and Jones (1998) studied a local linear estimator of the quantile regression in a nonparametric framework. To this end, a random sample of independent observations $(X_1,Y_1),\ldots,(X_n,Y_n)$ of the pair (X,Y) is supposed to be available. Then, the estimator will be $\widehat{q}_{\tau,h}(x) = \widehat{a}$, where $\widehat{a}$ and $\widehat{b}$ are the minimizers of

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - a - b(X_i - x)\right) K\left(\frac{X_i - x}{h_\tau}\right),$$

where K is a kernel function and h_τ is a bandwidth parameter.

Several authors have addressed the problem of bandwidth selection as Yu and Jones (1998), Abberger (1998) or Jones and Yu (2007). In this work a plug-in rule is designed to choose the bandwidth parameter, h_τ. The plug-in technique consists of minimizing the dominant terms of the

mean integrated squared error (MISE) of the estimator. For the local linear quantile regression, it can be written as (see Fan et al. (1994) and Yu and Jones (1998))

$$\begin{aligned} \text{MISE}\,(\widehat{q}_{\tau,h_\tau}) &= \mathbb{E}\int \left(\widehat{q}_{\tau,h_\tau}(x) - q_\tau(x)\right)^2 g(x)\,dx \\ &\cong \frac{1}{4}h_\tau^4 \mu_2(K)^2 \int q_\tau^{(2)}(x)^2 g(x)\,dx + \frac{R(K)\tau(1-\tau)}{nh_\tau}\int \frac{1}{f(q_\tau(x)|x)^2}\,dx \end{aligned} \tag{1}$$

where g is the density of X, $f(q_\tau(x)|x)$ is the conditional density of Y at $q_\tau(x)$ given $X = x$, $q_\tau^{(2)}$ is the second derivative of q_τ, $\mu_2(K) = \int u^2 K(u)\,du$ and $R(K) = \int K^2(u)\,du$.

An asymptotically optimal bandwidth can be derived as

$$h_{\text{AMISE},\tau} = \left[\frac{R(K)\ \tau(1-\tau)}{n\ \mu_2(K)^2\ \int q_\tau^{(2)}(x)^2\ g(x)\ dx}\ \int \frac{1}{f(q_\tau(x)|x)^2}\ dx\right]^{1/5}. \tag{2}$$

Note that $\mu_2(K)$ and $R(K)$ are obtained from the kernel function, while the two integrals in (2) are unknown and have to be estimated. Expression (2) is quite similar to the plug-in rule for mean regression. The curvature (integrated squared second derivative) is now calculated for the quantile regression function instead of the mean regression, while the integrated squared sparsity (where "sparsity" means the inverse of the conditional density) replaces the integrated conditional variance that appeared in mean regression.

Because of these similarities with mean regression, Yu and Jones (1998) proposed to use Ruppert et al. (1995) bandwidth with some simple transformations based on the assumptions of homoscedasticity and error normality. Homoscedasticity is useful to have the same curvature for any τ as in mean regression, while normality allows to estimate the sparsity from the conditional variance.

The purpose of this work is to provide a plug-in bandwidth for local linear quantile regression without imposing restrictions on the conditional variability and the error distribution. Instead, nonparametric estimations of the curvature at the given quantile τ will be used, as well as nonparametric estimations of the sparsity.

Other proposals in literature for bandwidth selection in nonparametric quantile regression were given, based on cross-validation techniques. In particular, Abberger (1998) proposed to minimize in h the cross-validation function given by:

$$CV(h) = \sum_{i=1}^{n} \rho_\tau\left(Y_i - \widehat{q}_{\tau,h}^{-i}(X_i)\right) \tag{3}$$

where $\widehat{q}_{\tau,h}^{-i}(X_i)$ is the classical leave-one-out estimator.

In Section 2 a preliminary rule of thumb is obtained, and afterwards the proposed plug-in rule is derived. In Section 3 mean squared errors of curvature and sparsity estimators are derived. In Section 4, a simulation study and an application to real data are given to explore the virtues of the new bandwidth selectors in comparison with natural competitors. Section 5 contains the conclusions.

2. PROPOSED SELECTORS

As any plug-in rule, the crucial ingredients of our proposed selectors will be the estimators of unknown quantities, which in our case are curvature and sparsity. Our first proposal will consist of a rule of thumb, where the estimators are defined on a simple partition of the sample in blocks.

2.1. Rule of thumb

Following the ideas in Ruppert et al. (1995), a rule of thumb can be constructed by doing the next steps:

1. Partition the range of X into N blocks with the same number of observations. The original sample $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ is subsequently split into the N blocks. A polynomial of order 4 is adjusted at each block, then providing N polynomials that will be denoted by $\widehat{q}_{\tau,j}$ with $j = 1, \ldots, N$. The number of blocks will be chosen as $\widehat{N}$ following the Mallows's C_p criterion (see Mallows (1973)) adapted to the quantile frameworks, that is, $\widehat{N}$ will minimize

$$C_p(N) = \frac{\mathrm{RSQ}(N)}{\mathrm{RSQ}(N_{\max})/(n - 5N_{\max})} - (n - 10N)$$

where $\mathrm{RSQ}(N)$ is the residual sum of quantile losses given by ρ_τ and summed over each blocked quartic fit, when the number of blocks is N, $N_{\max} = \max\{\min([n/20], N^*), 1\}$ and $N^* = 5$. Here $[\cdot]$ denotes the integer part of a number.

Remark 1 *Following the ideas of Ruppert et al. (1995) in the mean regression context, we have considered polynomials of order 4 because it is the lowest degree polynomial that admits estimates of the quantity* $\vartheta_{24} = \int q_\tau^{(2)}(x)\, q_\tau^{(4)}(x)\, g(x)\, dx$ *other than zero. This integral* ϑ_{24} *is involved in the mean squared error of curvature estimator, see equation* (5).

2. If we denote by $\vartheta = \int q_\tau^{(2)}(x)^2\, g(x)\, dx$, the following curvature estimator is considered:

$$\widehat{\vartheta}_{\mathrm{B}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\widehat{N}} \widehat{q}_{\tau,j}^{(2)}(X_i)^2\, I\,(X_i \in \text{Block } j).$$

3. Estimate the sparsity at each block j by means of

$$\widehat{s}_j(\tau) = \frac{r_{[\tau+d_j]} - r_{[\tau-d_j]}}{2\, d_j}$$

where $r_{[\tau-d_j]}$ and $r_{[\tau+d_j]}$ are the sample quantiles of orders $(\tau - d_j)$ and $(\tau + d_j)$, respectively, of the residuals from the quartic fit at block j. This type of sparsity estimator was studied by Bloch and Gastwirth (1968). For the parameter d_j, the selector proposed by Bofinger (1975) will be used here. Finally, the integrated squared sparsity will be estimated by

$$\widehat{s}_{\mathrm{B}}^2 = \sum_{j=1}^{\widehat{N}} \widehat{s}_j^2\, l_j$$

where l_j denotes the length of block j.

4. Finally, the selector from the rule of thumb will be obtained as:

$$\widehat{h}_{\mathrm{RT}} = \left(\frac{R(K)\, \tau(1-\tau)\, \widehat{s}_{\mathrm{B}}^2}{n\, \mu_2(K)^2\, \widehat{\vartheta}_{\mathrm{B}}} \right)^{1/5}.$$

2.2. Plug-in rule

The plug-in rule will come from a more elaborated estimation of the curvature and the sparsity.

Curvature estimation: Now the second derivative of the regression will be nonparametrically estimated at each sample observation. In order to do this, a local polynomial of order 3 will be adjusted. Let us call $\widetilde{q}_{\tau,h_c}^{(2)}(X_i)$ to its second derivative at X_i, for $i = 1, \ldots, n$. Then, we can consider the following curvature estimator:

$$\widehat{\vartheta}_{h_c} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{q}_{\tau,h_c}^{(2)}(X_i)^2 \tag{4}$$

At this point, a pilot bandwidth h_c for curvature estimation should be selected. The criterion for selecting h_c will be the mean squared error of the curvature estimator. As in the case of classical mean regression, see Ruppert et al. (1995), the variance of the curvature estimator will be asymptotically negligible compared with its asymptotic squared bias. So, the asymptotic mean squared error coincides with the asymptotic squared bias, which is given by

$$\text{MSE}\left(\widehat{\vartheta}_{h_c}\right) \cong \left[\delta_1\ h_c^2\ \int q_\tau^{(2)}(x) q_\tau^{(4)}(x) g(x) dx\ + \delta_2\ \tau(1-\tau)\ \frac{1}{n h_c^5}\ \int \frac{1}{f(q_\tau(x)|x)^2} dx\right]^2 \tag{5}$$

where $\mu_i(K) = \int v^i K(v) dv$ and

$$\begin{aligned}
\delta_1 &= \frac{1}{6}(\alpha_{31}\mu_4(K) + \alpha_{33}\mu_6(K)) \\
\delta_2 &= 4\left(\alpha_{31}^2 \int K^2(v)dv + \alpha_{33}^2 \int v^4 K^2(v)dv + 2\alpha_{31}\alpha_{33}\int v^2 K^2(v)dv\right) \\
\alpha_{31} &= \frac{-\mu_2(K)^2\mu_6(K) + \mu_2(K)\mu_4(K)^2}{\mu_2(K)\mu_4(K)\mu_6(K) - \mu_4(K)^3 - \mu_2(K)^3\mu_6(K) + \mu_2(K)^2\mu_4(K)^2} \\
\alpha_{33} &= \frac{\mu_2(K)\mu_6(K) - \mu_4(K)^2}{\mu_2(K)\mu_4(K)\mu_6(K) - \mu_4(K)^3 - \mu_2(K)^3\mu_6(K) + \mu_2(K)^2\mu_4(K)^2}
\end{aligned}$$

Minimizing the last expression, the pilot bandwidth will be

$$h_c = C(K)\ \left(\frac{\tau(1-\tau)\int 1/f(q_\tau(x)|x)^2 dx}{|\int q_\tau^{(2)}(x) q_\tau^{(4)}(x) g(x) dx|\ n}\right)^{1/7}$$

where

$$C(K) = \begin{cases} C^I = \left(\frac{5\delta_2}{2\delta_1}\right)^{1/7} & \text{if } \int q_\tau^{(2)}(x) q_\tau^{(4)}(x) g(x) dx > 0 \\ C^{II} = \left(\frac{\delta_2}{\delta_1}\right)^{1/7} & \text{if } \int q_\tau^{(2)}(x) q_\tau^{(4)}(x) dx < 0 \end{cases}$$

To compute this pilot bandwidth, preliminary estimations of the integrated squared sparsity and the integral $\vartheta_{24} = \int q_\tau^{(2)}(x) q_\tau^{(4)}(x) g(x) dx$ are required. They will be obtained from blocked estimators as those considered for the rule of thumb. They can be denoted by $\widehat{s}_{\text{B}}^2$ and $\widehat{\vartheta}_{24,\text{B}}$. The resulting estimated pilot bandwidth will be

$$\widehat{h}_c = C(K)\ \left(\frac{\tau(1-\tau)\ \widehat{s}_{\text{B}}^2}{|\widehat{\vartheta}_{24,\text{B}}|\ n}\right)^{1/7}.$$

Finally, the curvature estimator will be given by

$$\widehat{\vartheta}_{\widehat{h}_c}^{\alpha} = \frac{1}{n}\ \sum_{i=1}^{n} \widehat{q}_{\tau,\widehat{h}_c}^{(2)}(X_i)^2\ \mathbb{I}\left((1-\alpha)a + \alpha b < X_i < \alpha a + (1-\alpha)b\right)$$

where the sample was trimmed at each border a and b, by a small proportion $\alpha \in [0,1]$, assuming that the covariate is supported in the interval $[a,b]$. This strategy was already used by Ruppert et al. (1995) in their estimation of similar quantities for mean regression. It is intended to prevent from the variability of local polynomial kernel estimates of high derivatives near the boundaries. Following their suggestion, we will take $\alpha = 0.05$.

Sparsity estimation: Since the sparsity, denoted by $s_\tau(x)$, is the derivative of the quantile regression function, $q_\tau(x)$, with respect to τ, we propose an estimate of this kind:

$$\widehat{s}_{\tau,d_s,h_s}(x) = \frac{\widehat{q}_{\tau+d_s,h_s}(x) - \widehat{q}_{\tau-d_s,h_s}(x)}{2\,d_s}$$

where $\widehat{q}_{\tau+d_s,h_s}$ and $\widehat{q}_{\tau-d_s,h_s}$ are local linear quantile regression estimates at the quantile orders $(\tau+d_s)$ and $(\tau-d_s)$, respectively, and h_s denotes their bandwidth.

Note that we need two pilot bandwidths, d_s and h_s. The bandwidth d_s is placed in the Y-axis and plays a similar role to that of the bandwidth d_j in the rule of thumb. The bandwidth h_s is necessary to compute the nonparametric estimations of the regression functions.

The choice of the two pilot bandwidths will be based on the asymptotic mean squared error, which comes from the asymptotic squared bias and variance:

$$\begin{aligned}\mathrm{MSE}\left(\int \widehat{s}^2_{\tau,d_s,h_s}(x)\,dx\right) \cong \left[\frac{1}{nd_sh_s}\int a(x)\,dx + d_s^2\int b(x)\,dx + h_s^2\int c(x)\,dx\right]^2 \\ + \frac{1}{nd_s}\int d(x)\;dx + \frac{1}{n^2d_s^2h_s}\;\int e(x)\;dx\end{aligned} \tag{6}$$

where

$$\begin{aligned} a(x) &= \frac{1}{2}\frac{R(K)\;s_\tau(x)^2}{g(x)} & d(x) &= 2\;\frac{s_\tau(x)^4}{g(x)} \\ b(x) &= \frac{1}{3}s_\tau(x)\;s_\tau^{(2,\tau)}(x) & e(x) &= \left(\frac{1}{2}R(K*K) - R(K)\right)\;\frac{s_\tau(x)^4}{g(x)^2} \\ c(x) &= \mu_2(K)\;s_\tau(x)\;\frac{\partial q_\tau^{(2)}(x)}{\partial\tau} \end{aligned} \tag{7}$$

where $*$ represents the convolution and $s_\tau^{(2,\tau)}(x) = \frac{\partial^2}{\partial\tau^2}s_\tau(x)$.

Minimization with respect to d_s and h_s can be carried out by means of optimization algorithms. The resulting pilot bandwidths will be denoted by $\widehat{d}_s$ and $\widehat{h}_s$, and the estimation of the integrated squared sparsity by $\widehat{s}^2_{\tau,\widehat{d}_s,\widehat{h}_s}$. Now details are given on how to estimate the unknown integrals.

- **Estimation of $\int \mathbf{a(x)\,dx}$.** We will make use of the sparsity estimation at each block, $\widehat{s}_j$, together with a simple estimation of covariate density at that block, that could be given by $n_j/(nl_j)$, where n_j is the number of observations at block j. Then, this integral can be estimated by:

$$\widehat{\int a(x)\,dx} = \frac{1}{2}R(K)\sum_{j=1}^{\widehat{N}}\widehat{s}_j^2\left(\frac{n\,l_j}{n_j}\right)l_j\ .$$

- **Estimation of $\int \mathbf{b(x)\,dx}$.** Recall that $b(x) = (1/3)\;s_\tau(x)\;s_\tau^{(2,\tau)}(x)$, where $s_\tau^{(2,\tau)}(x)$ is the second derivative of $s_\tau(x)$ with respect to τ. The problem of estimating the second derivative of the sparsity without covariates was considered by Bofinger (1975). We apply her proposal to the residuals at each block

$$\widehat{s}_j^{(2,\tau)} = \frac{1}{2\delta^3}\;\left(r_{([n\tau]+2m)} - 2\,r_{([n\tau]+m)} + 2\,r_{([n\tau]-m)} - r_{([n\tau]-2m)}\right)$$

  where the value of m is taken as $m = [n\delta] = [cn^{8/9}]$ with $c = 0.25$, following Sheather and Maritz (1983) proposal. Then, the considered integral is estimated as

$$\widehat{\int b(x)\,dx} = \frac{1}{3}\sum_{j=1}^{\widehat{N}}\widehat{s}_j\;\widehat{s}_j^{(2,\tau)}l_j\ .$$

- **Estimation of $\int \mathbf{c(x)\,dx}$.** The novel ingredient in $c(x)$ is $\partial q_\tau^{(2)}(x)/\partial\tau$. Since this is a derivative with respect to τ, it can be estimated by

$$\widehat{\frac{\partial q_\tau^{(2)}(x)}{\partial\tau}} = \frac{q^{(2)}_{\tau+d_c}(x) - q^{(2)}_{\tau-d_c}(x)}{2d_c}$$

  In order to choose the pilot bandwidth d_c, a location and scale model, given by $Y = q_\tau(X) + \sigma(X)\varepsilon$, is assumed. Here, ε is assumed independent of X and with a zero τ-th quantile. Note that under this model, for each $\tau_1, \tau_2 \in (0,1)$, $q_{\tau_2}(x) - q_{\tau_1}(x) = \sigma(x)(c_{\tau_2} - c_{\tau_1})$, where c_{τ_1} and c_{τ_2} are τ_1 and τ_2 quantiles of ε, respectively. Thus,

$$\frac{\partial q_\tau^{(2)}(x)}{\partial\tau} = \sigma^{(2)}(x)s_\tau(x).$$

This expression leads to consider for d_c the same selector proposed by Bofinger (1975) to estimate the sparsity without covariates. This selector will also be based on the assumption of normality for ε. Finally, we arrive at the following estimator at block j

$$\widehat{\left(\frac{\partial q_\tau^{(2)}}{\partial \tau}\right)}_j = \frac{1}{n_j} \sum_{i=1}^{n} \frac{\widehat{q}^{(2)}_{\tau+\widehat{d}_c,j}(X_i) - \widehat{q}^{(2)}_{\tau-\widehat{d}_c,j}(X_i)}{2\,\widehat{d}_c}\, \mathbb{I}\,(X_i \in \text{Block } j),$$

and the subsequent estimation of the integral

$$\widehat{\int c(x)\,dx} = \mu_2(K) \sum_{j=1}^{\widehat{N}} \widehat{s}_j \widehat{\left(\frac{\partial q_\tau^{(2)}}{\partial \tau}\right)}_j l_j$$

- **Estimation of $\int \mathbf{d}(\mathbf{x})\,\mathbf{dx}$ and $\int \mathbf{e}(\mathbf{x})\,\mathbf{dx}$.** Similarly to the previous integrals, these integrals can be estimated by

$$\widehat{\int d(x)\,dx} = 2\sum_{j=1}^{\widehat{N}} \widehat{s}_j^4 \left(\frac{n\,l_j}{n_j}\right) l_j \quad \text{and} \quad \widehat{\int e(x)\,dx} = \left(\frac{1}{2}R(K*K) - R(K)\right) \sum_{j=1}^{\widehat{N}} \widehat{s}_j^4 \left(\frac{n\,l_j}{n_j}\right)^2 l_j.$$

Finally, the plug-in bandwidth selector is obtained as

$$\widehat{h}_{\tau,\text{PI}} = \left(\frac{R(K)\,\tau(1-\tau)\,\widehat{s}^2_{\tau,\widehat{d}_s,\widehat{h}_s}}{n\,\mu_2(K)^2\,\widehat{\vartheta}_{\widehat{h}_c}}\right)^{1/5}.$$

Remark 2 *In the framework of local linear smoothing quantile regression, Yu and Jones (1998) presented a different approach based on inverting a local linear conditional distribution estimator, that is called double-kernel estimator. Later, Jones and Yu (2007) proposed an improvement of their previous double-kernel estimator. Both double-kernel estimators studied by Yu and Jones (1998) and Jones and Yu (2007) need a bandwidth selector for classical local linear regression to be computed. In this sense, the plug-in rule presented along this work, could be useful in order to perform the double-kernel estimator.*

3. THEORETICAL RESULTS: MEAN SQUARED ERROR OF CURVATURE AND SPARSITY ESTIMATORS

Here expressions (5) and (6) are derived. A complete development of these expressions can be seen in Chapter 3 of Conde-Amboage (2017).

Derivation of (5): In order to derive the asymptotic mean integrated squared error of the curvature estimator, it will be necessary to make the following assumptions:

C1: The density function of the explanatory variable X, denoted by g, is differentiable and its first derivative is a bounded function.

C2: The kernel function K is symmetric, non negative and has a bounded support and verifies that $\int K(u)\ du = 1$, $\mu_6(K) = \int u^6 K(u)\ du < \infty$ and $\int K^2(u)\ du < \infty$. Moreover, it is assumed that the bandwidth parameter h_c verifies that $h_c \to 0$ and $nh_c^5 \to \infty$ when $n \to \infty$.

C3: The conditional distribution function $F(y|X=x)$ of the response variable is three times derivable in x for each y and its first derivative verifies that $F^{(1)}(q_\tau(x)|X=x) = f(q_\tau(x)|X=x) \neq 0$. Moreover, there exist positive constants c_1 and c_2 and a positive function $\text{Bound}(y|X=x)$ such that

$$\sup_{|x_n - x| < c_1} f(y|X=x_n) \leq \text{Bound}(y|x)$$

and

$$\int |\psi_\tau(y - q_\tau(x))|^{2+\delta}\ \text{Bound}(y|X=x)\ dy < \infty$$

$$\int (\rho_\tau(y-t) - \rho_\tau(y) - \psi_\tau(y)t)^2\ \text{Bound}(y|X=x)\ dy = o(t^2), \quad \text{as } t \to 0$$

where $\psi_\tau(r) = \tau\mathbb{I}(r>0) + (\tau-1)\mathbb{I}(r<0)$.

C4: The function $q_{\tau_1}(x)$ has a continuous fourth derivative with respect to x for any τ_1 in a neighbourhood of τ. These derivatives will be denoted by $q_\tau^{(i)}$ with $i \in \{1,2,3,4\}$. Moreover, all these derivatives are bounded functions in a neighbourhood of τ.

Applying the arguments of the proof of Theorem 3 in Fan et al. (1994) to a local polynomial of order 3, the estimator of the second derivative can be approximated by

$$\widetilde{q}^{(2)}_{\tau,h_c}(x) \cong q^{(2)}_\tau(x) + 2h_c^{-2}\frac{1}{f(q_\tau(x)|x)g(x)}V_{n,\tau}(x)$$

where

$$V_{n,\tau}(x) = \frac{1}{nh_c}\sum_{i=1}^{n}\psi_\tau\left(Y_i^{(3)}\right)\left(\alpha_{31}+\alpha_{33}\left(\frac{X_i-x}{h_c}\right)^2\right)K\left(\frac{x-X_i}{h_c}\right),$$

where $Y_i^{(3)} = Y_i - q_\tau(x) - q_\tau^{(1)}(x)(X_i-x) - (1/2)q_\tau^{(2)}(x)(X_i-x)^2 - (1/6)q_\tau^{(3)}(x)(X_i-x)^3$ and $\psi_\tau(z) = \tau - \mathbb{I}(z<0)$. Note that assumptions C1-C4 were used here.

Now, expectation and variance of $\widetilde{q}^{(2)}_{\tau,h_c}(x)$ can be obtained by some algebraic calculations:

$$\mathbb{E}\left(\widetilde{q}^{(2)}_{\tau,h_c}(x)\right) \cong q^{(2)}_\tau(x) + \frac{1}{2}\delta_1 q^{(4)}_\tau(x)h_c^2 \quad \text{and} \quad \mathbb{V}\text{ar}\left(\widetilde{q}^{(2)}_{\tau,h_c}(x)\right) \cong \delta_2\frac{1}{nh_c^5}\frac{\tau(1-\tau)}{f(q_\tau(x)|x)^2 g(x)}$$

where δ_1 and δ_2 were defined in expression (5). Recall that the curvature estimator is given by 4, then, combining expectation and variance of $\widetilde{q}^{(2)}_{\tau,h_c}(X_i)$ conditionally to X_i, and taking expectation with respect to X_i, we obtain

$$\mathbb{E}\left(\widehat{\vartheta}_{h_c}\right) \cong \vartheta + \delta_1\, h_c^2\int q^{(2)}_\tau(x)q^{(4)}_\tau(x)g(x)\,dx + \delta_2\,\tau(1-\tau)\,\frac{1}{nh_c^5}\int\frac{1}{f(q_\tau(x)|x)^2}\,dx$$

where $\vartheta = \int q^{(2)}_\tau(x)g(x)\,dx$ is the curvature that is being estimated. Additional calculations show that the dominant terms in the variance of $\widehat{\vartheta}_{h_c}$ are of orders n^{-1} and $n^{-2}\,h_c^{-9}$. Then, the asymptotic variance is negligible with respect to the asymptotic squared bias, and as a result, the asymptotic mean squared error of curvature estimator coincides with its asymptotic squared bias. This fact, together with last expression for $\mathbb{E}\left(\widehat{\vartheta}_{h_c}\right)$, leads to expression (5).

Derivation of (6): In order to derive the asymptotic mean integrated squared error of the sparsity estimator, it will be necessary to assume the following conditions:

S1: The conditional density function $f(y|X=x)$ of the response variable is twice derivable in x for each y and $f^{(i)}(q_\tau(x)|X=x) \neq 0$ with $i=0,1,2$. Moreover, there exists positive constants c_1 and c_2 and a positive function $\text{Bound}(y|X=x)$ such that

$$\sup_{|x_n-x|<c_1} f(y|X=x_n) \leq \text{Bound}(y|X=x)$$

and

$$\int |\psi_\tau(y-q_\tau(x))|^{2+\delta}\,\text{Bound}(y|X=x)\,dy < \infty$$

$$\int (\rho_\tau(y-t)-\rho_\tau(y)-\psi_\tau(y)t)^2\,\text{Bound}(y|X=x)\,dy = o(t^2), \quad \text{as } t\to 0.$$

S2: The function q_{τ_1} has a continuous second derivative for any τ_1 in a neighbourhood of τ as a function of x. These derivatives will be denoted by $q_\tau^{(i)}$. Moreover, all these functions are bounded functions in a neighbourhood of τ.

S3: The density function of the explanatory variable X, denoted by g, is differentiable and this first derivative is a bounded function.

S4: The kernel K is symmetric, non negative, has a bounded support and verifies that $\int K(u)\,du < \infty$, $\int K(u)^2\,du < \infty$ and $\mu_2(K) < \infty$. Moreover, the bandwidth parameters verify that $d_s \to 0$, $h_s \to 0$ and $nd_sh_s \to \infty$ when $n \to \infty$.

S5: The function q_{τ_1} has a continuous and bounded forth derivative with respect to τ_1 for any τ_1 in a neighbourhood of τ. Moreover, $q^{(2)}_{\tau_1}$ has a continuous and bounded second derivative with respect to τ_1 for any τ_1 in a neighbourhood of τ.

Recall the definition of the proposed sparsity estimator $\widehat{s}_{\tau,d_s,h_s}(x) = \frac{\widehat{q}_{\tau+d_s,h_s}(x)-\widehat{q}_{\tau-d_s,h_s}(x)}{2\,d_s}$ where $\widehat{q}_{\tau+d_s,h_s}$ and $\widehat{q}_{\tau-d_s,h_s}$ are local linear quantile regression estimates at the quantile orders $(\tau+d_s)$ and $(\tau-d_s)$, respectively, and h_s denotes their bandwidth. Applying Fan et al. (1994)'s results, we have

$$\widehat{q}_{\tau+d_s,h_s}(x) \cong q_{\tau+d_s}(x) + \frac{1}{f(q_{\tau+d_s}(x)|x)g(x)} U_{\tau+d_s,h_s}(x)$$

where

$$U_{\tau+d_s,h_s}(x) = \frac{1}{nh_s}\sum_{i=1}^{n} \psi_{\tau+d_s}\left(Y_i^{(1)}\right) K\left(\frac{x-X_i}{h_s}\right), \tag{8}$$

and $Y_i^{(1)} = Y_i - q_{\tau+d_s}(x) - q^{(1)}_{\tau+d_s}(x)(X_i - x)$. Analogously for $\widehat{q}_{\tau-d_s,h_s}(x)$. Substituting these expressions in the definition of $\widehat{s}_{\tau,d_s,h_s}(x)$, we have

$$\widehat{s}_{\tau,d_s,h_s}(x) = A(x) + B(x) \tag{9}$$

with

$$A(x) = \frac{q_{\tau+d_s}(x) - q_{\tau-d_s}(x)}{2\,d_s}, \qquad B(x) = \frac{1}{g(x)}\left(\frac{U_{\tau+d_s,h_s}(x)}{f(q_{\tau+d_s}(x)|x)} - \frac{U_{\tau-d_s,h_s}(x)}{f(q_{\tau-d_s}(x)|x)}\right).$$

Note that $A(x)$ is not random and can be approximated by a Taylor expansion as

$$A(x) \cong s_\tau(x) + \frac{1}{6} s^{(2,\tau)}_\tau(x) d_s^2$$

if S2 follows. Moreover, based on arguments developed in Lemma 2 of Fan et al. (1994), the expectation and variance of $B(x)$ can be approximated by, if assumptions S1-S5 follow,

$$\mathbb{E}(B(x)) \cong \frac{1}{2}\mu_2(K)\frac{\partial q^{(2)}_\tau(x)}{\partial \tau} h_s^2 \quad \text{and} \quad \mathbb{V}\text{ar}(B(x)) \cong \frac{1}{2nd_sh_s}\frac{\int K^2(u)\,du}{f(q_\tau(x)|x)g(x)}.$$

From these results, the asymptotic bias of the estimated squared sparsity is given by

$$\text{Bias}\left(\int \widehat{s}^2_{\tau,d_s,h_s}(x)\,dx\right) \cong \left[\frac{1}{nd_sh_s}\int a(x)\,dx + d_s^2\int b(x)\,dx + h_s^2\int c(x)\,dx\right]^2$$

where $a(x)$, $b(x)$ and $c(x)$ are given in (7).

In view of expression (9), the asymptotic variance of sparsity estimator can be decomposed as follows

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\int \widehat{s}_{\tau,d_s,h_s}(x)^2\,dx\right] &\cong \mathbb{V}\text{ar}\left[\int \left(A(x)^2 + B(x)^2 + 2A(x)B(x)\right)\,dx\right] \\
&= \mathbb{V}\text{ar}\left[\int B(x)^2\,dx\right] + 4\,\mathbb{V}\text{ar}\left[\int A(x)B(x)\,dx\right] \\
&\quad + 4\,\mathbb{C}\text{ov}\left[\int B(x)^2\,dx, \int A(x)B(x)\,dx\right].
\end{aligned}$$

Each of the previous terms can be expressed as covariances of U-expressions like that given in (8) evaluated at different points x and quantiles $\tau+d_s$ and $\tau-d_s$. These covariances can be computed (under assumptions S1, S2, S4 and S5) using similar arguments to those employed by Fan et al. (1994), adapting their φ function (given in equation (2.1) on page 435) to each covariance. Then, the asymptotic variance of sparsity estimator can be approximated as follows

$$\mathbb{V}\text{ar}\left[\int \widehat{s}_{\tau,d_s,h_s}(x)^2\,dx\right] \cong \frac{1}{nd_s}\int d(x)\,dx + \frac{1}{n^2d_s^2h_s}\int e(x)\,dx$$

where $d(x)$ and $e(x)$ are given in (7). Then, in view of the computed asymptotic bias and variance, expression (6) can be derived.

4. SIMULATION STUDY AND APPLICATION TO REAL DATA

In this section a simulation study is presented to analyse the behaviour of the new bandwidth selectors in comparison with already existing selectors. Some theoretical considerations are useful as an orientation to a meaningful comparison. Observing that the asymptotically optimal bandwidth for mean regression is given by

$$h_{\text{AMISE,MEAN}} = \left[\frac{R(K)}{n\ \mu_2(K)^2\ \int m^{(2)}(x)^2\ g(x)\ dx}\ \int \sigma^2(x)\ dx \right]^{1/5},$$

where m is the mean regression and σ^2 is the conditional variance, Yu and Jones (1998) proposed to use the following selector:

$$\widehat{h}_{\text{YJ},\tau} = \widehat{h}_{\text{RSW}} \left[\frac{\tau(1-\tau)}{\phi(\Phi^{-1}(\tau))^2} \right]^{1/5}$$

where $\widehat{h}_{\text{RSW}}$ is the Ruppert et al. (1995)'s plug-in selector for local linear mean regression, and the last factor is a correction for quantile regression. Yu and Jones (1998)'s selector is based on assuming that quantile and mean regression have the same curvature and the error distribution is normal. The last factor then relates the conditional sparsity with the conditional variance under normality.

Since $\widehat{h}_{\text{RSW}}$ converges to $h_{\text{AMISE,MEAN}}$, $\widehat{h}_{\text{YJ}}$ converges to

$$h_{\text{AMISE,YJ},\tau} = h_{\text{AMISE,MEAN}} \left[\frac{\tau(1-\tau)}{\phi(\Phi^{-1}(\tau))^2} \right]^{1/5}$$

which is generally different from the asymptotically optimal bandwidth for quantile regression, $h_{\text{AMISE},\tau}$. Meanwhile, the proposed plug-in selector $\widehat{h}_{\text{NP}}$ converges to $h_{\text{AMISE},\tau}$. Then, for a sample size large enough, the new bandwidth is expected to outperform Yu and Jones (1998)'s selector, the latter selector being generally inconsistent. This simulation study will help to assess the consequences of these facts from smaller to larger sample sizes, and in models where the difference between $h_{\text{AMISE,YJ},\tau}$ and $h_{\text{AMISE},\tau}$ can be controlled.

In particular, for any homoscedastic quantile regression model $Y = q_\tau(X) + \varepsilon$, where the model error ε has τ-quantile zero and is assumed independent of X, the curvatures of mean and quantile regression coincide, and then the quotient between $h_{\text{AMISE,YJ},\tau}$ and $h_{\text{AMISE},\tau}$ will be

$$\text{Ratio} = \frac{h_{\text{AMISE,YJ},\tau}}{h_{\text{AMISE},\tau}} = \sqrt[5]{\frac{\sigma^2\ f_\varepsilon(F_\varepsilon^{-1}(\tau))^2}{\phi(\Phi^{-1}(\tau))^2}} \qquad (10)$$

where f_ε and F_ε are the density and distribution functions of ε, and σ^2 denotes the variance of ε. Then, the ratio between both AMISE bandwidths only depends on the error distribution. Figure 1 shows the values taken by the ratio defined in (10) as a function of the quantile τ and for three error distributions: exponential with expectation one, uniform on the interval $(0, 1)$ and beta with parameters 5 and 1.

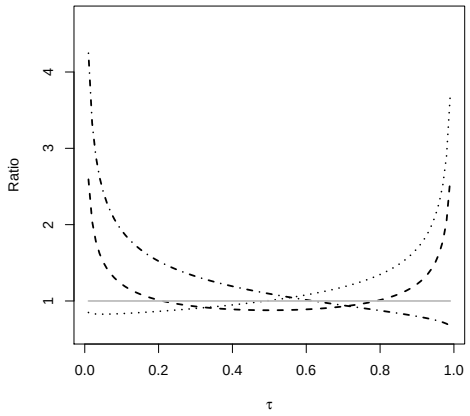


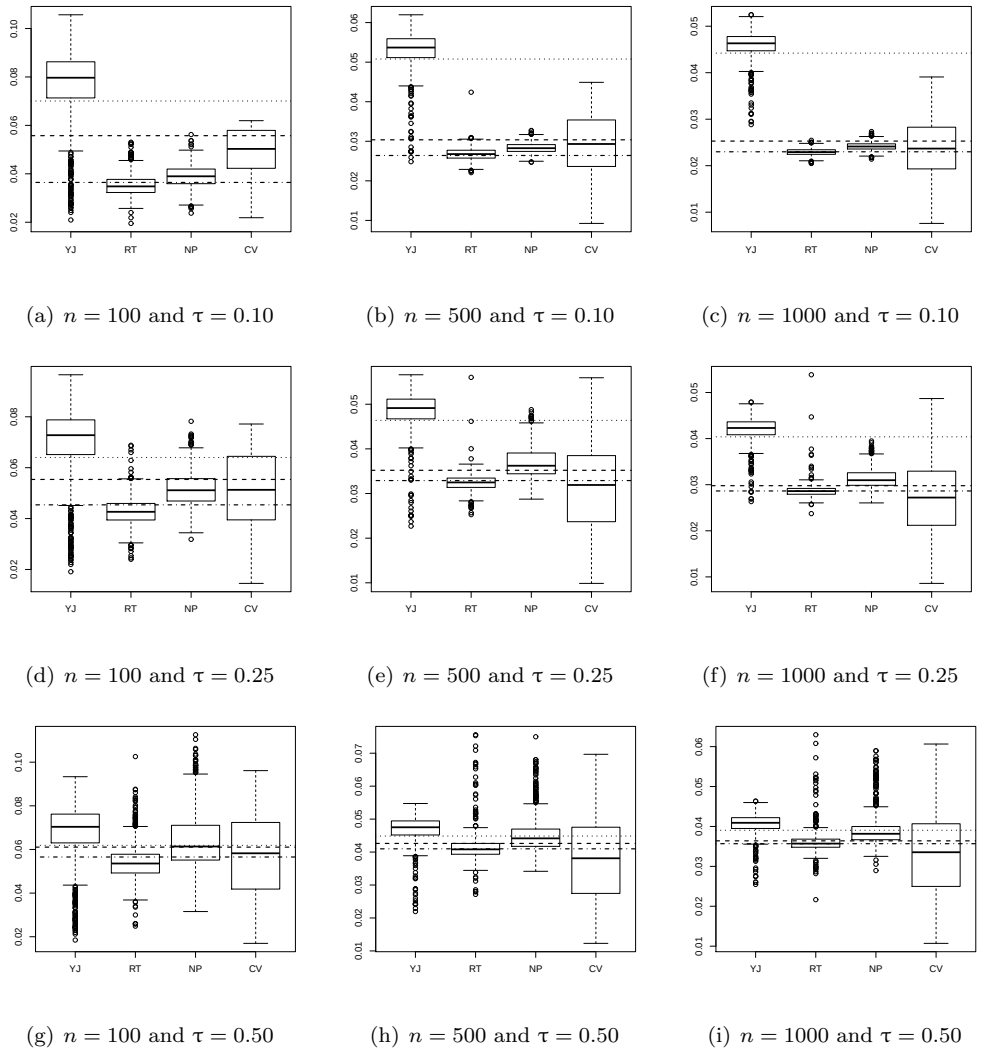
Figure 1: Representation of the ratio between the AMISE bandwidths as a function of the quantile τ and for three error distributions. The dashed line represents the uniform distribution, the dotted line represents the beta distribution and the dashed and dotted line represents the exponential distribution.

As shown in Figure 1, we observe that the differences between both AMISE bandwidth will be bigger as long as the error distribution differs from the Gaussian distribution. Furthermore, if we fix an error distribution, the behaviour of both bandwidths will depend on the quantile of interest.

Our first simulated model is given by

$$\text{Model 1:} \qquad Y = 10(X^4 + X^2 - X) + \varepsilon,$$

where X follows a uniform distribution on the interval $(0, 1)$ and ε is the unknown error, which is drawn independently of X. Note that in this case, $q_\tau(X) = 10(X^4 + X^2 - X) + c_\tau$ where c_τ represents the τ-quantile of the error distribution. In this model the error follows an exponential distribution with expectation 1, which is one of the distributions represented in Figure 1.



(a) $n = 100$ and $\tau = 0.10$ (b) $n = 500$ and $\tau = 0.10$ (c) $n = 1000$ and $\tau = 0.10$

(d) $n = 100$ and $\tau = 0.25$ (e) $n = 500$ and $\tau = 0.25$ (f) $n = 1000$ and $\tau = 0.25$

(g) $n = 100$ and $\tau = 0.50$ (h) $n = 500$ and $\tau = 0.50$ (i) $n = 1000$ and $\tau = 0.50$

Figure 2: Boxplot representations of Yu and Jones (1998)'s selector (YJ), the new rule of thumb (RT), the new plug-in selector (NP) and the cross-validation selector (CV), from 1000 replications of Model 1 for different values of τ and n. The dashed line represents the MISE bandwidth, the dotted line represents the Yu and Jones (1998)'s AMISE bandwidth and the dashed and dotted line represents the AMISE bandwidth.

Figure 2 represents the boxplots corresponding to the four bandwidth selectors: the plug-in selector proposed by Yu and Jones (1998), the selector based on the new rule of thumb, the new plug-in selector and the cross-validation selector. They are denoted by *YJ, RT, NP* and *CV*, respectively. The boxplots were obtained from 1000 replications of Model 1 for different values of τ, and sample sizes $n = 100, 500, 1000$. Three horizontal lines are added to the plots, representing the optimal bandwidths with three criteria: MISE (dashed line), Yu and Jones (1998)'s AMISE (dotted line) and AMISE (dashed and dotted line). The best of these bandwidths would be the one optimizing the MISE, so the performance of each selector is related to its approximation to this bandwidth. The AMISE bandwidth is an approximation to the MISE bandwidth. In fact both lines approaches to each other for increasing sample size. Meanwhile, Yu and Jones (1998)'s AMISE (YJ-AMISE) bandwidth do not approximate to MISE bandwidth even for a very large

sample size. This is the cause for inconsistency of Yu and Jones (1998)'s selector. However, for a small sample size, the errors of approximation between the three bandwidths can be confounded. As regards the value of τ, Figure 2 shows that for $\tau = 0.5$ the three bandwidths are quite similar, while for $\tau = 0.1$ they are far apart.

Yu and Jones (1998)'s selector estimates YJ-AMISE bandwidth, while the new selectors estimate AMISE bandwidth. For sample size $n = 500$, this leads to a clearer better performance of the new bandwidths, while for small sample size $n = 100$, the errors between optimal bandwidths are still confounded. The cross-validation bandwidth is generally centred to the MISE bandwidth, but its variability is clearly larger.

Now we are going to evaluate the performance of each selector in terms of the observed ISE. Following the arguments developed in Jones (1991), the sample mean of the integrated squared error (denoted by SMISE) will be employ to compare the different bandwidth selectors. To this aim, a Monte Carlo simulation study has been design where the number of replications is 1000. We expect that SMISE results will be a consequence of the selectors' distances from the MISE bandwidth, observed in the boxplots. To complete the presentation, a new model is included, again homoscedastic but with a larger curvature:

$$\text{Model 2:} \qquad Y = 1 - 48X + 218X^2 - 315X^3 + 145X^4 + \varepsilon,$$

where X follows a uniform distribution on the interval $(0, 1)$, and ε follows an exponential distribution with expectation 1, and is drawn independently of X.

		Model 1				Model 2			
τ	n	YJ	RT	NP	CV	YJ	RT	NP	CV
0.10	100	29.51	32.41	29.69	28.67	32.21	48.65	31.79	31.44
	500	6.93	4.71	4.65	5.48	7.16	5.99	5.49	6.26
	1000	4.01	2.41	2.39	2.78	4.50	3.06	2.95	3.34
0.25	100	48.81	45.82	43.49	52.64	51.66	59.66	50.58	56.84
	500	11.23	9.31	9.43	11.65	12.42	11.61	11.12	13.54
	1000	6.28	5.02	5.09	6.02	7.57	6.39	6.28	7.36
0.50	100	94.01	88.23	89.12	103.71	107.46	109.01	99.69	116.28
	500	21.97	21.61	21.98	26.62	26.23	26.35	26.36	31.56
	1000	12.04	11.80	12.03	14.23	14.98	14.86	14.99	17.41

Table 1: Sample mean of the integrated squared error (given values were multiplied by 10^3) associated with the considered bandwidth selectors, from 1000 replications of Model 1 and Model 2, with several sample sizes n and values of τ.

Table 1 contains the sample mean integrated squared error for the considered bandwidth selectors for several samples sizes and values of τ. We can observe that the new plug-in rule shows a better performance in terms of SMISE than the plug-in selector proposed by Yu and Jones (1998). It is interesting to emphasize the good behaviour of the rule of thumb, despite its simplicity. The cross-validation bandwidth shows a generally worst SMISE in the considered scenarios.

Summarizing, other scenarios are considered in order to show the good behaviour of the proposed bandwidth selector compare with natural competitors. The simulated sample mean integrated squared error shows the better performance of the new plug-in rule if: we fixed a quantile of interest and fit homoscedastic quantile regression models with different error distributions, we fit heteroscedastic regression models associated with different quantiles of interest, or we have considered a non-smooth model to check the robustness of the new plug-in selector to deviations from the assumed degree of smoothness of the quantile regression model. Thanks to these results we can conclude that the new plug-in rule outperforms the other three selectors.

Finally, the bandwidth selection method presented is also illustrated using real data. This data illustration allows us to show again the suitable performance of the new method in practice. In addition, we have developed an R package called `BwQuant` that enables any user to apply the techniques proposed: rule of thumb and plug-in rule. The natural competitors, cross-validation and Yu and Jones (1998)'s bandwidths, were also implemented.

5. CONCLUSIONS

We have proposed a new plug-in bandwidth selector for local linear quantile regression based on a nonparametric approach. This new method involves nonparametric estimation of the curvature of the quantile regression function and the integrated squared sparsity.

Thanks to a Monte Carlo simulation study, we have shown that the new proposal shows a good behaviour in terms of the sample mean of the integrated squared error compared with its natural competitors both in homoscedastic and heteroscedastic scenarios. Moreover, we have presented a simple rule of thumb that shows a quite good performance on a wide range of situations.

In order to facilitate the use of the new bandwidth selector an R package has been designed, called `BwQuant`. This package also implements already known bandwidth selectors for local linear quantile regression.

The developed methodology can be useful in order to apply the double-kernel estimator proposed by Yu and Jones (1998). Moreover, the presented ideas can be extended to the multivariate case, using nonparametric additive models in quantile regression context following the ideas of Yu and Lu (2004). Similarly to Yu and Jones (1998), Yu and Lu (2004) proposed to use Opsomer and Ruppert (1998)'s bandwidth with some transformation based on assumptions such as homoscedasticity or error normality.

Acknowledgements. The authors gratefully acknowledge the support of Projects MTM2013–41383–P (Spanish Ministry of Economy, Industry and Competitiveness) and MTM2016–76969–P (Spanish State Research Agency, AEI), both co–funded by the European Regional Development Fund (ERDF). Support from the IAP network StUDyS, from Belgian Science Policy, is also acknowledged. Work of M. Conde-Amboage has been supported by FPU grant AP2012-5047 from the Spanish Ministry of Education.

REFERENCES

Abberger, K. (1998). Cross-validation in nonparametric quantile regression. Allgemeines Statistisches Archiv, 82, 149–161.

Bloch, D.A. and Gastwirth, J.L. (1968). On a simple estimate of the reciprocal of the density function. The Annals of the Mathematical Statistics, 39, 1083–1085.

Bofinger, E. (1975). Estimation of a density function using order statistics. Australian Journal of Statistics, 17, 1–7.

Conde-Amboage, M. (2017). Statistical inference in quantile regression models, PhD Thesis. Universidade de Santiago de Compostela, Spain.

Fan, J., Hu, T.C. and Truong, Y.K. (1994). Robust nonparametric function estimation. Scandinavian Journal of Statistics, 21, 433–446.

Hall, P. and Sheather, S.J. (1988). On the distribution of a studentized quantile. Journal of the Royal Statistical Society. Series B (Methodological), 50: 381–391.

Jones, M.C. (1991). The roles of ISE and MISE in density estimation. Statistics and Probability Letters, 12, 51–56.

Jones, M.C and Yu, K. (2007) Improved double kernel local linear quantile regression. Statistical Modelling, 7: 377-389.

Koenker, R. and Bassett, G. (1978). Regression quantiles. Econometrica, 46, 33–50.

Mallows, C.L. (1973). Some comments on C_p. Technometrics, 15, 661–675.

Opsomer, J. D. and Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. Journal of the American Statistical Association, 93, 605–618.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An efective bandwidth selector for local least squares regression. Journal of the American Statistical Association, 90, 1257–1270.

Sheather, S.J. and Maritz, J.S. (1983). An estimate of the asymptotic standard error of the sample median. Australian Journal of Statistics, 25, 109–122.

Yu, K. and Jones, M.C. (1998). Local linear quantile regression. Journal of the American Statistical Association, 93, 228–237.

Yu, K. and Lu, Z. (2004). Local linear additive quantile regression. Scandinavian Journal of Statistics, 31, 333–346.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Plugin de R-Commander para o Ensino de Estatística Básica

Tomás R. Cotos-Yáñez[1], Manuel Alfredo Mosquera[1],Ana Pérez-González[1] e Benigno Reguengo Lareo

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

RESUMO

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. También se muestran las principales características del paquete señalando las ventajas más destacables tanto desde el punto de vista del docente y del alumno.

Palabras e frases chave: Estatística, R, R-Commander, plugin

1. INTRODUCIÓN

La idea de la creación de este plugin surgió de la necesidad de un software más adecuado para la docencia de las materias de Estadística en los grados de ADE, Informática y Turismo impartidos en el Campus de Ourense.

Los contenidos de dichas materias abarcan una parte de Estadística Descriptiva, Variables aleatorias y finalmente el bloque correspondiente a Inferencia Estadística donde se abordan los Intervalos de Confianza y Contastes de Hipotesis bajo hipótesis de Normalidad.

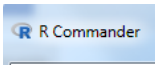
Hasta el momento el software que se utilizaba para impartir estas materias era de distinto tipo, SPSS, Excel, R, etc.

Parte de este software, a pesar de que su interfaz es muy amigable, su manejo es sencillo y calcula todo lo necesario para la docencia, presentaba la dificultad del acceso a licencias por parte de los alumnos para poder utilizarlo fuera de la Universidad. Dada la metodología docente actual en la cual el alumno tiene que realizar una parte importante del aprendizaje de la materia fuera del aula, el inconveniente de la licencia era muy importante.

Por otra parte, otro tipo de software como el R, cuya distribución es libre, necesitaba de unos conocimientos básicos en programación que lo hacía de muy difícil acceso a alumnos de titulaciones menos técnicas como ADE o Turismo.

Eso hizo que en los últimos años usáramos R (R Core Team, 2016) junto con la librería R-Commander (Fox, 2015).



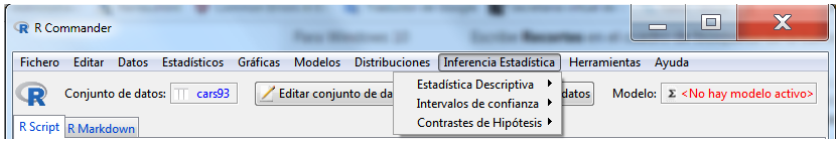


La utilización del R-Commander proporciona un interface más amigable para el alumno y su utilización es sencilla incluso para usuarios que no tienen conocimientos en programación.

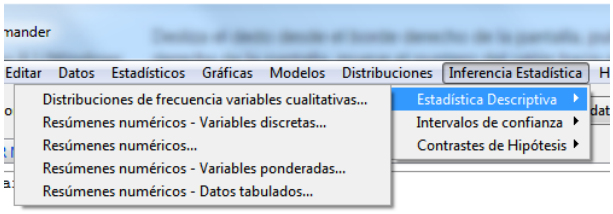
El inconveniente que presentaba la utilización de esta librería era que algunas técnicas estadísticas básicas que se manejan en la parte de Estadística descriptiva como Análisis de datos Tabulados o Medidas Descriptivas de datos ponderados, Intervalos de Confianza o contrastes para la varianza de una muestra, para la proporción, etc., no se encuentran disponibles en los menús. Esto motivó que pensásemos en un plugin (Fox, 2007) para Rcommander que incluyera estas metodologías. La elaboración del plugin constituyó la realización de un Proyecto Fin de Carrera de la Ingeniería Superior de Informática del Campus de Ourense realizado por Benigno Reguengo Lareo en 2016. Dicho proyecto fue dirigido por Manuel A. Mosquera Rodríguez y Tomás R. Cotos Yáñez

2. ESTRUCTURA DEL PLUGIN

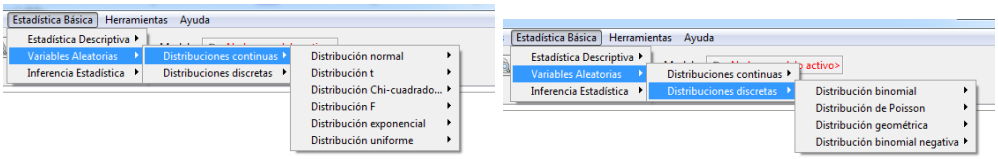
Con la instalación del plugin en Rcommander aparece en el menú un nuevo apartado con el nombre de Estadística Básica con tres epígrafes: Estadística Descriptiva, Variables Aleatorias e Inferencia Estadística.



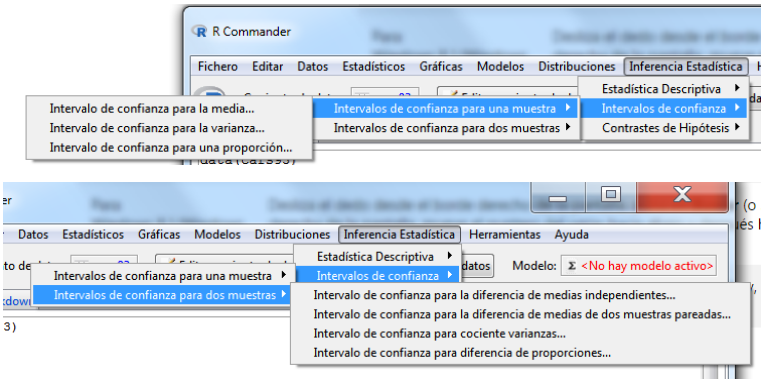
Dentro del submenú de Estadística descriptiva, podemos observar una serie de Resúmenes destinados a variables cualitativas, discretas, variables de tipo continuo, variables ponderadas, datos tabulados e incluso hemos incluido también el menú de Tablas de Contingencia que el R-Commander tiene en el apartado de Estadísticos. Este menú permite realizar un análisis descriptivo básico de prácticamente cualquier tipo de variable.



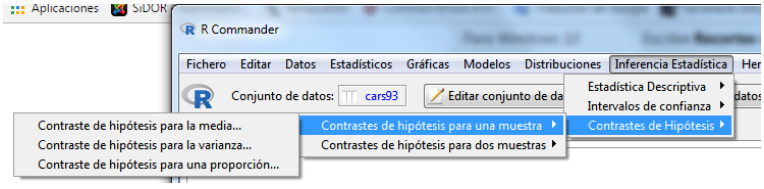
En el submenú de variables aleatorias hemos incluido parte de las funciones que aparecen en el menú de Distribuciones de R-Commander, relativas al cálculo de probabilidades, cuantiles y muestras aleatorias de las principales variables aleatorias discretas y continuas que se dan a conocer en los cursos de Estadística de las Titulaciones del Campus.

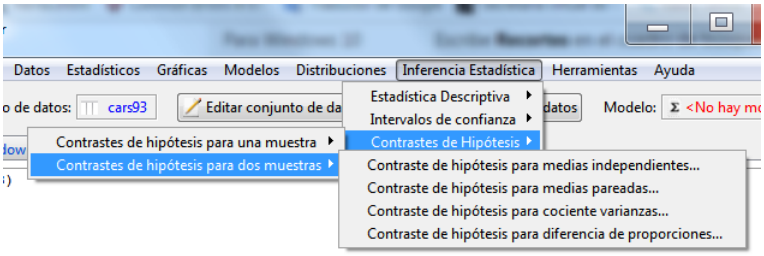


Finalmente en el menú de Inferencia Estadística se puede ver un Submenú de Intervalos de Confianza dónde se permite calcular los intervalos de confianza para una y dos muestras bajo hipótesis de Normalidad.

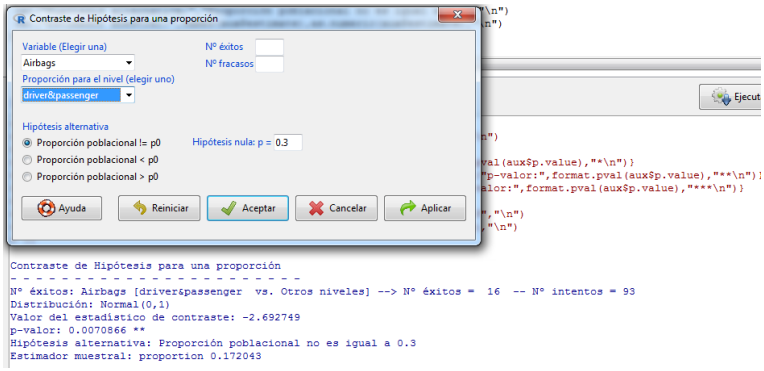


Análogamente para el caso de Contrastes de Hipótesis, incluyendo contrastes de hipótesis paramétricos para una y dos muestras:





Las salidas de las metodologías incluidas han sido adaptadas al objetivo de la materia, es decir, por ejemplo la salida de un contraste de hipótesis para una proporción, contiene los elementos básicos que se deben conocer al calcular un contraste de una proporción: Hipótesis alternativa, valor del estadístico, distribución muestral del estadístico, p-valor, estimador de la proporción, etc.



3. PRINCIPALES CARATERíSTICAS DEL plugin

Como principales características del plugin podemos señalar las siguientes: Realiza análisis descriptivo para datos ponderados; Realiza análisis descriptivos para datos tabulados; Realiza Intervalos de Confianza y Contrastes para la varianza de una muestra; Realiza Intervalos de Confianza y contrastes para medias con varianzas conocidas; Realiza Intervalos de Confianza y contrastes clásicos para proporciones, en este caso sobre proporciones de una categoría de una variable factor del archivo o para un valor determinado de éxitos sobre fracasos (véase figura anterior); Selección de factores de más de 2 niveles para establecer los grupos en inferencia de dos muestras independientes; Selección de factores de más de 2 niveles en inferencia para proporciones; Está disponible en Inglés, Español y en un futuro también en Gallego. Dado que este paquete ha sido creado en 2016, su utilización es muy reciente y no ha sido posible todavía la comparativa en cuanto a resultados académicos. Las mejoras que se han obtenido con su utilización en la docencia han sido las siguientes: Integración de todas las metodologías estadísticas impartidas en la docencia de las materias a las que hacíamos referencia en el principio en un único menú; Clara diferenciación entre Intervalos de Confianza y Contrastes de Hipótesis; La sencillez del manejo, permite que el alumno se concentre en los contenidos estadísticos salvando en gran parte la dificultad del conocimiento de programación R. Esto permite a su vez al profesor profundizar más en los razonamientos puramente estadísticos.

REFERENCIAS

Fox, J. (2005). The R Commander: A basic statistics graphical user interface to R. Journal of Statistical Software, 14, 1–42.

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/.

Fox, J. (2007). Extending the R Commander by plug-in packages. R News, 7, 46–52.

Reguengo Lareo, B. (2016). Plugin de R-Commander para la enseñanza de Inferencia Estadísitica básica. Universidade de Vigo.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

CLASIFICACION DE ACEITES COMESTIBLES SEGÚN SU COMPOSICIÓN Y ESTABILIDAD A LA OXIDACIÓN

Ana María Díaz Díaz[1], Jorge López Beceiro[1], Ramón Artiaga[1] y Javier Tarrío Saavedra [2]

[1] Grupo PROTERM, Escola Politécnica Superior, Universidade da Coruña.
[2] Grupo MODES, CITIC, Universidade da Coruña.

RESUMEN

En este trabajo se presenta un estudio de caracterización y comparación de aceites según su composición y resistencia a la oxidación. Es, por tanto, un trabajo enmarcado en las áreas de química analítica y análisis térmico, que precisa el uso de la estadística para obtener información relevante de los ensayos realizados. El presente estudio parte de la idea que el grado de resistencia a la oxidación que pueden sufrir distintos aceites comestibles y de uso alimentario es un indicador de su estabilidad y calidad, tanto durante su procesado y almacenamiento, como durante su utilización por la industria alimentaria o su consumo por particulares, como ingrediente para uso culinario. Actualmente se asume que la oxidación de aceites a alta temperatura está relacionada con su composición química aunque, si bien existen diversos estudios sobre los efectos de las diferentes proporciones de ácidos grasos y otros compuestos, se echa en falta un estudio cuantitativo de los efectos de compuestos minoritarios y de las interacciones entre éstos y los componentes químicos mayoritarios. También hay que tener en cuenta que los métodos más habituales para evaluar la susceptibilidad a lo oxidación se llevan a cabo a temperaturas muy inferiores a las que se alcanzan cuando son utilizados en restauración. En este trabajo se ha optado por reproducir las altas temperaturas de uso de los aceites mediante de un calorímetro diferencial de barrido a presión (PDSC, Differential Scanning Calorimeter). Este tipo de máquina de ensayos proporciona como resultado curvas de intercambio de energía o curvas DSC, que proporcionan información acerca de los procesos de oxidación sufridos por cada aceite. Se han seleccionado características relevantes de estas curvas como son el Onset Oxidation Temperature (OOT), obteniendo vectores de características a partir de los cuales se han aplicado métodos de reducción de dimensión y clasificación clúster con el fin de identificar las características más relevantes, caracterizar y comparar los distintos tipos de aceites. Este trabajo representa una primera aproximación al problema descrito, que en un futuro requerirá la aplicación de técnicas de regresión y clasificación para modelizar la relación entre oxidación y composición de los aceites.

Palabras y frases clave: Análisis térmico, estabilidad oxidativa, aceite vegetal, Estadística multivariante, Reducción de dimensión, Biplot.

1. INTRODUCCIÓN

Los procesos de oxidación que pueden sufrir distintos aceites comestibles y de uso alimentario determinan su estabilidad a la oxidación tanto durante el procesado y almacenamiento de los mismos, como durante su utilización por la industria alimentaria o consumo por particulares como ingrediente para uso culinario. Asimismo estos aceites pueden ser susceptibles control de calidad y de otros usos que hacen de la estabilidad oxidativa un indicador importante sobre estos productos. Se han seleccionado para su estudio diferentes aceites comerciales de origen vegetal de uso alimentario de diversas marcas españolas y francesas. También se ha considerado un aceite de origen animal, en concreto aceite de pescado, por su importancia dentro del grupo de los aceites de tipo terapéutico. Una primera clasificación de las muestras se puede hacer atendiendo al origen y usos de los distintos aceites es la siguiente: aceites terapéuticos (onagra, borraja, aceite de germen de trigo, aceite de pescado), aceites comestibles de frutos secos (aceite de avellana, de nuez, de cacahuete); aceites provenientes de semillas (aceite de pepitas de uva, aceite de semillas de calabaza); aceites de oliva (variedades hojiblanca, cornicabra, arbequina,

picual); como casos especiales se analiza también un aceite mineral que actúa como referencia o blanco, y oleína de palma, fracción líquida del aceite de palma, por su comportamiento especial (Aparicio y Luna, 2002; Dubois et al., 2007; Ferreira et al., 2014; Koushki et al., 2015; McGuire et al., 2004; FAO, 1999).

Los aceites comestibles son sustancias que están compuestas principalmente de glicéridos de ácidos grasos y han sido obtenidos mayoritariamente de fuentes vegetales. Pueden contener además otros lípidos como por ejemplo fosfátidos o constituyentes insaponificables, o ácidos grasos libres presentes originariamente en ese aceite o grasa. La oxidación de aceites a alta temperatura está relacionada con su composición química, existen diferentes estudios sobre los efectos de las diferentes proporcione de ácidos grasos y otros compuestos, aunque se echa en falta un estudio cuantitativo de los efectos de compuestos minoritarios y sobre las interacciones entre componentes químicos mayoritarios y minoritarios. Por todo lo anterior, se ha decidido evaluar la estabilidad de los aceites a través de análisis térmico con el método OOT (onset oxidation temperature). Las curvas obtenidas con este método o los parámetros asociados a dichas curvas pueden utilizarse para identificar diferentes aceites.

2. ANALISIS TÉRMICO SOBRE MUESTRAS DE ACEITES

Se han empleado técnicas de análisis térmico para la caracterización de los procesos de oxidación. Estas técnicas miden las propiedades de los materiales en función de la temperatura, tiempo y atmósfera presente durante los ensayos de los mismos. Los calorímetros diferenciales de barrido (DSC, Differential Scanning Calorimeter) miden las temperaturas y los flujos de calor asociados con las transiciones térmicas en un material. Entre las propiedades que los equipos de DSC miden, se incluyen las transiciones de material vítreo, la cristalización “fría”, los cambios de fases, la fundición, la cristalización, la estabilidad del material, el curado, la cinética del curado y la estabilidad oxidativa.

Los experimentos con los aceites se han realizado en una celda de presión de TA Instruments montada en un DSC Q2000 modulado (PDSC). Muestras de masa con un rango entre 2.7–3.3 mg se colocaron en celdas abiertas de aluminio dentro del PDSC. Las condiciones experimentales consistieron en aplicar una rampa de calentamiento de 10 °C por minuto desde temperatura ambiente hasta 300 °C. De acuerdo con lo indicado en la norma ASTM E 2009 del método para obtener el parámetro denominado Onset Oxidation Temperature (OOT, temperatura a la cual comienza el pico de la curva DSC correspondiente al proceso de oxidación), se ha mantenido una atmósfera de oxígeno a 3.5 MPa, con un flujo constante de 50 mililitros por minuto. Los experimentos se pararon manualmente una vez alcanzada la pendiente máxima de la exoterma. En algunos casos, el experimento se deja evolucionar hasta el final. Las curvas de cada aceite para este tipo de experimento presentan un perfil característico de brusco cambio de pendiente una vez que se alcanza la temperatura OOT. Cuanto mayor es el valor de la temperatura OOT para un aceite, mayor es su estabilidad frente a la oxidación. Para determinar exactamente la temperatura OOT, el equipo viene previsto de un software de análisis Universal Analysis 2000, en este caso versión 4.5A que ofrece este resultado. Gráficamente se determina extrapolando las rectas antes y después del cambio brusco de pendiente.

Las curvas PDSC representan el intercambio de energía en la muestra con respecto a la temperatura, dada una cierta presión. Proporcionan información relacionada con la estabilidad termooxidativa del material. La utilización de la técnica PDSC frente a otras técnicas de análisis térmico presenta ciertas ventajas; con PDSC se obtienen resultados en menos tiempo que con la técnica DSC, reduciéndose la evaporación de la muestra.

En la Tabla 1 se presenta un perfil de la composición de ácidos grasos de los distintos aceites ensayados. Estas composiciones han sido recopiladas de distintas fuentes bibliográficas, constituyendo una referencia de composición química teórica que en un futuro se puede sustituir por las composiciones químicas reales de cada aceite en concreto, lo que supone un análisis de cada uno de las muestras por otras técnicas específicas como por ejemplo cromatografía de gases. Para ello se necesitará una preparación previa de las mismas a través de una etapa de metilación y posterior medida en el aparato correspondiente.

Componente de los aceites	Palm olein (1)	Pumpkinseed (2)	Hazelnut (3)	Peanut (3)	Walnut (3)	Grape seed (4)	Borage (2)	Evening primrose (2)	Fish (5)	Wheatgerm (2)	Oliva (Arbequina) (6)	Oliva (Picual) (6)	Oliva(Hojiblanca) (6)	Oliva (Comicabra) (6)
TS	45,9	22	8,85	15,4	9,25	11,7	16,2	7,7	15,9	18,7	16,4	13,1	12,3	11
TU	54,3	77,6	90,7	84,6	90,3	88,9	86,8	92,6	93,7	82,2	84,4	83,5	85,4	86,8
MU	42,7	26,1	79,6	38,6	21,2	16,5	26	9,2	22,8	20,7	75,3	77,8	75,2	78,7
PU	11,6	51,5	11,1	46	69	72,4	60,8	83,4	70,9	61,5	9,1	5,7	10,2	8,1
C12:0	0,2		ND	ND	ND									
C14:0	1	0,2	0,13	0,03	0,13		0,1							
C16:0	39,8	12,8	5,82	11	6,7	7,4	11,9	6	12,4	17,5	14,3	9,9	9	8,4
C16:1	0,2	0,4	0,29	0,15	0,23	0,6	0,2	0,1	1,28	1,1				
C17:0		ND	ND	ND	ND					0				
C17:1		ND	ND	ND	ND					0				
C18:0	4,4	9	2,74	2,66	2,27	3,9	4,3	1,6	3,54	1	2,1	3,2	3,3	2,6
C18:1	42,5	25,7	79,3	38,4	21	15,6	18,8	8,9	21,5	19,1	75,3	77,8	75,2	78,7
C18:2	11,2	51,3	10,4	44,6	57,5	72,2	38,5	73,8	46,8	55,2	8,5	5	9,5	7,5
C18:3	0,4	0,2	0,46	0,58	11,6	0,24	22,3	9,6	18,3	6,1	0,6	0,7	0,7	0,6
C20:0	0,4	0	0,16	1,57	0,08		0,2	0,3		0,2				
C20:1	0	ND	ND	ND	ND	ND-0,3	3,9	0,3		1,6				
C20:2			ND	ND	ND	ND	0,2	0,1		0,2				
C20:3			ND	0,02	ND	ND								
C20:5			ND	0,02	0,06	ND			2,83					
C22:0	0,1	0	ND	0,1	0,07	ND-0,5	0,2	0,1		0				
C22:1			ND	ND	ND	ND-0,3	2,1	0,1		0,4				
C22:2			ND	ND	ND	ND								
C22:6			0,27	0,75	ND	ND			3,05					
C24:0		0	ND	ND	ND	ND-0,4				0				
C24:1			ND	ND	ND	ND	1,2			0,2				

Tabla 1: Composición de ácidos grasos de los aceites analizados. El número entre paréntesis en el nombre denota la referencia bibliográfica de origen

ANÁLISIS ESTADÍSTICO

El vector de características compuesto por las características extraídas de las curvas DSC obtenidas mediante el uso del análisis térmico y por los datos de composición (contenido en diversos ácidos grasos indicados en la Tabla 1) caracteriza cada una de las muestras de aceite estudiadas. Un objetivo importante de este trabajo es identificar las características más relevantes para diferenciar entre aceites, además de relacionar la composición de dichos aceites con su estabilidad a la oxidación. Con tal fin, se han aplicado técnicas de reducción de la dimensión como es el análisis por componentes principales (ACP), con ayuda del software estadístico R, en particular la librería FactoMineR (Kassambara, 2017; Lê et al., 2008). ACP permite representar los individuos en un nuevo conjunto de ejes ortogonales, conocidos como componentes principales, a partir de las variables originarias. La primera componente principal así construida será la componente que explique una mayor parte de la varianza global de los datos de partida,

por otro lado, la segunda será ortogonal con respecto a la primera y tendrá la dirección conforme explique la mayor varianza remante de los datos, y así se construirán sucesivamente todas las nuevas componentes incorreladas. Cada componente principal viene definida por el autovector de la matriz de varianza. Cada autovalor, la cantidad de varianza explicada por una componente dada, se utiliza para determinar las varianzas de los componentes principales más importantes. El uso de gráficos como el biplot permite ver la relación entre variables originales y componentes principales, además, la aplicación del algoritmo de clasificación clúster jerárquico proporciona información acerca de los diferentes grupos existentes en la base de datos de aceites. En particular, se ha aplicado el algoritmo clúster jerárquico con distancia promedio no ponderada (Lê et al., 2008), incluyendo también un procedimiento automático de identificación del número de particiones, correspondiente a la mayor pérdida relativa de inercia entre clústeres (i(clústeres n+1)/i(clústeres n)).

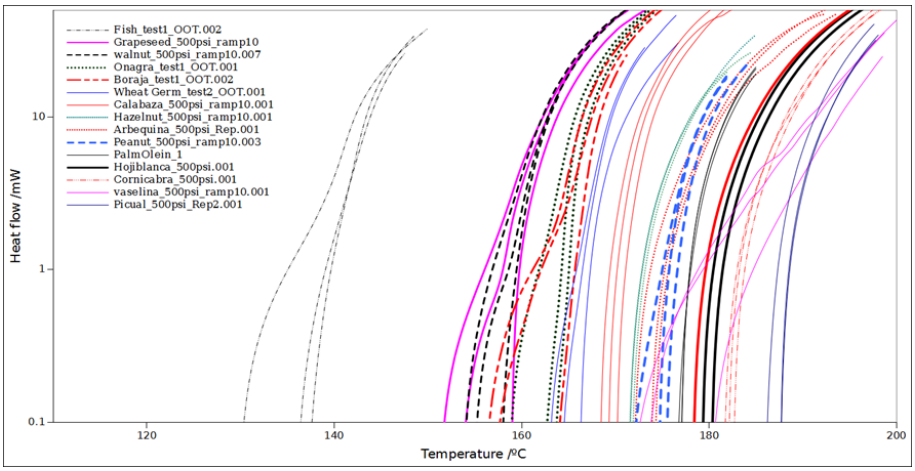


Figura 1: Flujo de calor (heat flow) en función de la temperatura.

3. RESULTADOS Y DISCUSIÓN

Las siguientes características de las curvas DSC han sido utilizadas para distinguir entre diferentes tipos de aceite de oliva: OOT, pendiente máxima, temperatura a máxima pendiente, valor de la pendiente de la curva de flujo de calor frente a la temperatura entre 5 y 10 W g^{-1} °C $^{-1}$ (Beceiro et al., 2011). En la Figura 1 se muestra la superposición de curvas experimentales DSC obtenidas mediante la técnica PDSC para los distintos tipos de aceites estudiados. A partir de la observación del gráfico, se puede decir que algunos aceites se pueden clasificar atendiendo a su estabilidad térmica: el aceite de pescado es menos estable que el aceite de nuez y que el de semillas de uva; borraja y onagra más estables que el aceite de semillas de calabaza, y el aceite de avellana es menos estable que el aceite de oliva variedad hojiblanca. De todas formas se necesitaría analizar más en profundidad para distinguir aceites de semillas de uva y nuez, y entre aceite de borraja, onagra y germen de trigo. En la Figura 2 se observa el biplot correspondiente al análisis de componentes principales, representando las dos primeras componentes el 65% de la variabilidad de los datos. Se han representado las variables originales más relevantes con diferentes grados de transparencia según los valores de la calidad de representación de las variables en las componentes principales. En la Tabla 2 se muestran las etiquetas de cada una de las características estudiadas. Se han identificado dos grupos principales de variables relevantes. El primero está representado por el OOT (X3) y tiene más peso en la primera componente. Relacionadas directamente con el OOT se encuentran los contenidos en C18:0 (ácido esteárico) y C18:1 (ácido oleico). Por lo tanto, se puede inferir que a mayores contenidos de ácidos esteárico y oleico el aceite tenderá a presentar un mayor OOT y, por lo tanto, tenderá a ser más estable con respecto a la oxidación. Este es el caso para los distintos tipos de aceite de oliva, los más estables, incluyéndose en este grupo también el aceite de avellana. Por otro lado, existe un segundo grupo de variables relevantes, relacionadas con la componente 2, prácticamente independientes de las anteriores, relacionadas sólo con la composición del aceite, representadas por el total de ácidos saturados, X10, e insaturados, X11, además del contenido en ácido palmítico, X16. Es importante destacar que el biplot muestra que ambas variables son prácticamente independientes: el grupo donde está el OOT representa la estabilidad a la degradación por oxidación,

mientras el segundo grupo está definido más bien por características de composición, contenido en ácidos saturados e insaturados, que no tienen relación con la resistencia a la degradación por oxidación.

Por otro lado, también es interesante destacar que los diversos tipos de aceite se distinguen sobre todo por la forma en la que se oxidan, y por los ácidos que influyen en la estabilidad a la oxidación, más que por su composición en general. Esa conclusión tendría que ser refrendada por posteriores análisis incluyendo las composiciones reales de cada tipo de aceite, en lugar de las obtenidas a partir de otros trabajos publicados en la literatura científica de este campo.

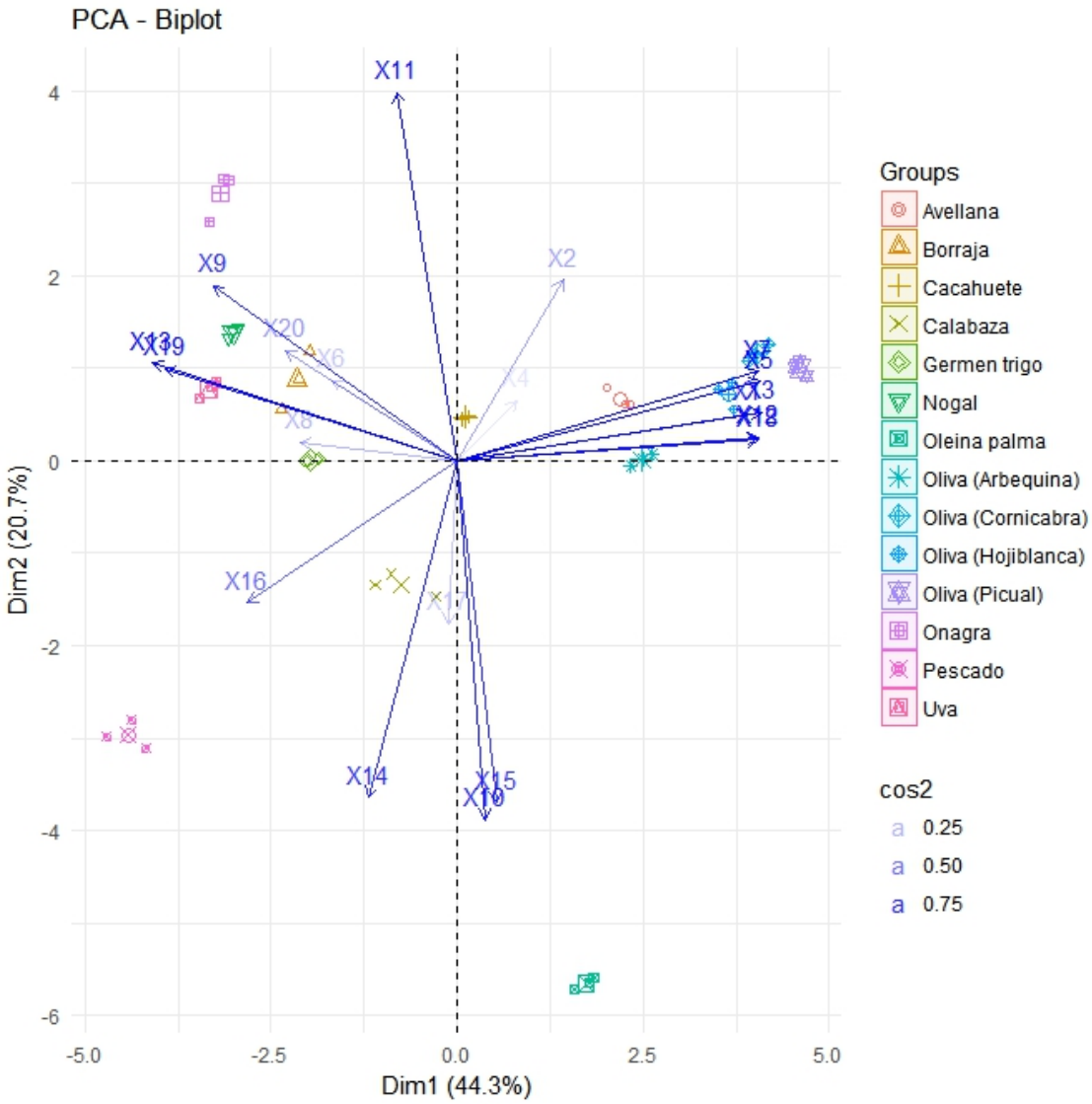


Figura 2: Biplot del análisis de componentes principales a partir del vector de variables originales asociadas a la estabilidad a la oxidación y composición de los aceites. Se han representado las variables originales más relevantes con diferentes grados de transparencia según los valores de la calidad de representación de las variables en las componentes principales.

La Figura 3 muestra la aplicación del algoritmo clúster jerárquico con detección automática del número de particiones (panel izquierdo) y el resultado de la clasificación para 4 grupos identificados (panel derecho). Se puede observar dos grupos principales diferenciados en su resistencia a la estabilidad térmica, caracterizado por estar compuesto de aceites comestibles como son todas las variedades de aceite de oliva, así como el de cacahuete. Se incluye aquí también el aceite de avellana, ya muy cerca del segundo gran grupo compuesto mayormente por aceites con aplicaciones cosméticas y medicinales como el aceite de calabaza, borraja o nogal. Los dos clústeres restantes están relacionados con sólo un tipo de

aceite cada uno: el aceite de uva (cosmético y medicinal), en el semieje horizontal común al grupo de los aceites cosméticos (semieje izquierdo), y el aceite de oleína de palma, utilizado para cocinar y situado en el semieje derecho correspondiente a los aceites comestibles. El diferente contenido en ácidos saturados e insaturados es lo que condiciona que estos últimos se distingan en grupos aparte de los dos grupos principales.

Tabla 2: Características relevantes de las curvas DSC y sustancias constituyentes de los aceites, con sus respectivas etiquetas.

Característica de las curvas DSC	Temperatura de inicio (ºC)	X1
	Temperatura precio (ºC)	X2
	OOT (°C)	X3
	Máxima pendiente curva DSC ($Wg^{-1}s^{-1}$)	X4
	Temperatura a la máx. pendiente de la curva DSC (ºC)	X5
	Pendiente entre 5-10 ($Wg^{-1}ºC^{-1}$)	X6
	Temperatura de pico (°C)	X7
Composición, contenidos en diverso ácidos	sum	X8
	Ratio C18:2/C16:0	X9
	Total ácidos saturados	X10
	Total ácidos insaturados	X11
	MU	X12
	PU	X13
	C14:0	X14
	C16:0	X15
	C16:1	X16
	C18:0	X17
	C18:1	X18
	C18:2	X19
	C18:3	X20

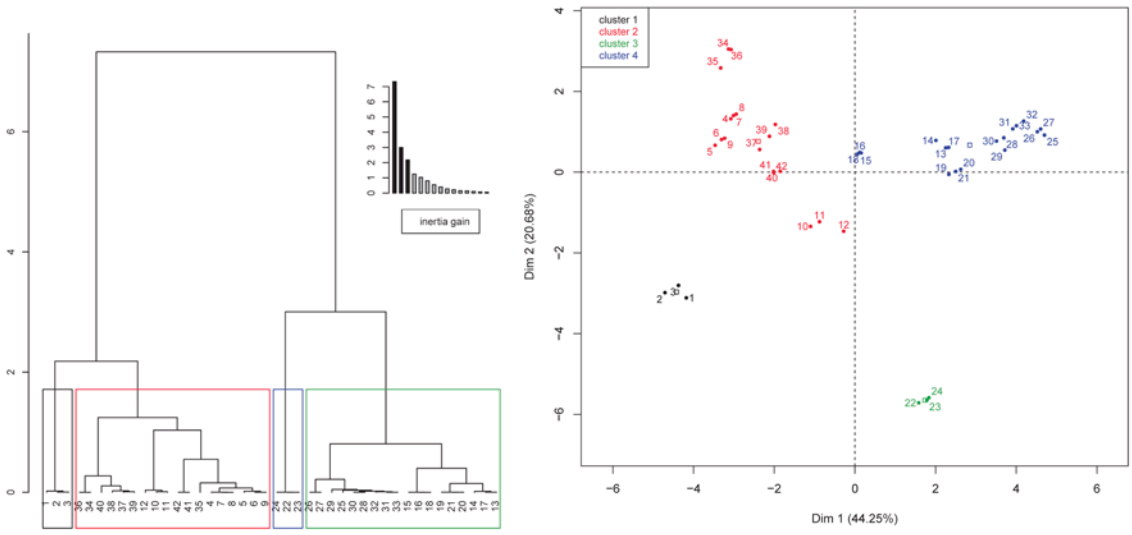


Figura 3: Dendograma correspondiente al análisis clúster jerárquico y obtención de número de particiones óptimo a partir del cálculo de la inercia de cada grupo (izquierda). Grupos identificados en el gráfico de componentes principales (derecha).

REFERENCIAS

Aparicio R., Luna G. (2002) Characterisation of monovarietal virgin olive oils, European Journal of Lipid Science and Technology, 104, 9–10, 614–627.

Dubois V., Breton S., Linder M., Fanni J., Parmentier M. (2007) Fatty acid profiles of 80 vegetable oils with regard to their nutritional potential, European Journal of Lipid Science and Technology, 109, 710–732.

FAO, Codex Standards for Fats and Oils from Vegetable Sources, CODEX-STAN 210 – 1999. Available: http://www.fao.org/docrep/004/y2774e/y2774e04.htm#bm4.1. [Accessed: 18-May-2015].

Ferreira E. M. et al. (2014) Growth, feed intake, carcass characteristics, and meat fatty acid profile of lambs fed soybean oil partially replaced by fish oil blend, Animal Feed Science and Technology, 187, 9–18.

Kassambara A. (2017). Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra (Vol. 2). STHDA.

Koushki M., Nahidi M., Cheraghali F. (2015) Physico-chemical properties, fatty acid profile and nutrition in palm oil, Journal of Paramedical Sciences.vol. 6, 117–134.

Maguire L. S., O'Sullivan S. M., Galvin K., O'Connor T. P., O'Brien N. M. (2004) Fatty acid profile, tocopherol, squalene and phytosterol content of walnuts, almonds, peanuts, hazelnuts and the macadamia nut, International Journal of Food Sciences and Nutrition, 55, 3, 171–178.

Lê S., Josse J., Husson F. (2008) FactoMineR: A Package for Multivariate Analysis, Journal of Statistical Software, 25, 1-18.

López-Beceiro J., Artiaga R., Gracia C., Tarrío-Saavedra J., Naya S., Mier J. L. (2011) Comparison of olive, corn, soybean and sunflower oils by PDSC, Journal of Thermal Analysis and Calorimetry, 104, 169–175.

XII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 y 28 de octubre de 2017

A compositional approach to predict small area proportions

María Dolores Esteban[1], María José Lombardía[2], Esther López-Vizcaíno[3] y Domingo Morales[1]

[1]Centro de Investigación Operativa, Universidad Miguel Hernández de Elche
[2]Departamento de Matemáticas, Universidade da Coruña
[3]Instituto Galego de Estatística

ABSTRACT

Compositional data analysis deals with vectors, called compositions, with nonnegative elements representing proportions or counts of some partition of a given population that fulfil a size constraint. Many surveys have categorical variables that produce compositional data after calculating the direct weighted estimators of the domain totals or proportions of categories. For the case of a classification variable with four categories, a trivariate Fay-Herriot model is fitted to the additive log-ratio transformation of the compositions and the corresponding EBLUPs are derived. The model-based estimators of proportions are obtained by applying the inverse additive logistic transformation to the vectors of EBLUPs. This approach allows positive and negative correlations between the direct estimators of the category proportions. This is an advantage with respect to the multinomial mixed models that only permit negative correlations. The mean squared errors of the proposed predictors of category proportions are estimated by parametric bootstrap. The communication presents the required mathematical derivations and the results of three simulation experiments. Simulation 1 investigates the behaviour of the REML estimators of the model parameters (regression parameters and six variance components). Simulation 2 studies the performance of the proposed predictors of category proportions and makes comparisons with the predictors derived under multinomial mixed models. Simulation 3 analyses the implemented parametric bootstrap procedure. Finally, an application to data from a survey on health conditions in the region of Galicia (north-west of Spain) is given.

Key words and sentences: Compositional data, small area estimation, multivariate Fay-Herriot model, prediction of category proportions.

1. COMPOSITIONAL DATA

Any vector $(z_1, \ldots, z_q) \in R^q$ with nonnegative elements representing proportions of some partition of a given population fulfill the constraint $z_1 + \ldots + z_q = 1$. Let us define the simplex

$$\mathcal{S}^{q-1} = \big\{(z_1, \ldots, z_{q-1}) \in R^{q-1} : z_1 > 0, \ldots, z_{q-1} > 0, z_1 + \ldots + z_{q-1} < 1\big\}. \tag{1}$$

The vectors $z \in \mathcal{S}^{q-1}$ are called compositions. For compositions, the Dirichlet probability density function (p.d.f.) is widely used. The correlation structure of the Dirichlet distribution is completely negative, with $\text{corr}(z_i, z_j) < 0$ for every $i \neq j$, and so it cannot be appropriate to data patterns for which some correlations are positive. Further, it cannot support a sufficient degree of compositional dependence. This is why the transformation technique is employed for introducing new p.d.f.'s on $\mathcal{S}^{q-1}$. The additive logratio transformation, $y = \text{alr}(z)$, is

$$y_i = \log(z_i/z_q), \quad i = 1, \ldots, q-1,$$

with inverse the additive logistic transformation, $z = \text{alogist}(y)$, which is the one-to-one transformation from $y \in R^{q-1}$ to $z \in \mathcal{S}^{q-1}$ defined by

$$z_i = \frac{\exp\{y_i\}}{1 + \exp\{y_1\} + \ldots + \exp\{y_{q-1}\}}, \; i = 1, \ldots, q-1; \quad z_q = 1 - z_1 - \ldots - z_{q-1}.$$

2. PREDICTION OF SMALL AREA PROPORTIONS

Consider a population $U = \cup_{d=1}^{D} U_d$ partitioned in D domains U_d. Assume that each domain is partitioned in subsets U_{dk}, $k = 1, \ldots, q$, defined by a classification variable, like "labor force status" or "education level", that classifies units into a finite number of categories. Let N and N_d be the sizes of U and U_d respectively. Consider a study variable taking the values $z_{dkj} = 1$ if the unit j from the domain U_d is in the category k and $z_{dkj} = 0$ otherwise. The target parameters are the domain proportions $\overline{Z}_{dk} = \frac{Z_{dk}}{N_d}$, $d = 1, \ldots, D$, $k = 1, \ldots, q$. For any sample $s \subset U$ extracted from the population, s_d denotes the subsample from U_d of size n_d and w_{dj} represents the sampling weight of the unit j from the subsample s_d. Direct estimators of $\overline{Z}_{dk}$ are

$$\hat{\overline{Z}}_{dk}^{dir} = \frac{\hat{Z}_{dk}^{dir}}{\hat{N}_d^{dir}}, \quad \hat{Z}_{dk}^{dir} = \sum_{j \in s_d} w_{dj} z_{dkj}, \quad \hat{N}_d^{dir} = \sum_{j \in s_d} w_{dj}.$$

Särndal et al. (1992), section 10.3, give estimators of the design-based covariances $\mathrm{cov}_\pi(\hat{Z}_{dk_1}^{dir}, \hat{Z}_{dk_2}^{dir})$, $k_1, k_2 = 1, \ldots, q-1$.

For ease of exposition, we use the notation $z_{dk} = \hat{\overline{Z}}_{dk}^{dir}$, $z_d = (z_{d1}, \ldots, z_{dq-1})'$. This work applies the additive logratio transformation $y_d = \mathrm{alr}(z_d)$ and fits a multivariate Fay-Herriot (MFH) model to the y-data, where the estimated design-based covariances are assumed to be the error term covariance. The fitted MFH model allows introducing plug-in and empirical best (EB) predictors of the category proportions. The plug-in predictors are biased as they are obtained by applying the alogist transformation to the empirical best linear unbiased predictors of the category means in the MFH model. The EB predictors do not require the application of the alogist transformation, they are unbiased and they are approximated by Monte Carlo simulation. The introduced estimators are compared in simulation experiments to the multinomial-based estimators introduced by López-Vizcaíno et al. (2013). A parametric bootstrap procedure is introduced for estimating the mean squared errors of the proposed predictors and the effect of the number of bootstrap resamples is empirically investigated. Finally we give an application to survey data from the Behavioral risk factors information system from Galicia in 2015.

3. CONCLUSIONS

The new compositional approach to small area estimation presents a good performance in simulations and gives additional flexibility to model real data.

REFERENCES

Dantzig, G. B. (1940) On the non-existence of "Student's" hypothesis having power functions independent of σ. Annals of Mathematical Statistics, 11, 186–192.

Aitchison (1986). The statistical analysis of compositional data. Chapman and Hall..

López-Vizcaíno, E., Lombardía, M.J. and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling*, **13**, 153-178.

Särndal, C.E., Swensson, B. and Wretman J. (1992). *Model assisted survey sampling.* Springer-Verlag.

XII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 y 28 de octubre de 2017

A Trivariate Fay-Herriot model to predict small area poverty indicators

María Dolores Esteban[1], María José Lombardía[2], Esther López-Vizcaíno[3], Domingo Morales[1] and Agustín Pérez[4]

[1]Centro de Investigación Operativa, Universidad Miguel Hernández de Elche
[2]Departamento de Matemáticas, Universidade da Coruña
[3]Instituto Galego de Estatística
[4]Departamento de Estudios Económicos y Financieros, Universidad Miguel Hernández de Elche

ABSTRACT

A Trivariate FayHerriot models for estimating small area indicators is introduced. Among the available procedures for fitting linear mixed models, the residual maximum likelihood is employed. The empirical best linear unbiased predictor of the vector of area means is derived. The developed methodology is applied to data from the 2013 Spanish living condition surveys of the region of Valencia. The target of the application is the joint estimation of annual mean equivalised household incomes, poverty proportions and poverty gaps at domain level.

Key words and sentences: Small area estimation, multivariate Fay-Herriot model, prediction of poverty indicators.

1. TRIVARIATE FAY-HERRIOT MODEL

The trivariate Fay-Herriot (TFH) model is defined in two stages. The sampling model is

$$y_d = \mu_d + e_d, \quad d = 1, \ldots, D,$$

where the vectors $e_d \sim N_3(0, V_{ed})$ are independent and the 3×3 covariance matrices $V_{ed} = \left(\sigma_{dij}\right)_{i,j=1,2,3}$ are known. the linking model

$$\mu_d = X_d\beta + u_d, \quad u_d \sim N_3(0, V_{ud}), \quad d = 1, \ldots, D, \tag{1}$$

where the vectors u_d's are independent and independent of the vectors e_d's. The 3×3 covariance matrices V_{ud} depend on 6 unknown parameters, $\theta_1 = \sigma^2_{u1}$, $\theta_2 = \sigma^2_{u2}$, $\theta_3 = \sigma^2_{u3}$, $\theta_4 = \rho_{12}$, $\theta_5 = \rho_{13}$, and $\theta_6 = \rho_{23}$, i.e.

$$V_{ud} = \begin{pmatrix} \sigma^2_{u1} & \rho_{12}\sigma_{u1}\sigma_{u2} & \rho_{13}\sigma_{u1}\sigma_{u3} \\ \rho_{12}\sigma_{u1}\sigma_{u2} & \sigma^2_{u2} & \rho_{23}\sigma_{u2}\sigma_{u3} \\ \rho_{13}\sigma_{u1}\sigma_{u3} & \rho_{23}\sigma_{u2}\sigma_{u3} & \sigma^2_{u3} \end{pmatrix}.$$

Let I_n be the $n \times n$ identity matrix and $\delta_{\ell d}$ be the Kronecker delta. We define

$$y = \underset{1\le d\le D}{\text{col}}(y_d),\ u = \underset{1\le d\le D}{\text{col}}(u_d),\ e = \underset{1\le d\le D}{\text{col}}(e_d),\ u_d = \underset{1\le k\le 3}{\text{col}}(u_{dk}),\ e_d = \underset{1\le k\le 3}{\text{col}}(e_{dk}),$$

$$X = \underset{1\le d\le D}{\text{col}}(X_d),\ Z_d = \underset{1\le \ell\le D}{\text{col}}(\delta_{\ell d} I_3),\ Z = \underset{1\le d\le D}{\text{col}'}(Z_d) = I_{3D},\ V_u = \underset{1\le d\le D}{\text{diag}}\,(V_{ud}),$$

where col and col$'$ are matrix operators stacking by columns and rows respectively.

In matrix form, the trivariate Fay-Herriot model is

$$y = X\beta + Zu + e = X\beta + Z_1u_1 + \cdots + Z_Du_D + e, \tag{2}$$

where $e, u_1, \ldots, u_D$ are independent with distributions

$$e \sim N\left(0, V_e\right), \quad u \sim N(0, V_u) \quad \text{and} \quad u_d \sim N(0, V_{ud}), \quad d = 1, \ldots, D.$$

Under model (2), it holds that

$$E\left(y\right) = X\beta \quad \text{and} \quad V = \text{var}(y) = Z'V_uZ + V_e = V_u + V_e = \underset{1 \leq d \leq D}{\text{diag}}\left(V_d\right),$$

where $V_d = V_{ud} + V_{ed}$, $d = 1, \ldots, D$. Further, the best linear unbiased estimator (BLUE) of β, and the best linear unbiased predictors (BLUP) of u and μ are

$$\hat{\beta}_B = (X'V^{-1}X)^{-1}X'V^{-1}y, \ \hat{u}_B = V_uZ'V^{-1}(y - X\hat{\beta}_B), \ \hat{\mu}_B = X\hat{\beta}_B + Z\hat{u}_B. \tag{3}$$

The model is fitted by the REML method. The output of REML Fisher scoring algorithm, $\hat{\beta}$ and $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_6)$, are the REML estimators of β and $\theta = (\theta_1, \ldots, \theta_6)$ respectively. By plugging $\hat{\theta}$ in V_u, we get $\hat{V}_u = V_u(\hat{\theta})$ and $\hat{V} = \hat{V}_u + V_e$. By substituting $\hat{V}_u$ in (3), we obtain the EBLUP of $\mu = X\beta + Zu$, i.e.

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y, \quad \hat{u} = \hat{V}_uZ'\hat{V}^{-1}(y - X\hat{\beta}), \quad \hat{\mu} = X\hat{\beta} + Z\hat{u}. \tag{4}$$

The EBLUP $\hat{\mu}$ is employed to estimate the domain means of the target variables.

2. ESTIMATION OF POVERTY INDICATORS

The Spanish Living Conditions Survey, promoted by the Spanish Statistical Office, is designed to obtain reliable direct estimates at the autonomous community level. The objective of this work is the modeling and estimation of indicators of poverty and living conditions by using estimation techniques in small areas based on a trivariate Fay-Herriot model. To do this, we calculate direct estimators (vector y_d in the TFH model) of some indicators of income and poverty referring to comarcas (counties) of the the autonomous community of Valencia crossed by sex for the reference year 2013. We also estimate the design-based variances and covariances of the direct estimators to construct the matrix V_{ed} of the TFH model. Similarly as Esteban et al. (2012) and Benavent and Morales (2016), this paper investigates the small area estimation of the indicators: poverty proportion, poverty gap and annual mean equivalised household income.

3. CONCLUSIONS

The multivariate small area estimation based on the TFH model presents a good performance in relation with the univariate approach and gives additional flexibility to model real data.

REFERENCES

Benavent, R. and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, **13**, 372-390.

Esteban, M.D., Morales, D., Pérez, A. and Santamaría L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, **56**, 2840-2855.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

COMPARING ROC CURVES IN THE PRESENCE OF COVARIATES

Arís Fanjul Hevia[1], Wenceslao González Manteiga[1] and Juan Carlos Pardo Fernández[2]

[1] Dpt. de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela
[2] Dpt. de Estadística e Investigación Operativa, Universidad de Vigo

SUMMARY

The receiver operating characteristic (ROC) curve is a statistical tool that analyses the accuracy of a diagnostic test in which a variable is used to decide whether an individual is healthy or not. In a medical environment it is usual the situation in which more than one variable is available to diagnose a certain disease. In such cases, the ROC curve analysis is a commonly accepted way to compare the accuracy of the different methods of diagnosis. Along with the diagnostic variables it is usual to have information of some covariates. In those situations it is advisable to incorporate that information into the study, as the performance of the ROC curves can be affected by them. Several articles have been written concerning the estimation of the ROC curve incorporating covariate information, as well as the comparison of two or more ROC curves without covariates. However, to our knowledge, little has been done to combine these subjects. Here, we discuss a way of comparing different ROC curves taking into account the presence of one continuous covariate and focusing on a nonparametric approach.

Keywords: AUC; bootstrap; conditional ROC curve.

1. INTRODUCTION

The correct diagnosis of a patient plays a critical role in any medical study. In order to select an appropriate diagnostic variable, one must take into account the two types of error measurements involved in the decision process: the sensitivity (the ability of correctly detecting the condition of interest) and the specificity (the ability of correctly diagnosing a patient as healthy). Those two concepts are directly related to the cumulative distribution function of the groups in which we desire to divide the population (usually called healthy and diseased population).

The receiver operating characteristic (ROC) curve is a statistical tool that represents the sensitivity of a diagnostic test against the complementary of its specificity. It is a commonly accepted way of analysing the discriminatory capacity of a diagnostic variable, and it has been involved in the solution of different problems (Pepe, 2003; Krzanowski & Hand, 2009). The most common way of representing this curve is by the formula:

$$ROC(t) = 1 - F(G^{-1}(1-t)), \quad 0 \leq t \leq 1,$$

where F and G are the cumulative distribution functions of the diagnostic variable in the diseased and in the healthy population respectively.

One of the main applications of these kind of curves is the comparison between different diagnostic markers. In a medical environment, a situation may arise where more than one method is available to diagnose a certain disease. In those situations, one would like to determine which method of diagnosis is the most accurate, and the ROC curve analysis provides the appropriate tools to make that comparison.

Moreover, the presence of a covariate related to the diagnostic variable can affect the discriminating power of the ROC curve. What may appear as an appropriate diagnostic test for a certain value of one covariate might not be so good at discriminating other values of that covariate. For

example, the diagnosis of diabetes by glucose measurements tends to worsen with age (González-Manteiga et al. 2011).

Given the relevance of the ROC comparison methods and the influence of covariates in the performance of the diagnostic tests, it is only natural to combine those two methodologies. We would like to be able to compare two or more ROC curves conditioned to the value of one covariate. In the previous example, this could be useful to check if, at a certain age, the discriminatory capacity of the glucose is the same for men and women (thus, we would compare the ROC curve of the men's population the ROC curve correspondent to the women's conditioned to the age of each individual).

In this work we first go through the ROC comparison methods that exist in the literature (Section 2) and the different ways of taking into account the information of the covariates to estimate the ROC curve (Section 3), focusing always on the nonparametric approach. The final goal is to develop a test capable of comparing two or more ROC curves for a fixed value of a continuous covariate (Section 4). Simulations will be run to check the nominal level and the power of the test.

2. COMPARISON OF TWO OR MORE ROC CURVES WITHOUT COVARIATES

In order to compare the accuracy or K different diagnostic variables, we check the hypotesis

$$H_0 : ROC_1 = ... = ROC_K.$$

This problem has been studied widely among the last decades. The first proposals that appeared in the literature address the issue through the comparison of the area under the ROC curves (AUCs) (DeLong et al. 1988). After that, other methods were developed to test the equality among ROC curves through some equivalent expression (Venkatraman & Begg, 1996; Venkatraman, 2000; Antoch et al. 2010). One of the most recent approaches use the fact that the ROC curve can be viewed as a cumulative distribution function (CDF) to make the comparison (Martínez-Camblor et al. 2011 and 2013).

Some of these proposals are parametric approaches, others are non-parametric. Some of them are prepared to deal with paired data, some others do not, while others can compare more than two ROC curves. As can be noted, there are quite an amount of different methodologies that approach the problem from different perspectives, as well as several ways to determine or approximate the distribution behind those methods. In Fanjul-Hevia & González-Manteiga (2016) a simulation study is made to compare several methodologies of ROC curve comparison, all of them without covariates.

3. ESTIMATION OF THE CONDITIONAL ROC CURVE

When some covariate information is available, a natural extension of the ROC curve for continuous diagnostic test is the conditional or covariate-specific ROC curve. Given a covariate value x, it is defined as

$$ROC^x(t) = 1 - F(G^{-1}(1-t|x)|x), \quad 0 \leq t \leq 1.$$

Different approaches have been proposed in the literature to deal with the estimation of the conditional ROC curve (Pardo-Fernández et al. 2014). One obvious way of approaching the problem is to construct a estimator based on conditional distribution functions (López-de Ullibarri et al. 2008). An alternative is to incorporate the covariates to the ROC curve through regression models, either by direct (Rodríguez-Álvarez et al. 2011b) or induced (Gonzlez-Manteiga et al. 2011; Rodríguez-Álvarez et al. 2011a) methodologies.

Each approach has its own advantages. Here we focus on the last one, the induced methodology in ROC analysis. It consists on modelling the effect of the covariates (X^F and X^G) through regression models of the classification variable in each population (Y^F and Y^G) separately:

$$Y^F = \mu^F(X^F) + \sigma^F(X^F)\varepsilon^F,$$
$$Y^G = \mu^F(X^G) + \sigma^F(X^G)\varepsilon^G,$$

where, for $d = F, G$, $\mu^d(\cdot) = E(Y^d|X^d = \cdot)$, $(\sigma^d)^2(\cdot) = Var(Y^d|X^d = \cdot)$ and the error ε^d is independent of X^d. In González-Manteiga et al. (2011) the estimating process makes use of the estimation of the distribution of the regression errors, and several theoretical results concerning the asymptotic behaviour of the estimator are shown.

4. COMPARING ROC CURVES WITH COVARIATES

The main goal here is to build a method capable of testing the equality among K ROC curves given a covariate value x:

$$H_0 : ROC_1^x = ... = ROC_K^x.$$

In order to do so, we combine methods discussed in the previous sections. Given a certain covariate value x, we estimate the conditional ROC curve ROC_x^k for each $k = 1, ..., K$, and we use them to build a statistic that captures the distance between them:

$$S_N^x = \sum_{k=1}^{K} d_N \left(\sqrt{g_k N_k} \{ \widehat{ROC}_k^x(t) - \widehat{ROC}^x(t) \} \right),$$

where N_k is the sample size of the $k-th$ population that is being compared; g_k is a bandwidth parameter involved in the estimation of the conditioned ROC curve; d_N is some sort of distance function ($d_N(g(t)) = ||g(t)||_\infty$ or $d_N(g(t)) = \int g(t)^2 dt$); and $\widehat{ROC}^x = K^{-1} \sum_{k=1}^{K} \widehat{ROC}_k^x$.

As the asymptotic distribution of this statistic is difficult to estimate (it depends on the unknown densities and cumulative distribution functions of the healthy and diseased populations), a bootstrap algorithm is proposed. A simulation study is carried out to check the power of the test under different scenarios.

5. CONCLUSIONS

One of the main applications of the ROC curves is the comparison between different diagnostic tests, but, to our knowledge, little has been done to extend that comparison to the case in which we have covariates. In this work we approach this problem by combining existing methods for comparing ROC curves without covariates and for estimating the conditioned ROC curve.

The methodology used here can be extended to test whether a covariate has a significant effect on the ROC curve or not, meaning

$$H_0 : ROC^x(t) = ROC(t) \quad \forall t \in (0, 1).$$

Furthermore, in a future work longitudinal data could be considered, with the goal of taking all the advantage possible from the increasingly detailed data that real life problems can provide.

REFERENCES

Antoch, J., Prchal, L. & Sarda, P.(2010). Nonparametric comparison of ROC curves: Testing equivalence. IMS Collections, 7, 1224.

DeLong, E.R., DeLong, D.M. & Clarke-Pearson, D.L. (1988). Comparing the Areas Under two or more correlated Receiver Operating Characteristic Curves: a nonparametric approach. Biometrics, 44, 837-845.

Fanjul-Hevia, A. & González-Manteiga, W. (2016). A comparative study of methods for testing the equality of two or more ROC curves. Manuscript submitted for publication.

González-Manteiga, W., Pardo-Fernández, J.C. & Van Keilegom, I. (2011). Scandinavian Journal of Statistics, 38, 169-184.

Krzanowski W.J. & Hand D.J. (2009). ROC Curves for continuous data. Chapman & Hall/CRC.

López-de Ullibarri, I., Cao, R., Cadarso-Suárez, C. & Lado, M.J. (2008). Nonparametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. Computational Statistics and Data Analysis, 52, 2623-2631.

Martínez-Camblor, P., Carleos, C. & Corral, N. (2011). Powerful nonparametric statistics to compare k independent ROC curves. Journal of Applied Statistics, 38(7), 13171332.

Martínez-Camblor, P., Carleos, C. & Corral, N. (2013). General nonparametric ROC curve comparison. Journal of the Korean Statistical Society, 43, 71-81.

Pardo-Fernández, J.C., Rodríguez-Álvarez, M.X., & Van Keilegom, I. (2014). A review on ROC curves in the presence of covariates. Statistical Journal, 12 (1), 21-41.

Pepe, M.S. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford Statistical Science Series.

Rodríguez-Álvarez M.X., Roca-Pardiñas, J. & Cadarso-Suárez, C. (2011a). ROC curve and covariates: extending induced methodology to the non-parametric framework. Stat Comput, 21, 483-499.

Rodríguez-Álvarez, M.X., Roca-Pardiñas, J. & Cadarso-Suárez, C. (2011b). A new flexible direct ROC regressioon model: Application to the detection of cardiovascular risk factors by anthropometric measures. Computational Statistics and Data Analysis, 55, 3257-3270.

Venkatraman, E.S. & Begg, C.B. (1996). A distribution- free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika, 83(4), 835-848.

Venkatraman, E.S. (2000). A permutation test to compare receiver operating characteristic curves. Biometrics, 56, 1134-1138.

Mobilidade e coñecemento de idiomas dos titulados no Sistema Universitario de Galicia nos cursos 2011-2012 e 2012-2013

Mª Carmen Fernández Montes[1], José Eduardo López Pereira[1], Mª Patrocinio Morrondo Pelayo[1,] Pedro Faraldo Roca[2], Rosa María Crujeiras Casais[2], , María José Lombardía Cortiña[3], Salvador Naya Fernández[3], José María Matías Fernández[4], Verónica Carreira Figueiras[5], Belén María Fernández de Castro[6], Adela Martínez Calvo[7]

1 Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG)
2 Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela
3 Departamento de Matemáticas. Universidade da Coruña
4 Departamento de Estatística e Investigación Operativa. Universidade de Vigo
5 ABANCA
6 CESUGA. Universidad San Jorge
7 Instituto Tecnolóxico de Matemática Industrial

RESUMO

A Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG) comenzou o seu longo camiñar no sendeiro da mellora continua da calidade do Sistema Universitario de Galicia (SUG) no ano 2001, articulando actividades encamiñadas á orientación e apoio ás institucións universitarias. Entre elas, como actividade consolidada cunha periocidade anual atópase a realización de estudos de inserción laboral aos titulados no SUG de 1º, 2º ciclo e graduados. Coa nova configuración do mapa de titulacións no SUG, froito da adaptación ao Espazo Europeo de Educación Superior (EEES), incrementase cada vez máis a representatividade dos titulados en graos.

No contexto do EEES, tras a aprobación no 2007 da estratexia de internacionalización do EEES, o Comunicado Ministerial de Bucarest de 2012 aprobou a "Mobility for Better Learning. Mobility Strategy 2020 for the European Higher Education Area" onde se establece que se deben desenvolver estratexias nacionais e institucionais de internacionalización e mobilidade da educación superior. A mobilidade de estudantes xoga un papel importante na internacionalización das universidades, polo que se pon de manifesto a importancia de recadar información relevante sobre a mobilidade dos estudantes universitarios e directamente ligada a mesma, sobre o seu coñecemento de idiomas.

Neste traballo presentase unha análise da participación dos estudantes en programas de mobilidade, tanto nacional como internacional, centrándose no tipo de programa e na localización da universidade de acollida. Así como se analiza que idiomas coñecen os estudantes e son quen de usar a maiores das linguas oficiais do SUG (galego e castelán).

Palabras e frases chave: Sistema Universitario de Galicia, inserción laboral, titulados universitarios, mobilidade, idiomas.

1. INTRODUCIÓN

A periodicidade anual destes estudos garante a obtención de históricos de resultados permitindo analizar a evolución no tempo de distintos factores relacionados co ámbito académico e laboral do titulado. Coa actual estrutura do mapa de titulacións: grao, máster e doutoramento, así como froito dos cambios acaecidos na homoxeneización dos programas formativos a nivel europeo, considerouse oportuno realizar unha revisión do cuestionario coa finalidade de recoller información sobre aspectos relevantes na actualidade, como son, a mobilidade e os coñecementos de idiomas. Así, no ámbito académico analízanse aspectos relacionados coa formación académica e a obtención de bolsas de estudo, a realización de prácticas e o traballo durante os estudos, a participación en programas de mobilidade e o coñecemento de idiomas. Neste traballo, centrámonos na análise da mobilidade e o coñecemento de idiomas dos titulados no SUG nos dous últimos estudos realizados.

Na Sección 2 preséntase un resumo dos datos técnicos dos estudos realizados sobre as cohortes de titulados no 2011-2012 e 2012-2013 (véxase Referencias). A análise da mobilidade e o coñecemento de idiomas dos titulados no SUG correspondese coa Sección 3.

2. DATOS TÉCNICOS DOS ESTUDOS

A poboación obxecto de estudo está constituída polos titulados no SUG durante os cursos académicos 2011-2012 e 2012-2013 (EIL1112 e EIL1213, onde as siglas EIL designan "Estudo de Inserción Laboral"). A información básica e de contacto correspondente aos titulados (que coinciden con unidade de mostraxe e unidade informante) é facilitada polas universidades do SUG. Os datos de poboación, mostra e período de realización das distintas enquisas achéganse na Táboa 1.

Curso académico	EIL1112	EIL1213
Poboación	8.106	8.395
Mostra	4.687	4.473
Traballo de campo	**Decembro 2014**	**Decembro 2015**

Táboa 1: Titulados no SUG. Tamaños de poboación e mostra e período do traballo de campo.

Para cada titulación e campus (subpoboación) seleccionouse unha mostra aleatoria, determinando o tamaño para a estimación da proporción (con varianza máxima) e fixando un erro máximo absoluto do 10%, cun nivel de confianza do 95%. Posteriormente, os resultados por ramas de coñecemento e para o total do SUG obtéñense mediante elevación proporcional ao tamaño de cada subpoboación. Considéranse as seguintes ramas de coñecemento: Ciencias da Saúde, Ciencias, Enxeñaría e Arquitectura, Ciencias Sociais e Xurídicas (I, II, III e IV) e Artes e Humanidades (I e II).

As enquisas realizáronse mediante entrevista telefónica, fixando un número máximo de 5 chamadas por titulado, escollendo estes de xeito aleatorio. O cuestionario empregado foi o mesmo para cada un dos estudos. Na Táboa 1 pódese consultar as datas para o desenvolvemento do traballo de campo.

3. ANÁLISE DA MOBILIDADE E O COÑECEMENTO DE IDIOMAS DOS TITULADOS NO SUG (2011-2012 E 2012-2013)

Mobilidade

Nos programas de mobilidade (nacionais e internacionais) de estudantes atópanse, entre outros, como obxectivos a adquisición de aprendizaxes, competencias lingüísticas e experiencias, tanto académicas como culturais, o fomento de cooperación entre institucións universitarias e a transferencia e recoñecemento de créditos.
O 19,46% dos titulados durante o curso académico 2012-2013 participou durante a súa estadía na universidade nalgún programa de mobilidade, obténdose unha porcentaxe similar para o curso académico 2011-2012 (19,28%) como se pode observar na Táboa 2 e Figura 1.

Mobilidade SUG	EIL1112	EIL1213
Si participou en mobilidade	19,28%	19,46%
Non participou en mobilidade	80,72%	80,54%

Táboa 2: Participación en programas de mobilidade durante a titulación.
Comparativa dos estudos: EIL1112 e EIL1213.

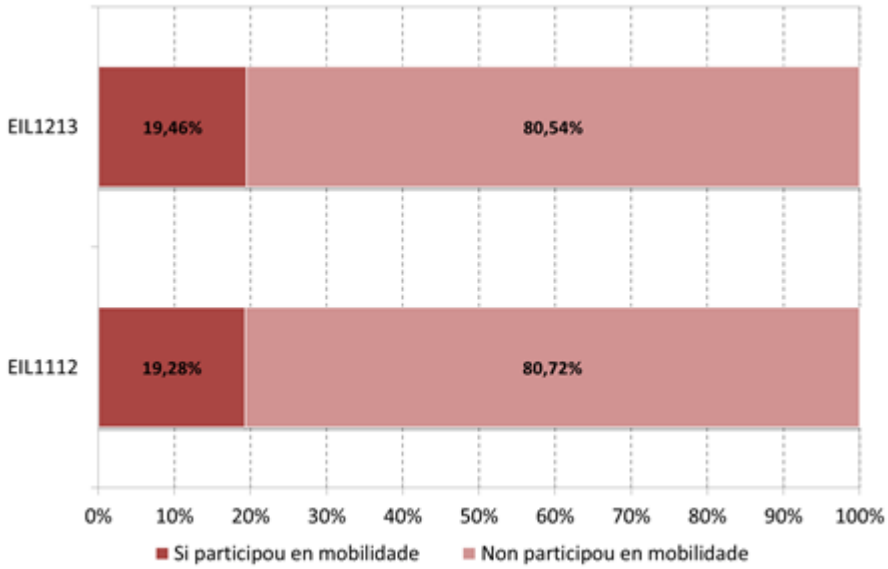


Figura 1: Participación en programas de mobilidade durante a titulación.
Comparativa dos estudos: EIL1112 e EIL1213.

Recolle a Táboa 3 e a Figura 2 o tipo de programa de mobilidade no que participou durante a titulación. Destaca o programa Erasmus, cun 80,24% no curso académico 2012-2013 e un 69,15% no curso académico 2011-2012 fronte aos programas nacionais, Sicue-Séneca cunhas porcentaxes do 16,47% e 15,93%, respectivamente. Así mesmo, destacase o incremento da participación no programa Erasmus no curso académico 2012-2013 con respecto a anterior.

Tipo de programa de mobilidade	EIL 1112	EIL1213
Sicue-Séneca	16,47%	15,93%
Erasmus	69,15%	80,24%
Bilateral	4,64%	6,87%
Outros	16,83%	1,37%

Táboa 3: Tipo de programa de mobilidade nos que participou durante a titulación.
Comparativa dos estudos: EIL1112 e EIL1213.

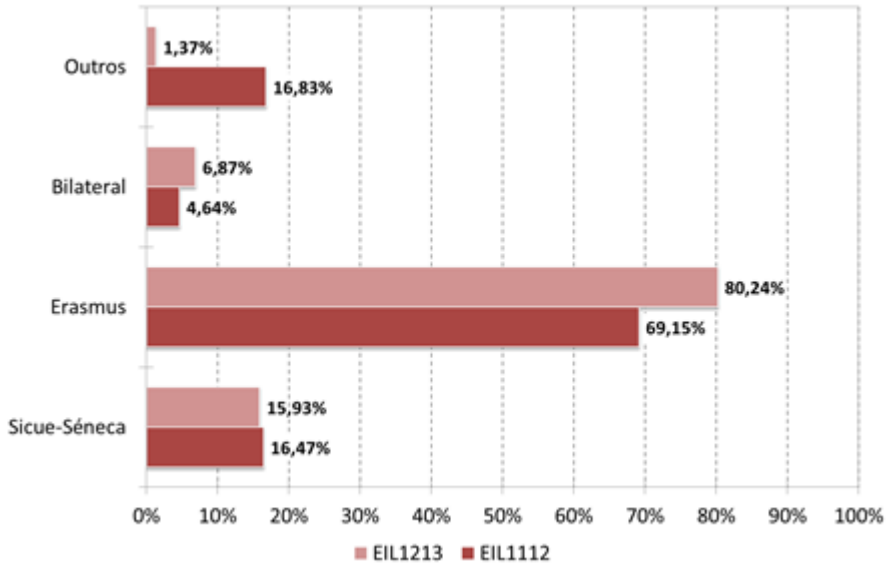


Figura 2: Tipo de programa de mobilidade nos que participou durante a titulación. Comparativa dos estudos: EIL1112 e EIL1213.

Con respecto a localización da universidade de destino Táboa 4 e Figura 3, as porcentaxes gardan coherencia co tipo de programa no que se participou, de xeito o 81,07% (curso académico 2012-2013) e 80,05% (curso académico 2011-2012) dos casos, o seu destino foi unha universidade do resto da Unión Europea.

Localización da universidade á que se desprazou	EIL1112	EIL1213
Resto de España	17,05%	16,17%
Resto da Unión Europea	80,05%	81,07%
Resto do mundo	8,65%	6,89%

Táboa 4: Localización da universidade á que se desprazou co programa de mobilidade durante a titulación. Comparativa dos estudos: EIL1112 e EIL1213.

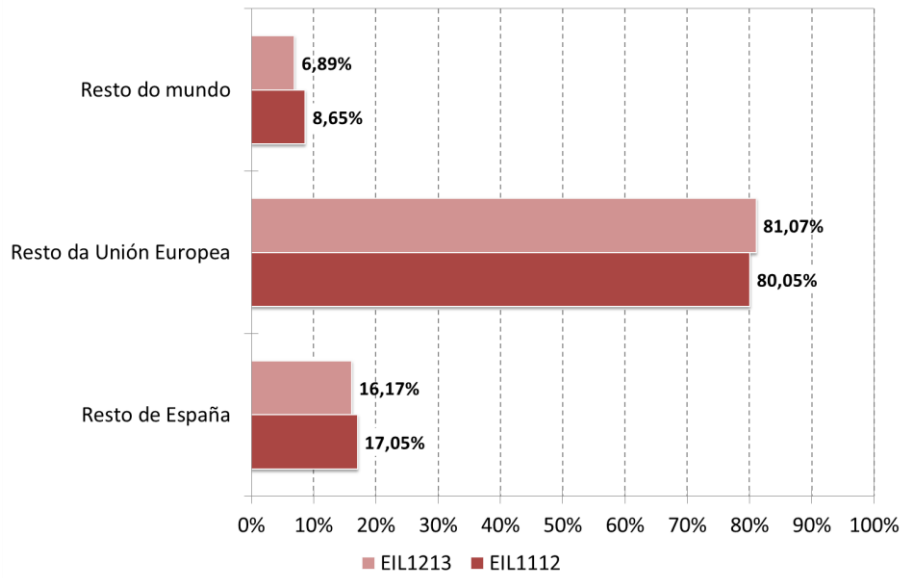


Figura 3. Localización da universidade á que se desprazou co programa de mobilidade durante a titulación. Comparativa dos estudos: EIL1112 e EIL1213.

Coñecemento de idiomas

Un factor determinante á hora de participar nun programa de mobilidade durante a titulación, que directamente contribuirá o éxito do programa, é coñecer o idioma do país no que se localiza a universidade a cal se desprazan os estudantes. Nestes dous últimos estudos, analizouse que idiomas coñecen e son quen de usar a maiores das linguas oficiais do SUG (galego e castelán).
O 93,12% dos titulados durante o curso académico 2012-2013 coñecía algún idioma a maiores das linguas oficias do SUG (galego e castelán), obténdose un 91,36% para o curso académico 2011-2012 como se pode ver na Táboa 5 e Figura 4.

Coñecemento de idiomas adicionais	EIL1112	EIL 1213
Non coñece outro idioma	8,64%	6,88%
Si coñece outro idioma	91,36%	93,12%

Número de idiomas adicionais que coñece	EIL1112	EIL 1213
1 idioma	56,45%	52,88%
2 idiomas	26,58%	29,67%
3 idiomas	6,54%	8,29%
4 idiomas	1,48%	1,96%
Máis de 4 idiomas	0,31%	0,33%

Táboa 5: Coñecemento de idiomas adicionais ás linguas oficiais do SUG e o seu número. Comparativa dos estudos: EIL1112 e EIL1213.

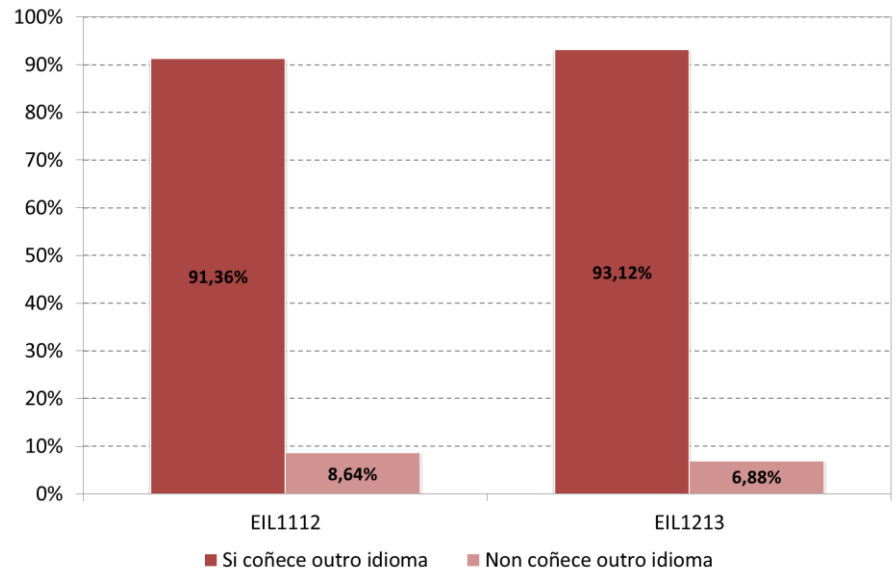


Figura 4. Coñecemento de idiomas adicionais ás linguas oficiais do SUG. Comparativa dos estudos: EIL1112 e EIL1213.

Con respecto ao número de idiomas adicionais que coñecen, na Figura 5 observase que o 56,78% e o 61,79% para os cursos académicos 2012-2013 e 2011-2012, respectivamente, afirmou coñecer só un idioma adicional. Así mesmo, incrementase as porcentaxes de titulados que afirman coñecer dous idiomas, pasando do 29,09% no curso académico 2011-2012 ao 31,86% no curso académico 2012-2013.

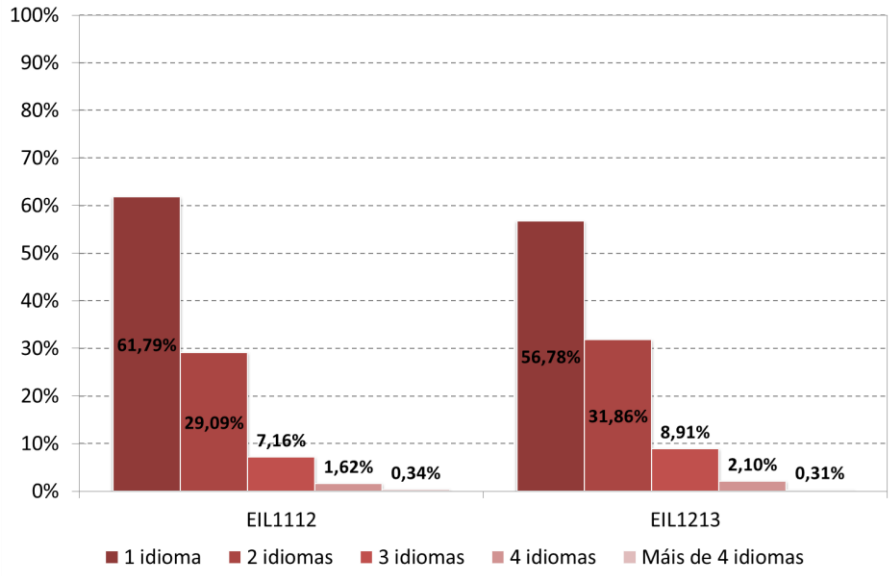


Figura 5. Número de idiomas adicionais que coñece.
Comparativa dos estudos: EIL1112 e EIL1213.

O idioma que os titulados dixeron coñecer mellor é o inglés, seguido de lonxe polo francés e o portugués en ambos estudos. A maioría de definiuse como usuario independente (niveis B1 e B2) no caso do inglés e usuario básico nos outros idiomas estudados.

REFERENCIAS

Faraldo, P., Crujeiras, R.M., Martínez, A., Lombardía, M.J., Naya, S., Matías, J.M., Carreira V., Fernández, B.M., Morrondo, M.P., Pereira, J.E., Fernández, M.C. (2016). Estudo da inserción laboral dos titulados no Sistema Universitario de Galicia 2011-2012. Axencia para a Calidade do Sistema Universitario de Galicia.

Faraldo, P., Crujeiras, R.M., Martínez, A., Lombardía, M.J., Naya, S., Matías, J.M., Carreira V., Fernández, B.M., Morrondo, M.P., Pereira, J.E., Fernández, M.C. (en imprenta). Estudo da inserción laboral dos titulados no Sistema Universitario de Galicia 2012-2013. Axencia para a Calidade do Sistema Universitario de Galicia.

Grupo de Trabajo de Internacionalización de Universidades (2014), Estrategia para la internacionalización de las Universidades Españolas 2015-2020. Ministerio de Educación, Cultura y Deporte.

European Higher Education in the World. EC. July 2013

Mobility for Better Learning. Mobility Strategy 2020 for the European Higher Education Area.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

A inserción laboral dos titulados en Máster: motivación e valoración do título

Mª Carmen Fernández Montes[1], José Eduardo López Pereira[1], Mª Patrocinio Morrondo Pelayo[1] , Jose Carlos de Miguel Domínguez[2], David Rodríguez González[3], Beatriz Valcárcel Aguiar[2], Alexandre Vecino Aguirre[2], Luís Pedro Pereira Andrade[4], Manuel Guisado Tato[5]

1 Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG)
2 Departamento de Economía Cuantitativa. Universidade de Santiago de Compostela
3 Departamento de Empresa. Universidade da Coruña
4 Departamento de Economía. Universidade da Coruña
5 Departamento de Organización de Empresas e Marketing. Universidade de Vigo

RESUMO

Desde o ano 2001, unha das actividades da Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG) é a realización dos estudos da inserción laboral dos titulados no Sistema Universitario de Galicia (SUG). Neste contexto, a ACSUG desenvolveu, de xeito periódico e continuo desde a súa creación, a análise da situación laboral dos titulados no SUG de 1º, 2º ciclo e graduados. Ampliando no ano 2013 o alcance dos seus estudos de inserción laboral aos titulados en másteres oficiais no SUG, en consonancia coa configuración actual dos títulos de educación superior -grao, máster e doutoramento-, froito da progresiva harmonización dos sistemas universitarios esixida no proceso de adaptación ao Espazo Europeo de Educación Superior (EEES).

Estes estudos proporcionan información estatística dos titulados en másteres desde dous puntos de vista: a súa etapa formativa durante a súa estadía na universidade e a súa etapa laboral. Coa definición do Marco Español de Cualificación para a Educación Superior (MECES), establécese o nivel de Máster no nivel 3 do MECES, "no que se inclúen aquelas cualificacións que teñen como finalidade a adquisición por parte dos estudantes dunha formación avanzada, de carácter especializado ou multidisciplinar, orientada á especialización académica ou profesional, ou ben promover a iniciación en tarefas investigadoras", o que pon de manifesto a relevancia dos estudos de máster para a consecución ou especialización nun emprego, ou como iniciación ás tarefas investigadoras previa á realización dun doutoramento.

Neste traballo analizase a motivación principal pola que os titulados universitarios cursaron un máster, a valoración que os titulados realizan sobre distintos elementos relacionados coa planificación e o desenvolvemento, a situación laboral posterior a realización do máster, así como a localización do traballo actual.

Palabras e frases chave: titulados en másteres, inserción laboral, motivación, planificación, desenvolvemento, situación laboral.

1. INTRODUCIÓN

En consonancia con dous dos cinco obxectivos establecidos pola Unión Europea para o ano 2020[1] relacionados co emprego e a educación (o emprego: garantirlle o emprego ao 75% das persoas de 20 a 64 anos; e mais a educación: lograr que cando menos un 40% das persoas de 30 a 34 anos remataran estudos superiores), establecese como finalidade dos estudos de inserción laboral dos titulados en másteres analizar o valor engadido dos másteres na empregabilidade dos titulados universitarios, e proporcionarlles información ás universidades para elaborar estratexias de mellora na planificación dos seus títulos que contribúan ao incremento da empregabilidade dos seus titulados.
Os obxectivos destes estudos son proporcionar información dos titulados universitarios en másteres oficiais no SUG sobre dous aspectos fundamentais: a súa percepción sobre a formación e a universidade (aspectos organizativos e de desenvolvemento do máster, motivación para realizar o máster, etc.) e a súa situación laboral, tanto previa como posterior á realización do máster. Neste traballo, centrámonos na análise da motivación principal que leva a un titulado a cursar un máster e a súa valoración sobre o desenvolvemento do mesmo, así como a situación laboral posterior ao máster e a localización do traballo actual.

2. DATOS TÉCNICOS DOS ESTUDOS

A poboación obxecto de estudo está constituída polos titulados en másteres no SUG correspondentes aos cursos académicos 2009-2010, 2010-2011, 2011-2012 e 2012-2013. A información básica e de contacto correspondente aos titulados en másteres é facilitada polas universidades do SUG, realizando a enquisa aproximadamente aos tres anos de remataren os seus estudos de máster. Na Táboa 1 pódese consultar os datos de poboación, mostra e as datas para o desenvolvemento do traballo de campo.

Curso académico	EILMasteres0910	EILMasteres1011	EILMasteres1112	EILMasteres1213
Poboación	1.633	2.099	2.728	2.432
Mostra	1.261	1.526	1.907	1.723
%	77,22%	72,70%	69,90%	70,85%
Traballo de campo	maio-xuño 2013	xuño-xullo 2014	xuño-xullo 2015	febreiro-marzo 2016

Táboa 1: Tamaño da poboación e mostra dos titulados en másteres no SUG e período do traballo de campo.

A poboación e a mostra foron agrupadas nas seguintes ramas de coñecemento: Artes e Humanidades, Ciencias, Ciencias da Saúde, Ciencias Sociais e Xurídicas e Enxeñaría e Arquitectura. Considerouse a rama de adscrición do máster no proceso de verificación para realizar a agrupación nas distintas ramas de coñecemento.

Realizouse un estudo exhaustivo, contactando con todos os individuos que constituían a poboación para intentar acadar a maior representatividade para cada un dos másteres no SUG, polo que non se realizou unha mostraxe aleatoria. Finalmente, o erro de mostraxe para os titulados en másteres no SUG varía entre o +/-0,99% e o +/-1,27% para cada un dos estudos, cun nivel de confianza do 95%.

A recollida de información realizouse mediante enquisa telefónica empregando a metodoloxía CATI (enquisa telefónica asistida por ordenador), cun número máximo de oito chamadas con cada unidade da mostra, realizadas en distintos días e franxas horarias. O cuestionario empregado foi moi similar para cada un dos estudos, o que permite realizar unha comparativa dos resultados acadados (véxase Referencias).

[1] *Europa 2020: a estratexia europea do crecemento*

2. ANÁLISE DA MOTIVACIÓN PARA REALIZAR O MÁSTER, VALORACIÓN DO SEU DESENVOLVEMENTO E SITUACIÓN LABORAL POSTERIOR AO MÁSTER

Motivación para realizar o Máster

Na Táboa 2 e Figura 1 pódese observar que o motivo máis destacado para realizar estudos de máster é complementar a formación académica para obter unha maior especialización no mercado de traballo, 64,3% (EILMasteres1213), 63,7% (EILMasteres1112), 57,6% (EILMasteres1011) e 58,8% (EILMasteres0910). Á súa vez, complementar a formación académica previa para realizar o doutoramento e dedicarse á investigación sitúase como segundo motivo de importancia, con valores comprendidos entre o 25,4% (EILMasteres1213) e 31,9% (EILMasteres1112), seguido dos titulados en másteres que afirmaron que o máster era preciso ou obrigatorio para o seu futuro emprego ou perspectiva laboral (profesores de ensinanza secundaria, avogados, etc.) con valores inferiores ao 19% en todos os estudos.

Motivacións para realizar o máster	EILMasteres0910	EILMasteres1011	EILMasteres1112	EILMasteres1213
Complementar a formación académica (Investigación)	29,0%	30,1%	31,9%	25,4%
Complementar a formación académica (Traballo)	58,8%	57,6%	63,7%	64,3%
Detectar necesidades formativas na experiencia laboral anterior	4,6%	5,4%	3,4%	5,1%
Era necesario ou obrigatorio para o futuro emprego	18,7%	17,0%	15,0%	18,5%
Outros	1,5%	5,4%	2,4%	2,6%

Táboa 2: Motivos para realizar o máster. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213

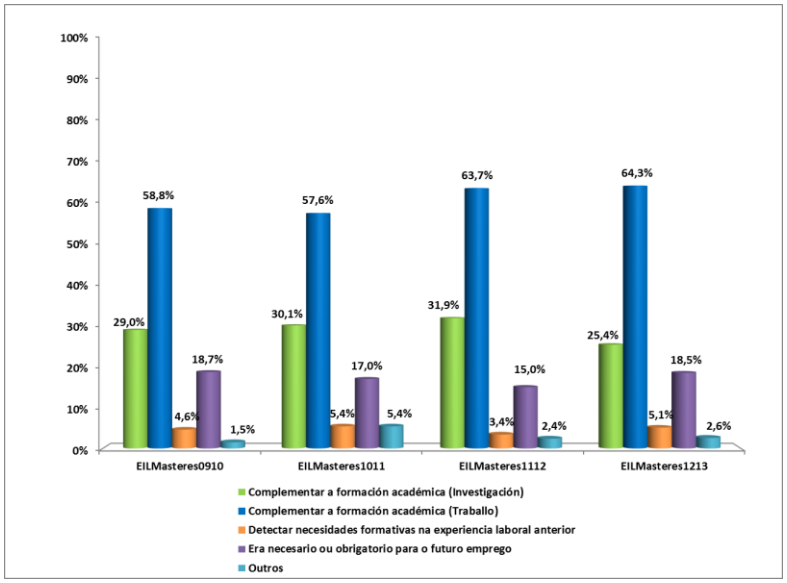


Figura 1: Motivos para realizar o máster. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213

Valoración do desenvolvemento do Máster

Co obxecto de analizar a experiencia dos titulados en másteres con respecto a súa planificación e desenvolvemento, pedíuselles que valoraran aspectos tales como a planificación das ensinanzas, as competencias, as metodoloxías de ensino-aprendizaxe, os sistemas de avaliación, o profesorado, os recursos materiais, a coordinación e o traballo de fin de máster. As análises acadadas para a planificación das ensinanzas, competencias, profesorado, coordinación e o Traballo de Fin de Máster pódese consultar os valores obtidos desde a Figura 2 ata a Figura 6. En relación coa valoración da planificación das ensinanzas (estrutura do máster, horarios, distribución da carga de traballo, etc.) e as competencias (coñecementos, destrezas e habilidades) máis do 50% dos titulados de máster teñen unha valoración positiva dela, e valóraa como boa ou moi boa.
Con respecto á valoración do profesorado, o 63,8% dos titulados de máster no curso académico 2012-2013 valórano como bo ou moi bo, incrementándose a súa valoración positiva nos tres últimos estudos.
Outro aspecto a valorar dentro da organización do máster, é a súa coordinación, aspecto de maior relevancia nos másteres interuniversitarios. As valoracións sobre a coordinación foron mellorando nos respectivos estudos, 35,4% (EILMasteres0910) ata o 46,5% (EILMasteres1213) que consideran que hai unha moi boa ou boa coordinación.
O Traballo Fin de Máster (contidos, desenvolvemento, etc.) xunto co profesorado son os dous ítems mellor valorados. Os titulados en másteres no SUG dan unha valoración moi boa ou boa superior ao 60% en todos os estudos do Traballo Fin de Máster.

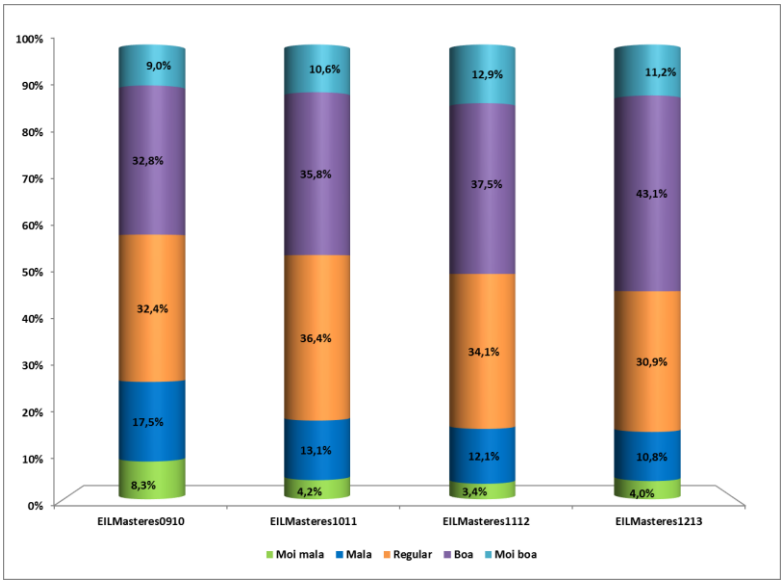


Figura 2: Valoración da planificación das ensinanzas. Resultados para o total do SUG. Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

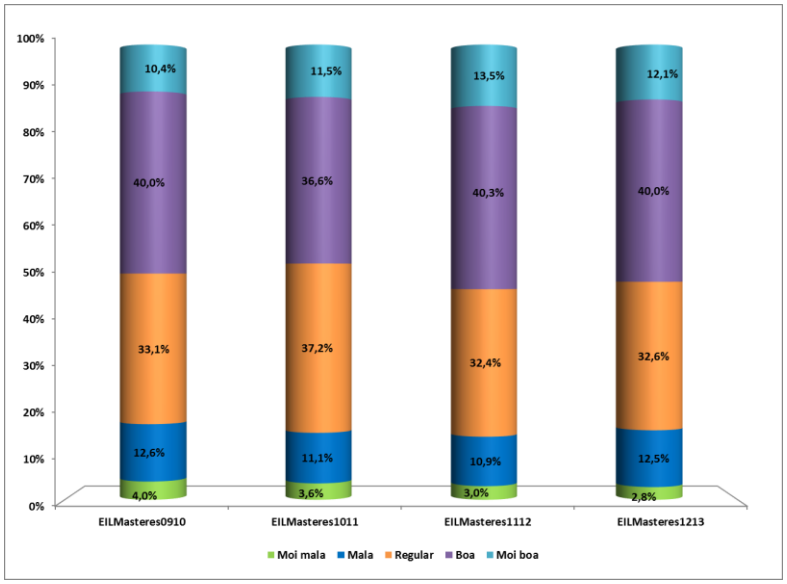


Figura 3: Valoración das competencias. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

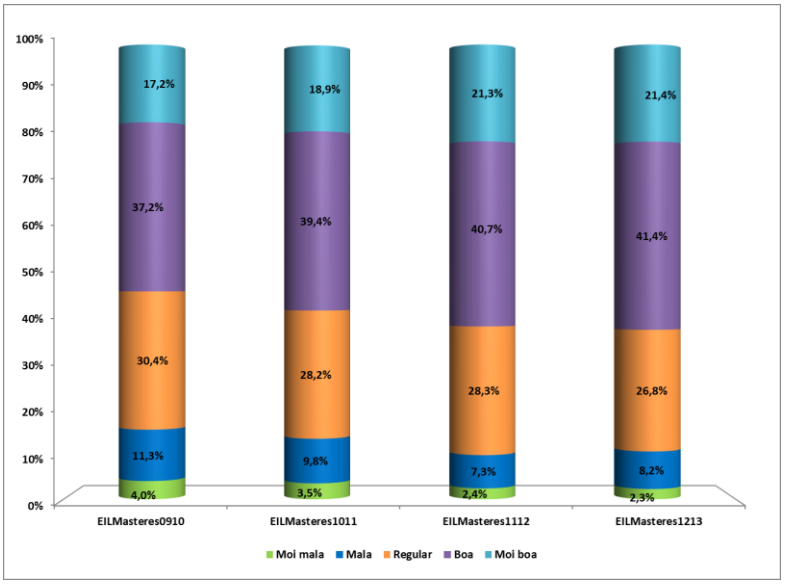


Figura 4: Valoración do profesorado. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

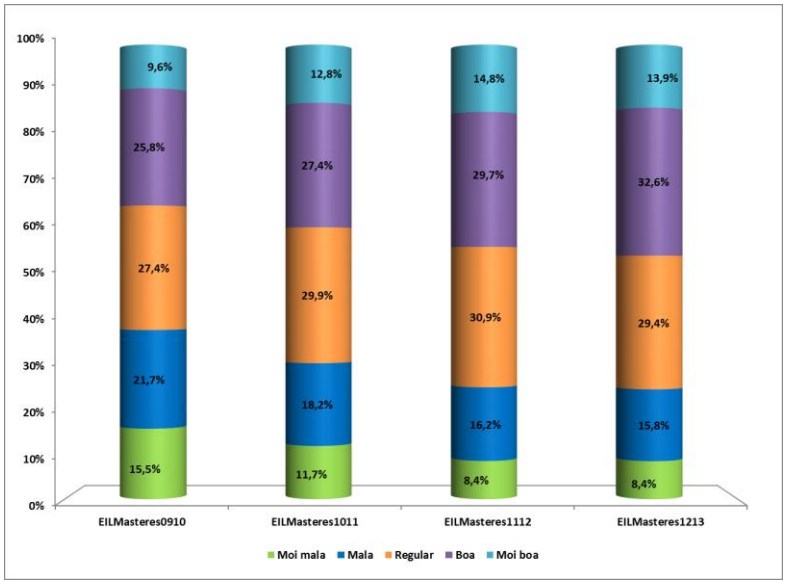


Figura 5: Valoración da coordinación. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

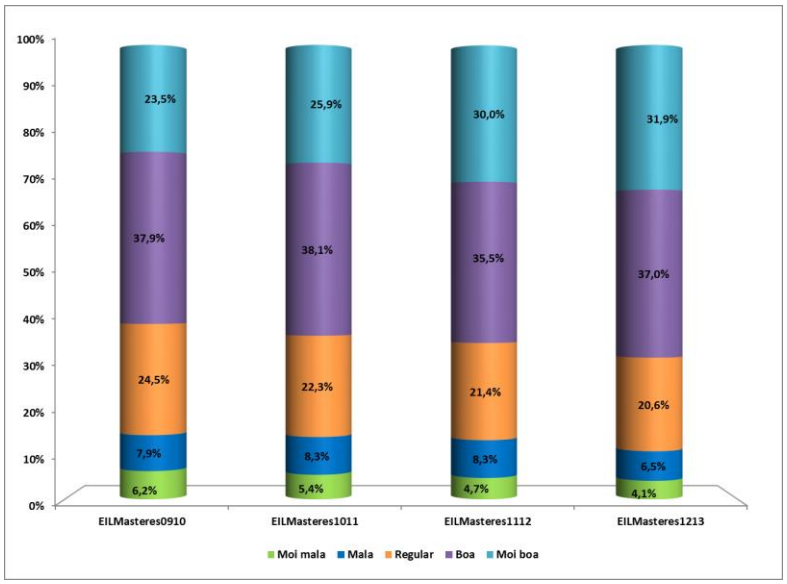


Figura 6: Valoración do Traballo de Fin de Máster. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

Situación laboral e lugar de traballo

Na Figura 7 e na Figura 8 presentase a situación laboral e o lugar de traballo no momento de realizar a enquisa dos titulados en máster. Un 73,8% (EILMasteres1213), 77,6% (EILMasteres1112), 73,6% (EILMasteres1011) e 73,7% (EILMasteres0910) dos titulados en másteres no SUG atópase traballando no momento de realizar a enquisa, dos cales máis do 50%, en todos os estudos, traballan nun emprego relacionado co máster.

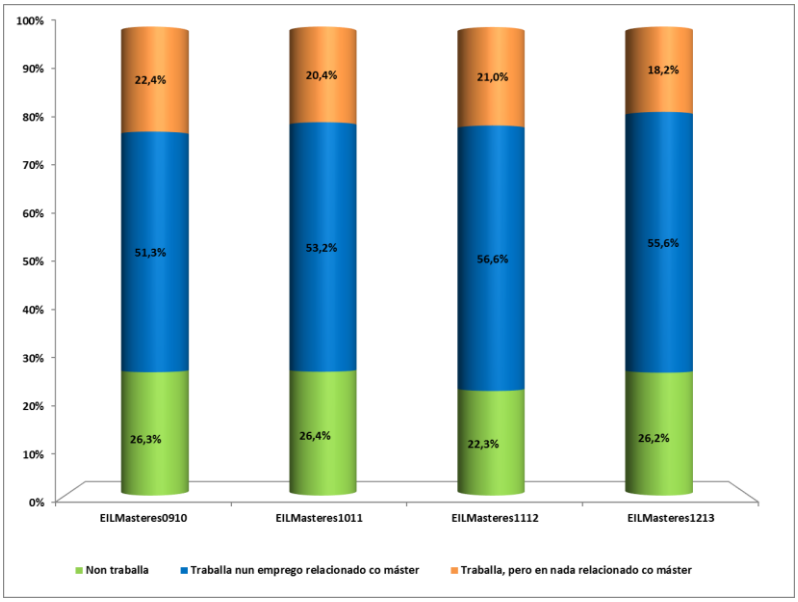


Figura 7: Situación laboral actual dos titulados en máster. Resultados para o total do SUG. Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

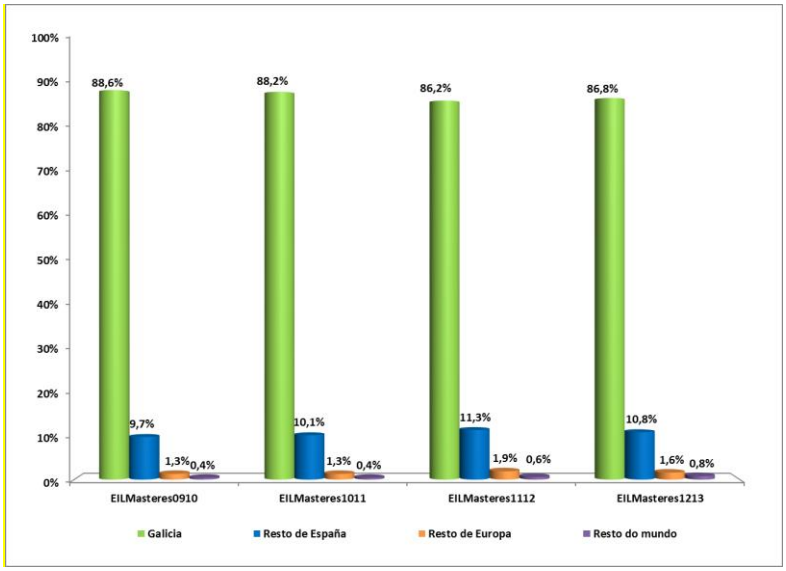


Figura 8: Lugar de traballo actual. Resultados para o total do SUG. Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

Satisfacción co máster

Os titulados en másteres no SUG está satisfeitos cos estudos universitarios cursados. Tres cuartas partes dos titulados en másteres volverían a realizar o máster.

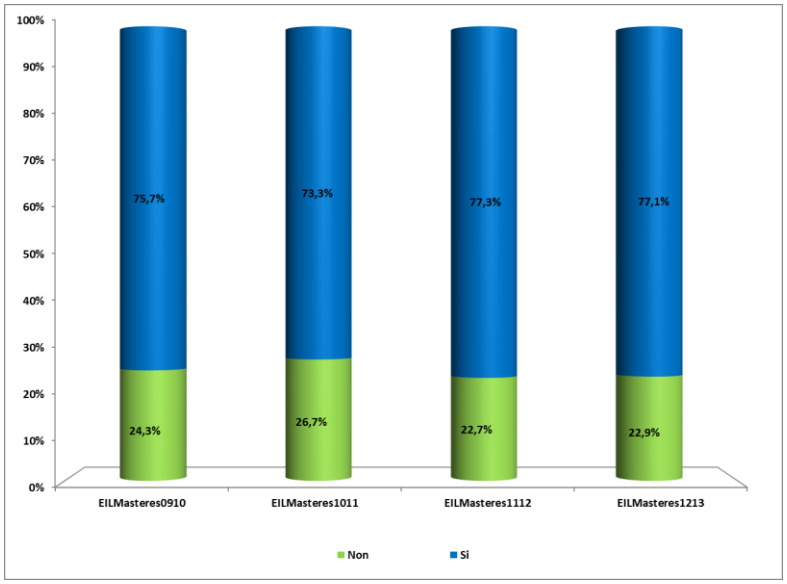


Figura 9: Volverían realizar o máster. Resultados para o total do SUG.
Comparativa dos estudos: EILMasteres0910 ata EILMasteres1213.

REFERENCIAS

De Miguel Domínguez, J. C; Rodríguez González, D.; Valcárcel Aguiar, B.; Vecino Aguirre, A.; Pedreira Andrade, L. P.; Guisado González, M.; Morrondo Pelayo, M. P.; López Pereira, J. E.; Fernández Montes, M. C. (2015) Estudo da inserción laboral dos titulados en Másteres no Sistema Universitario de Galicia 2007-2008, 2008-2009 e 2009-2010.

De Miguel Domínguez, J. C; Rodríguez González, D.; Valcárcel Aguiar, B.; Vecino Aguirre, A.; Pedreira Andrade, L. P.; Guisado González, M.; Morrondo Pelayo, M. P.; López Pereira, J. E.; Fernández Montes, M. C. (2016) Estudo da inserción laboral dos titulados en Másteres no Sistema Universitario de Galicia 2010-2011.

De Miguel Domínguez, J. C; Rodríguez González, D.; Valcárcel Aguiar, B.; Vecino Aguirre, A.; Pedreira Andrade, L. P.; Guisado Tato, M.; Morrondo Pelayo, M. P.; López Pereira, J. E.; Fernández Montes, M. C. (2017) Estudo da inserción laboral dos titulados en Másteres no Sistema Universitario de Galicia 2011-2012.

De Miguel Domínguez, J. C; Rodríguez González, D.; Valcárcel Aguiar, B.; Vecino Aguirre, A.; Pedreira Andrade, L. P.; Guisado Tato, M.; Morrondo Pelayo, M. P.; López Pereira, J. E.; Fernández Montes, M. C. (En imprenta) Estudo da inserción laboral dos titulados en Másteres no Sistema Universitario de Galicia 2012-2013.

XIII Congreso Galego de Estatística e Investigación de Operacines
Ferrol, 26, 27 e 28 de outubro de 2017

ILS: Un paquete en R para realizar estudios Interlaboratorio

Miguel Flores[1], Salvador Naya[2], Javier Tarrío-Saavedra[2], y Rubén Fernández-Casal[2]

[1]Departamento de Matemática, Facultad de Ciencias, Escuela Politécnica Nacional. Quito, Ecuador

[2] Grupo MODES, Departamento de Matemáticas, ITMATI, CITIC, Universidade da Coruña, España.

RESUMEN

En este trabajo se presenta una librería de R para realizar estudios interlaboratorio (ILS). Este paquete permite detectar aquellos laboratorios que proporcionan resultados no consistentes al compararlos con otros. Además permite trabajar simultáneamente con diversos materiales de prueba, desde las perspectivas del análisis de datos univariante y funcional (FDA).

Las técnicas implementadas en el paquete `ILS` permiten realizar Estudios Interlaboratorios con el enfoque univariante, basado en las normas ASTM E691 e ISO 5725, que consiste en estimar los estadísticos h y k de Mandel para identificar aquellos laboratorios que proporcionan resultados significativamente diferentes, así también, permite probar la presencia de valores atípicos mediante las pruebas de Cochran y Grubbs. Además, proporciona técnicas de Análisis de Varianza (ANOVA), incluyendo los test F y de Tukey para probar las diferencias entre las medias de la variable de prueba correspondientes a diferentes laboratorios y a cada material ensayado.

Una de las novedades de esta librería es la de incorporar herramientas para realizar un estudio ILS a partir de datos funcionales. Por tanto, este paquete permite tener en cuenta la naturaleza funcional de los datos obtenidos por técnicas experimentales correspondientes a la química analítica, física aplicada e ingeniería (espectros, termogramas, etc.).

Especificamente, el paquete ILS proporciona herramientas para estimar con un enfoque FDA los estadísticos h y k de Mandel. Estos estadísticos funcionales $H(x)$ y $K(x)$ facilitan la realización de estudios de Repetibilidad y Reproducibilidad (r&R) cuando los datos tienen un carácter funcional.

Palabras y frases claves: Estudio Interlaboratorio, Análisis de Datos Funcional, Detección de atípicos, Bootstrap, Profundidad de datos.

1. INTRODUCCIÓN

Este trabajo se enmarca dentro de las herramientas para el control estadístico de calidad (CEC). Concretamente se abordarán, por un lado, programación de herramientas informáticas en el contexto de Estudios Interlaboratorio y, por otro, elaboración, adaptación, aplicación y puesta en marcha de metodologías de utilidad en la metrología.

El paquete `ILS` permite aplicar técnicas estadísticas de control de la calidad en el marco de Estudios Interlaboratorio que son utilizadas para evaluar la consistencia (homogeneidad) de los resultados experimentales de un laboratorio, obtenidos utilizando un procedimiento experimental bien definido.

Un Estudio Interlaboratorio se puede definir como procedimientos de control para evaluar el desempeño mediante pruebas colaborativas[3,4]. En un Estudio Interlaboratorio se elige un número adecuado L de laboratorios para participar en el experimento, con el objetivo de que éstos analicen las muestras que se les proporcionan y así obtener n resultados. Con estos resultados se evalúan

los métodos analíticos realizados por los laboratorios, ya sea para la verificación de laboratorios, de un procedimiento experimental concreto o una norma, entre otras aplicaciones. Por ejemplo, se puede estudiar las propiedades de un material, un análisis clínico, o cualquier tipo de estudio que requiera el contrastarse entre distintos laboratorios.

Los laboratorios participantes reciben muestras (homogeneizadas o que los laboratorios deberán homogeneizar) para su análisis, para posteriormente estudiar los resultados conjuntos e individuales de dichos análisis y valorar la consistencia de los resultados obtenidos. Algunos de los factores más comunes causantes de variabilidad son el equipamiento de los laboratorios, los operadores, los materiales, la temperatura ó la humedad, entre otros.

Actualmente se utilizan diversas técnicas estadísticas univariantes para estudiar la consistencia de los resultados de las pruebas procedentes de varios laboratorios. La norma ASTM E 691 (normas para estudios de precisión de un método de medición) recomienda aplicar sólo una técnica gráfica a partir de los estadísticos k y h de Mandel [3], mientras que la norma ISO 5725 (práctica estándar para conducir un Estudio Interlaboratorio para determinar la precisión de un método analítico) recomienda además de la técncia gráfica utilizar los test de Cochran y Grubbs [4].

Entre las utilidades que ofrece el paquete `ILS`, están la evaluación de la consistencia de los resultados obtenidos en laboratorio mediante la aplicación de contrastes y herramientas gráficas para la detección de atípicos, además del análisis de la varianza (ANOVA) mediante los contrastes de la F y Tuckey, a través del paquete `multcomp`.

A partir de las propuestas en [2] y [7] de nuevas técnicas en el campo del Análisis de Datos Funcionales (FDA) para realizar Estudios Interlaboratorios, se implementan en el paquete `ILS` funciones para aplicar el método gráfico y analítico pero con un enfoque FDA para los estadísticos h y k de Mandel. Estas funciones permiten estimar los estadísticos funcionales $H(x)$ y $K(x)$, así como estadísticos de contraste que facilitan la realización de estudios de Repetibilidad y Reproducibilidad (r&R) cuando los datos tienen un carácter funcional.

El resto del documento se organiza de la siguiente manera. En la sección 2, se describen 4 conjuntos de datos experimentales que son el resultado de pruebas que generan datos escalares y funcionales (curvas). En la Sección 3, se definen las funciones incluidas en el paquete `ILS` que permiten al usuario aplicar las técnicas desarrolladas. En las secciones 4 y 5 se ilustra la funcionalidad del paquete a través de una serie de aplicaciones utilizando dos conjuntos de datos. Específicamente, en la sección 4 se presenta el uso y resultados para el análisis estadístico en el caso univariante y en la sección 5 para el caso funcional.

2. DATOS EXPERIMENTALES

El paquete `ILS` contiene 4 conjuntos de datos. Para el caso del procedimiento estándar para análisis de resultados univariantes se cuenta con la base de datos **Glucosa** utilizada para explicar el uso de la norma ASTM E 691 [1]. Las bases se han descrito ampliamente en el artículo de Naya et al. [2], estas son IDT para el caso escalar y GT y DSC para el caso funcional.

- **Glucosa:** Contenido de glucosa en la sangre. Conjunto de datos correspondiente a la prueba de glucosa en suero (mediciones de la concentración de glucosa en la sangre utilizada para controlar la diabetes). En el estudio participan ocho laboratorios donde se realizaron cinco pruebas diferentes de muestras de sangre etiquetadas con diferentes referencias, que van desde un bajo contenido de azćar a muy alto. Se obtuvieron tres réplicas para cada muestra.

- **IDT:** Conjunto de datos compuesto por la temperatura de descomposición inicial (IDT) de diferentes muestras de Oxalato de Calcio, obtenidas por 7 diferentes laboratorios. La temperatura de descomposición inicial (IDT) es un parámetro definido por la temperatura a la que se produce una pérdida de material del 5 % de su peso cuando se calienta usando una velocidad constante. 105 muestras de oxalato de calcio se analizaron aplicando técnicas termogravimétricas (TG), obteniendo 105 curvas TG a partir de las cuales se extrae el IDT. En resumen, el conjunto de datos IDT se compone de los valores de IDT del oxalato de calcio obtenido de 7 diferentes laboratorios que analizan 15 muestras de oxalato cada uno.

  El laboratorio 1 utiliza un analizador térmico simultáneo (STA) con un programa de calibración desfasado. El Laboratorio 6 utiliza un analizador SDT simultáneo con una calibración antigua y el Laboratorio 7 utiliza un analizador simultáneo SDT con un sesgo en temperatura con respecto a los valores reales (2°C desplazada con respeto al punto de fusión del zinc).

- **TG:** Curvas termogravimétricas. Se analizaron 105 muestras de oxalato de calcio mediante técnicas termogravim'etricas (TG), obteniéndose 105 curvas TG que muestran la pérdida de masa de oxalato en funciún de la temperatura cuando se calientan las muestras de oxalato a 20°C/min. El conjunto de datos está compuesto por 15 curvas TG compuestar por más de 1000 observaciones de cada uno de los 7 laboratorios.

  El Laboratorio 1 utiliza un analizador térmico simultáneo (STA) con un programa de calibración obsoleto, el Laboratorio 2 al Laboratorio 4 utiliza en condiciones similares el mismo analizador SDT simultáneo con la misma calibración, el laboratorio 6 utiliza un analizador SDT simultáneo con una calibración obsoleta, y el laboratorio 7 utiliza un analizador SDT simultáneo con un calibración sesgada en temperatura (2°C desplazada del punto de fusión de zinc).

- **DSC:** Curvas obtenidas por calorimetría diferencial de barrido. Se analizaron 90 muestras de oxalato de calcio mediante técnica términca calorimétrica diferencial de barrido (DSC), obteniéndose 90 curvas DSC que muestran a partir de un dispositivo SDT la diferencia de energía entre una referencia y la muestra de oxalato. Muestran el intercambio de energía entre muestra y referencia en función de la temperatura. La temperatura varía linealmente con el tiempo con pendiente de 20°C/min. El conjunto de datos está compuesto por 15 curvas TG compuestas por más de 1000 observaciones de cada uno de los 6 laboratorios.

3. ESTRUCTURA DEL PAQUETE

El paquete `ILS` consta de dos grupos de funciones (Tabla 1) para aplicar los enfoques escalar y funcional de detección de laboratorios atípicos en un estudio interlaboratorio. Existen procedimientos gráficos y númericos (contrastes) para este fin.

En el caso escalar, se detectan aquellos laboratorios que proporcionan resultados inconsistentes con respecto a los demás mediante gráficos de los estadísticos h y h de Mandel y a través de las pruebas de Grubbs y Cochran de detección de laboratorios atípicos. En el caso funcional se proporciona una extensión de los gráficos h y k de Mandel y además se proponen pruebas desarrollados por los propios autores para contrastar la hipótesis de repetibilidad y reproducbilidad mediante la propuesta de estadísticos de contraste que se definen como la norma de las curvas de los estadísticos funcionales $H(x)$ y $K(x)$.

Tanto en el procedimiento gráfico como en el análitico, se obtienen los valores críticos de los estadísticos a niveles de significación del 5% y 1%. La norma ISO 5725-2 también proporciona los valores críticos dependiendo del número de laboratorios L y del número de mediciones n. El paquete `ILS` permite calcular los valores críticos para cualquier valor de α, L y n para el caso escalar y funcional.

El paquete `ILS` consta de las clases `lab.qcd` (quality control data) y `lab.fqcd` (functional quality control data) para definir objetos de datos en el caso escalar y, funcional respectivamente, que son necesarios para poder aplicar las funciones presentadas en la Tabla 1. Además de estas funciones el paquete `ILS` consta de las funciones plot, summary, print que permiten gráficar y resumir las estimaciones realizadas.

El objeto `lab.qcd` tiene la estructura de la clase data.frame. En la primera columna debe contener la variable respuesta, mientras que en las siguientes columnas desde la segunda hasta la cuarta se tiene respectivamente, el índice de repetición por cada laboratorio, el índice del material con el que se realizó la prueba y el índice del laboratorio donde se ejecuto el procedimiento.

Mientras que el objeto `lab.fqcd` tiene la estructura de una lista de objetos FDA. El tamaño de la lista es igual al número de laboratorios que participan en el estudio. Se implementaron funciones especificas para realizar gráficos y resúmenes de este tipo de objetos.

Para realizar el cálculo de los estadísticos para los métodos gráficos y analíticos para el caso escalar y funcional se desarrollaron las funciones `lab.qcs` (quality control statistical) y `lab.fqcs` (functional quality control statistical) respectivamente. Estas funciones calculan las medidas estadísticas (media, varianza, etc.) escalares y funcionales necesarias para la estimación de los estadísticos h y k para el caso escalar y $H(t)$ y $K(t)$ para el caso funcional.

4. ESTUDIOS INTERLABORATORIO: ENFOQUE ESTÁNDAR

Enfoque	Técnica	Función	Descripción
Escalar	Gráfica	`h.qcs` `k.qcs`	Estadísticos, h y k de Mandel

	Test	`Cochram.test`
		`Grubbs.test`
		`lab.aov`
			Pruebas de Cochram, Grubbs y ANOVA

Funcional	Gráfica	`h.fqcs` `k.fqcs`	Estadísticos, h y k de Mandel
	Test	`h.norm` `k.norm`	Estadísticos de contraste norma de $H(x)$ y $K(x)$ de Mandel

Tabla 1: Funciones incorporadas en el paquete `ILS` para realizar Estudios Interlaboratorio con un enfoque escalar y funcional.

Como ya se comentó, entre las metodologías utilizadas para evaluar la consistencia de los resultados de los laboratorios, se tienen los estudios r&R, que permiten cuantificar la variabilidad entre laboratorios (reproducibilidad) y la variabilidad entre resultados (reproducibilidad). A continuación se describe la reproducibilidad y repetibilidad desde un contexto de Estudios Interlaboratorios.

La repetibilidad es la variabilidad entre los resultados de las pruebas independientes obtenidas por cada laboratorio individual, aquí se evalúa la variación que se produce por el sistema de medida, ya que las pruebas las obtiene un único operario en cada laboratorio y sin que transcurra demasiado tiempo entre varias mediciones. Por otra parte, la reproducibilidad se refiere a la variabilidad entre los resultados de las pruebas individuales obtenidas en los diferentes laboratorios. Con esto, se puede determinar el sesgo que se produce debido a los operarios o trabajadores del laboratorio en cuestión.

Para realizar estudios r&R mediante métodos gráficos a través del paquete `ILS`, se cuenta con los estadísticos h y k de Mandel [5,6] que se utilizan intensivamente en ILS para detectar los laboratorios que proporcionan resultados inconsistentes. El estadístico h explica la variabilidad intra laboratorios, es decir, estima el sesgo, que es la diferencia de las medias de cada laboratorio con respecto a la media global, mientras que el estadístico k estima la variabilidad intra-laboratorio comparando la repetibilidad correspondientes a cada laboratorio.

La regla decisión para detectar un laboratorio como inconsiste consiste comparar el valor del estadístico h o k sí es mayor que el valor crítico calculado con un nivel de significación generalmente de 0.5%, que es el recomendado por la norma ASTM E-691.

A continuación se definen los estadísticos h y k y posteriormente se explica el uso del paquete `ILS` para calcular los estadísticos h y k de Mandel con sus respectivos valores críticos, así como para realizar una representación gráfica de estos en el caso escalar.

4.1. ESTADÍSTICO h DE MANDEL

Sea $(x_1, x_2, \ldots, x_L)$ una muestra de L observaciones. Los $x_l; l = 1, \ldots, L$ son modelados como realizaciones de variables aleatorias $X_l; l = 1, \ldots, L$ idénticamente distribuidas e independientes distribuidas acorde a una distribución normal $N(\mu, \sigma^2)$. Se denota:

$$\bar{X} = \frac{\sum_{l=1}^{L} X_l}{L} \tag{1}$$

como la media de los X_l,

$$S^2 = \frac{1}{L-1} \sum_{l=1}^{L} (X_l - \bar{X})^2 \tag{2}$$

como la varianza muestral de los X_l. Se define el estadístico h de Mandel [8]

$$H_l = \frac{X_l - \bar{X}}{S}; l = 1, \ldots, L \tag{3}$$

tienen la misma distribución para todo $l = 1, \ldots, L$. El valor crítico [9] es:

$$h_{l;1-\frac{\alpha}{2}} = \frac{(L-1)t_{L-2;1-\frac{\alpha}{2}}}{\sqrt{L(t^2_{L-2;1-\frac{\alpha}{2}} + L - 2)}} \tag{4}$$

Donde $t_{L-2;1-\frac{\alpha}{2}}$ es el cuantil $(1-\frac{\alpha}{2})$ de la distribución t con $v = L-2$ grados de libertad. Para el caso cuando se tienen L laboratorios y n réplicas en cada uno, se define el estadístico h como:

$$h_l = \frac{d_l}{S_{\bar{x}}} = \frac{\bar{x}_l - m}{\sqrt{\sum_{l=1}^{L} \frac{(\bar{x}_l - m)^2}{L-1}}}; l = 1, \ldots, L \tag{5}$$

Donde $\bar{x}_l$ es el promedio de los n resultados de cada laboratorio y m es la media global de todos los resultados de los L laboratorios. Se detecta un laboratorio como inconsiste cuando el valor del estadístico h_l es mayor que el valor crítico, es decir $h_l \geq h_{l;1-\frac{\alpha}{2}}$.

4.2. ESTADÍSTICO k DE MANDEL

Sea $(S_1^2, S_2^2, \ldots, S_L^2)$ una serie de L varianzas muestrales cada una basada en n valores observados. Bajo la hipótesis de que los valores de las observaciones $x_{ji}; j = 1, 2 \ldots, L, i = 1, 2, \ldots, n$ son realizaciones de variables aleatorias X_{ji} las cuales cada j son idénticamente distribuidas e independientes acorde a una distribución normal $N(\mu_i, \sigma^2)$, las varianzas muestrales $S_j^2; j = 1, \ldots, L$ divididas por σ^2 siguen una distribución χ^2/v con $v = n-1$ grados de libertad. El estadístico k de Mandel se define [8]:

$$k_l = \frac{S_l}{\sqrt{\bar{S^2}}}; j = 1, 2, \ldots, L \tag{6}$$

con

$$\bar{S^2} = \frac{\sum_{l=1}^{L} S_l^2}{L} \tag{7}$$

tienen la misma distribución para todo $l = 1, \ldots, L$. El valor cr'itico [9] es:

$$k_{l,n;1-\alpha} = \sqrt{\frac{L}{(1 + \frac{L-1}{F_{v_1,v_2;\alpha}})}} \tag{8}$$

Donde $F_{v1,v2;\alpha}$ es el α-cuantil de la distribución F con $v_1 = (L-1)(n-1)$ y $v_2 = n-1$ grados de libertad.

Para el caso cuando se tienen L laboratorios y n réplicas en cada uno, se define el estadístico k como:

$$k_l = \frac{S_l}{S_r} = \frac{\sqrt{\sum_{l=1}^{L} \frac{(x_l - \bar{x})^2}{n-1}}}{\sqrt{\frac{\sum_{l=1}^{L} S_l}{L}}} \tag{9}$$

Donde S_l es la desviación estándar del número de replicaciones de cada laboratorio estudio. Se detecta un laboratorio como inconsiste cuando el valor del estadístico k_l es mayor que el valor crítico, es decir $k_l \geq k_{l,n;1-\alpha}$.

4.3. ANÁLISIS DE LOS RESULTADOS DEL ESTUDIO DE GLUCOSA

Para explicar el funcionamiento del paquete `ILS` se analiza una base de datos (Glucose) sobre un estudio interlaboratorio que incluye 8 laboratorios. El paquete `ILS` permite llegar a conclusiones en cuanto a la precisión de 8 laboratorios a través de los métodos gráficos implementados.

Cada uno de estos laboratorios mide 5 niveles de concentración diferentes de un determinado material (niveles “A”, “B”, “C’, “D” y “E”) y en cada uno de esos niveles se toman tres medidas (3 réplicas). Cada laboratorio proporciona entonces un total de 15 mediciones (3 para cada nivel) teniendo por tanto, tenemos 120 mediciones.

Para poder trabajar con el paquete `ILS` y con la base de datos Glucose primero cargamos la librería y la base de datos mediante las siguientes sentencias en R:

R> library(ILS)
R> data(Glucose)

Posteriormente, se construye un objeto de la clase `lab.qcd` que se utiliza a través de la función lab.qcs para calcular las medidas estadísticas necesarios para realizar las estimaciónes de los estadísticos h y k de Mandel para cada laboratorio y posteriormente comparar con el valor crítico correspondiente con un nivel de significación α.

R> muestra <− lab.qcd(Glucose)
R> est <− lab.qcs(muestra)

Esta última operación permite estimar simultáneamente tanto la media y desviación global (considerando los resultados de todos los laboratorios), así como la media y desviación para cada laboratorio por tipo de material. Además de la desviación de repetibilidad S_r y reproducibilidad S_R.

Para realizar la estimación del estadístico k que permite estudiar la variabilidad dentro de los laboratorios y así, se utiliza la función k.qcs que necesita como argumento el objeto *est*.

R> k <− k.qcs(est, alpha = 0.005)

El objeto k es de la clase lab.qcs y entre los valores que calcula se encuentran los siguientes:

- k: El estadístico k para cada laboratorio y por material.
- k.critical: El valor crítico para el α definido en la función k.qcs.
- violations: Matriz de dimensión $L \times R$, número de laboratorios por número de materiales.

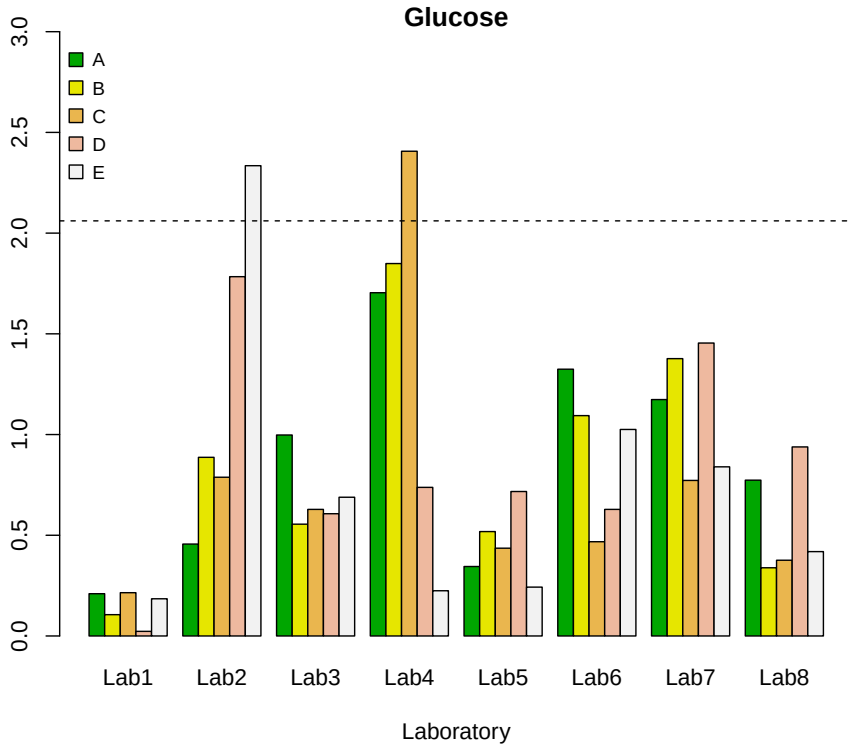


Figura 1: Estadísticos k por laboratorio y material correspondientes a las medidas de glucosa en sangre disponibles en la base de datos Glucosa.

En la matriz "violations" se almacenan valores lógicos resultado de las comparaciones entre los valores críticos y los valores del estadístico k. Es "False" cuando el laboratorio no es consistente, es decir el valor crítico es menor que el valor del estadístico. Para este ejemplo el valor crítico es 2.06. Para realizar el gráfico correspondiente del estadístico k para cada laboratorio y material se utiliza la función `plot(k)`. En la Figura 1, se puede apreciar como los materiales 5 y 3 de los laboratorios 2 y 4 respectivamente presentan resultados inconsistentes dado que los correspondientes valores de los estadísticos k son mayores que el valor crítico calculado a partir de $L = 8, n = 15$ y $\alpha = 0.005$.

De igual forma se procede a utilizar las funciones `h.qcs` y `plot(h)` para estimar y gráficar respectivamente los estadśiticos h para cada laboratorio y material.

R> h <- h.qcs(est, alpha = 0.005)
R> plot(h)

Para este caso el valor crítico es 2.15, a partir de este resultado se observa en la Figura 2, que a un nivel de significancia $\alpha = 0.005$ todos los resultados por laboratorio y material son consistentes.

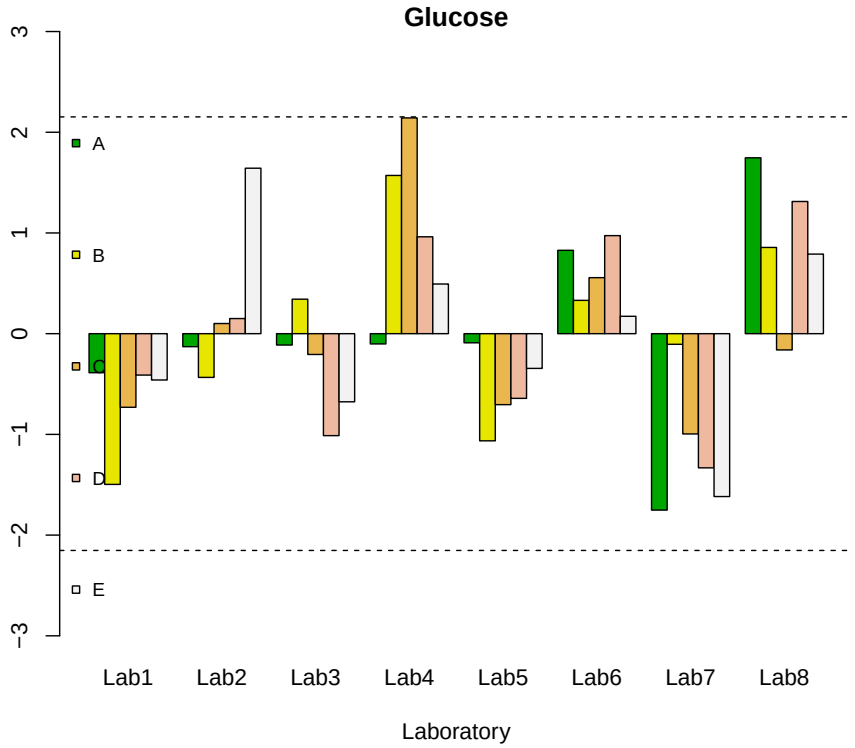


Figura 2: Estadísticos h por laboratorio y material.

5. ESTUDIOS INTERLABORATORIO: NUEVO ENFOQUE FUNCIONAL

En esta sección, se presenta el uso del paquete `ILS` para realizar las estimaciones y representación gráfica de los estadísticos funcionales, $H(t)$ y $K(t)$, y mediante estos realizar un estudio r&R para uno de los conjuntos de datos funcionales que mantiene el paquete `ILS`, específicamente el conjunto de curvas TG. En este trabajo, sólo se aplicará el método gráfico funcional, sin embargo el paquete `ILS` incorpora las funciones para aplicar otros métodos que se pueden encontrar en Flores et al. [7]. En [7] se tiene una descripción de los métodos analíticos utilizados para contrastar las hipótesis de repetibilidad y reproducibilidad a través de los correspondientes estadísticos de contraste d_H y d_K. Además se presenta el algoritmo bootstrap implementado para estimar los valores críticos correspondientes a los test de reproducibilidad y repetibilidad c_h y c_k respectivamente.

En la siguiente sección se definen las hipótesis de reproducibilidad y repetibilidad que posteriormente son contrastadas mediante los gráficos fucionales de los estadísticos $H(t)$ y $K(t)$.

5.1. HIPÓTESIS DE REPRODUCIBILIDAD Y REPETIBILIDAD

En el ILS, un conjunto de observaciones $\left\{X_1^l(t), \ldots, X_n^l(t)\}\right\}$ se obtienen para cada $l, l = 1, \ldots, p$. Cada laboratorio realiza n muestras experimentalmente, obteniendo n curvas diferentes. Los estadísticos funcionales $H_l(t)$ y $K_l(t)$ se calculan para cada laboratorio y considerando la hipótesis nula de que no hay datos estadísticamente diferentes entre las mediciones de laboratorio. Las hipótesis nulas para r&R se describen a continuación.

La hipótesis nula de reproducibilidad afirma que:

$$H_0 : \mu_1(t) = \mu_2(t) = \ldots = \mu_p(t), \tag{10}$$

Donde $\mu_l(t)$, $l = 1 \ldots p$ es la media funcional poblacional para cada laboratorio l. Para probar la reproducibilidad de los resultados de los laboratorios, el cálculo del estadístico $H(t)$ es necesario. Este se define como:

$$H_l(t) = \frac{X_i^l(t) - \bar{X}(t)}{S_l(t)}; l = 1, \ldots, p,$$

Donde $\bar{X}_l(t)$ y $S_l(t)$ son la media y la varianza funcional punto a punto calculada para el laboratorio l.

La hipotésis nula de repetibilidad afirma que:

$$H_0 : \sigma_1^2(t) = \sigma_2^2(t) = \ldots = \sigma_p^2(t), \tag{11}$$

Donde $\sigma_l(t)$, $l = 1 \ldots p$ son las varianzas funcionales teóricas correspondientes para cada laboratorio l.

La prueba de repetibilidad está basada en el estadístico ($K(t)$, definido como:

$$K_l(t) = \frac{S_l(t)}{\sqrt{\bar{S^2}(t)}}; l = 1, \ldots, p,$$

Donde, $\bar{S^2}(t) = \frac{1}{p}\sum_{l=1}^{p} S_l^2(t)$.

Por otro lado, para probar la hipótesis de reproducibilidad, se define el estadístico de prueba d_H como:

$$d_l^H = \|H_l(t)\| = \left(\int_a^b H_l(t)^2 dt\right)^{\frac{1}{2}},$$

Teniendo en cuenta que los valores m'as grandes de d_K corresponden a laboratorios no consistentes. Así también, para la hipotésis de repetibilidad, se define $d_l^K = \left\|K_(t)\right\|$, y de igual forma los valores grandes de d^K corresponden a laboratorios no consistentes.

5.3. ANÁLISIS DE LOS RESULTADOS DEL ESTUDIO TERMOGRAVIMÉTRICO (TG)

En esta subsección nos limitamos a utilizar el procedimiento funcional para estudios de r&R, descrito anteriormente, en el análisis termogravimétrico con muestras oxalato de calcio. Como se mencionó en la Sección 3, el conjunto de datos TG, contiene resultados de pruebas termogravimétricas de 7 laboratorios, de los cuales hay 4 laboratorios que proporcionan resultados similares, mientras que los laboratorios 1, 6 y 7 proporciona diferentes resultados, es decir son inconsistentes.

El primer paso consiste en estimar los estadísticos funcionales $H(t)$ y $K(t)$, posteriormente son graficados y comparados con respecto a sus correspondientes valores críticos c_H y c_K (límites de control), definidos como los cuantiles correspondientes a $\alpha = 0.01$ (como es habitual en los estudios ILS), y obtenidos por el procedimiento bootstrap descrtio en [7]. Para esto, cargamos la librería ILS, así como el conjunto de datos TG que es un arreglo de dimensión $15 \times 1000 \times 7$. Es decir, contiene 7 matrices de dimensión 15x1000, donde 15 es el número de replicas por cada laboratorio

y 1000 los diferentes grados en las que fueron evaluada la perdida de masa del material durante las pruebas.

```
R> library(ILS)
R> data(TG)
```

De igual forma que en el caso escalar se contruye un propio objeto (lab.fqcd) del paquete que permite realizar la estimación de los estadísticos funcionales necesarios para estimar los estadísticos funcionales $H(t)$ y $K(t)$.

```
R> delta <- seq(from = 40 ,to = 850 ,length.out = 1000 )
R> list.fdata <- lab.fqcd(TG, argvals = delta)
```

A partir de este objeto se puede realizar un gráfico (ver Figura 3) con las siguientes instrucciones en R.

```
R> xlab <- "Temperatura (Grados)"
R> ylab <- "Masa (%)"
R> main <- "Curvas TG obtenidas del oxolato de calcio"
R> p <- dim(curves.fqcd$curves)[3]
R> col <- terrain.colors(p)
R> plot(x = curves.fqcd, main, xlab, ylab,legend = FALSE,col = col)
R> legend(45,70,c(paste("Lab",c(1:7))), col = col,lty = 1, lwd = c(rep(1,7),2), cex = 0.7)
```

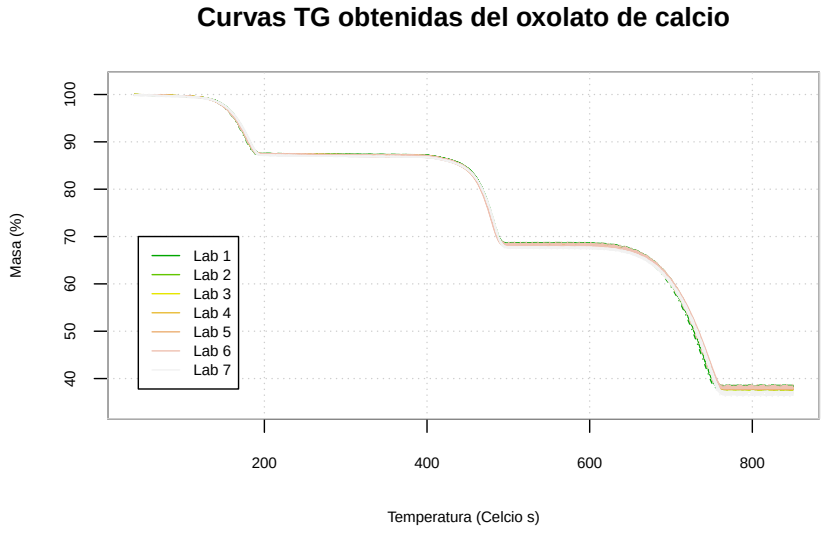


Figura 3: Curvas TG obtenidas del oxolato de calcio.

A partir del objeto de la clase lab.fqcd se construye otro objeto de la clase lab.fqcs. En la Figura 4, se presentan las diferentes medias y varianzas funcionales para cada laboratorio así como la media y la varianza global.

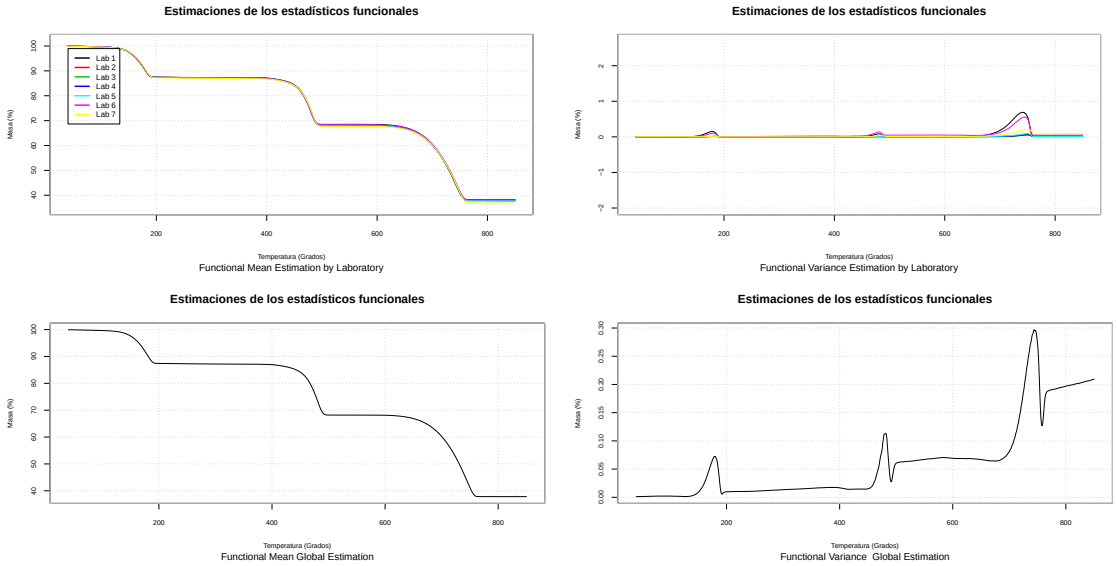


Figura 4: Estadística descriptiva funcional: medias y varianzas funcionales para cada laboratorio así como la media y la varianza global por laboratorio de las curvas TG obtenidas del oxolato de calcio.

R> est <- lab.qcs(muestra)

R> xlab <- "Temperatura (Grados)"

R> ylab <- "Masa (%)"

R> main <- "Estimaciones de los estadísticos funcionales"

R> p <- dim(curves.fqcd$curves)[3]

R> col <- 1:p

R> plot(curves.fqcs,main = main, xlab = xlab, ylab = ylab, col = col)

Finalmente para realizar la estimación de los estadísticos funcionales se ejecuta los siguientes comandos en R.

R> k.fdata <- k.fqcs(est, delta, alpha = 0.01)

R> h.fdata <- h.fqcs(est, delta, alpha = 0.01)

La hipótesis de reproducibilidad se prueba usando directamente el estadístico funcional H y la de repetibilidad el estadístico funcional K. Por ejemplo, en la Figura 5 se muestran las curvas de los estadísticos K para cada laboratorio con $\alpha = 0.01$. Los laboratorios 1, 6 y 7 son detectados como laboratorios at'ipicos en la primera iteración de la metodología. La región correspondiente a la primera, segunda y tercera etapas de degradación del oxalato de calcio está fuera de las bandas de confianza de 99%.

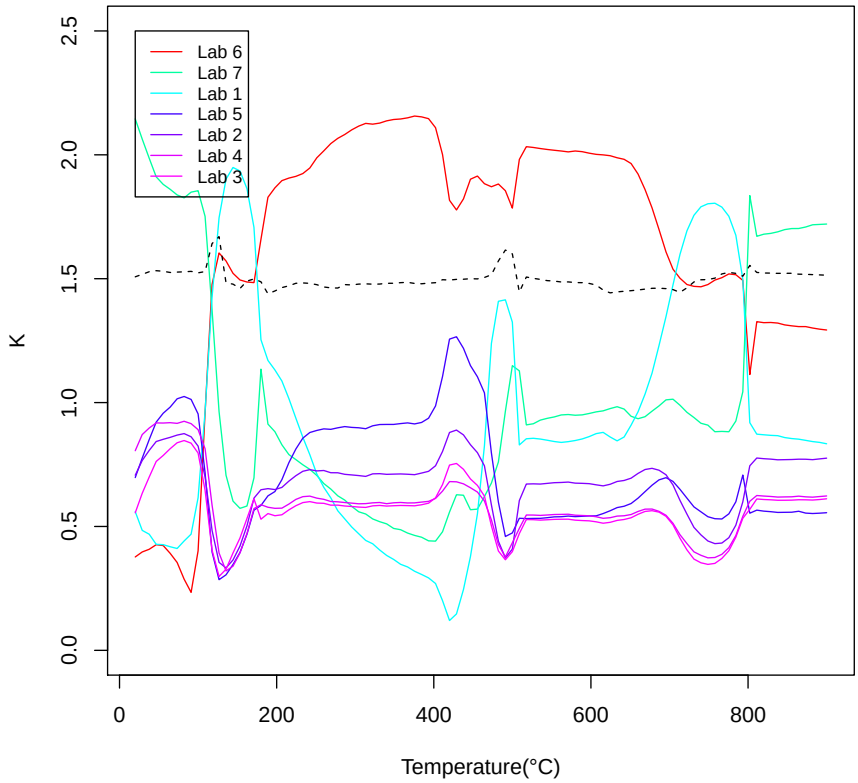


Figura 5: Estadísticos $K(x)$ para cada laboratorio.

REFERENCIAS

1. ASTM E 691. Standard practice for conducting an interlaboratory study to determine the precision of a test method. American Society for Testing and Materials. West Conshohocken, PA, USA, (1999).

2. Naya, S., Tarrio-Saavedra. J., Lopez- Beceiro, J., Francisco Fernandez, M., Flores, M. and Artiaga, R., "Statistical functional approach for interlaboratory studies with thermal data". Journal of Thermal Analysis and Calorimetry, 118,1229-1243, (2014).

3. Practice for conducting and interlaboratory study to determine the precision of a test method. Annual Book of ASTM Standards. West Conshohocken, PA: ASTM International E691; 2004.

4. ISO. International Standard ISO 5725-4-1994. Accuracy (Trueness and precision) of measurement methods and results Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. Geneva, 1994.

5. Hund E, Massart DL, Smeyers-Verbeke J. Interlaboratory studies in analytical chemistry. Ann Chim Acta. 2000;423:14565.

6. Lysiak-Pasttuszak E. Interlaboratory analytical performance studies: a way to estimate measurement uncertainty. Oceanologia. 2004;46(3):42738.

7. Flores M. , Naya S., Tarrío-Saavedra J., Fernández-Casal R., Functional data analysis approach of Mandel's h and k statistics in Interlaboratory Studies, 2017, pp. 123130

8. Mandel, J., A new analysis of interlaboratory test results. In: ASQC Quality Congress Transaction-Baltimore, pp. 360366 (1985)

9. Wilrich P., Critical values of Mandel's h and k, the Grubbs and the Cochran test statistic,AStA Adv Stat Anal, 97:110, (2013).

10. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria, (2008). http://www.R-project.org.

11. Ramsay J.O., Silverman B.W.: Applied Functional Data Analysis: Methods and Case studies. Springer-Verlag, New York (2002)

12. Ramsay J.O., Silverman B.W.: Functional Data Analysis. 2nd edition. Springer-Verlag, New York (2005)

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

ANÁLISIS ESTADÍSTICO DE LOS INCENDIOS DE ALBERTA

Laura Freijeiro Fernández[1] y María José Ginzo Villamayor[2]

[1]Instituto Tecnológico de Matemática Industrial (ITMATI), España, laura.freijeiro@rai.usc.es
[2]Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela, Grupo de Investigación Modestya, Instituto Tecnológico de Matemática Industrial (ITMATI), Santiago de Compostela, España, mariajose.ginzo@usc.es

RESUMEN

Uno de los principales problemas medioambientales del siglo XXI son los incendios. Estos contribuyen al calentamiento global, influyen en la contaminación del aire, aumentan la desertificación y fomentan la pérdida de biodiversidad, con lo que se hace latente la necesidad de implantar políticas eficientes de prevención y actuación para enfrentarse a este conflicto.

Para ilustrar este problema se han seleccionado datos de incendios de Alberta, provincia de Canadá, http://wildfire.alberta.ca/resources/default.aspx. Esta región es propensa a sufrir, anualmente, un considerable número de incendios de gran envergadura, denominados "Wildfires", siendo de gran importancia la disminución de estas catástrofes. Con un fin introductorio acerca del análisis y tratamiento de bases de datos de incendios se analizan y tratan estos a través de diversas técnicas estadísticas. Los datos de los que se dispone han sido registrados entre los años 1996 y 2014 y cuentan con un gran número de variables que caracterizan a cada incendio. En primer lugar se analizan las distintas variables con las que cuenta la base de datos, con el objetivo de entender su comportamiento, determinando sus posibles causas, establecer qué factores podrían ser relevantes y detectar aquellas variables a tener en cuenta si se busca prevenir incendios en el futuro. Además, se analizan las posibles relaciones dos a dos entre distintas variables con la finalidad de ver si existe algún tipo de dependencia entre ellas que permita explicar mejor estos fenómenos.

En ocasiones se pueden observar determinados comportamientos espaciales en la distribución de los incendios, como zonas más propensas donde se han registrado un número mayor de fuegos, debidos al clima, vegetación u otro factor referente a su causa o en contraste, zonas menos afectadas debido a factores similares.

Con la finalidad de conocer si existe algún patrón espacial en la disposición de los incendios sobre el mapa de Alberta, así como de otras variables relacionadas que pueden presentar esta característica, se realiza un análisis espacial de tipo exploratorio, con el objetivo de testar, gráficamente, si es posible afirmar que existe algún tipo de patrón respecto a la distribución de sus valores en determinadas localizaciones.

Una de las principales características de las épocas de incendios es que tienden a estar ligadas a climas favorables, registrándose un mayor número de estos en estaciones más cálidas y secas. A pesar de este hecho, pueden existir otras causas ajenas a la climatología que sean las culpables de que existan determinados períodos donde se observa un crecimiento del número de fuegos. A continuación, para conocer la posible existencia de patrones temporales tanto en los incendios de la provincia de Alberta, como en el número de hectáreas quemadas a lo largo del período considerado, se aplican técnicas de series de tiempo, construyendo un modelo que permita explicar el comportamiento de los datos a lo largo de los años. Se estudia también si existe algún tipo de estacionariedad que determine la existencia de temporadas de menor o mayor número de fuegos a lo largo de cada año.

Palabras y frases clave: Alberta, incendios, dependencia, ANOVA, estadística espacial y series de tiempo.

1. INTRODUCIÓN

En este trabajo se pretende mostrar una primera toma de contacto acerca de cómo se puede abordar un problema como es la actuación y prevención de incendios utilizando diversas técnicas estadísticas.

Con esta finalidad se toma una base de datos real de incendios, los ocurridos en Alberta (Canadá) y se analiza la información que esta proporciona. Dado que estos conjuntos de datos tienden a contar con un gran número de variables es importante realizar un análisis exploratorio exhaustivo que permita conocer las principales características de los incendios con los que se va a trabajar. Este procedimiento permitirá entender como manipular las variables de forma correcta y proporcionará información que permita determinar cuales hay que tener en cuenta según un fin determinado.

Una vez que la base de datos está perfectamente caracterizada es posible utilizar técnicas estadísticas de forma eficiente utilizando la información extraída en este primer paso.

1.1. BASE DE DATOS

Como ya se ha comentado anteriormente se trabajará con la base de datos de los incendios registrados en la provincia de Alberta (Canadá), proporcionada por el Gobierno de Alberta.

Esta base de datos contiene información de 25.311 incendios que tuvieron lugar en Alberta entre los años 1996 y 2014. Está formada por 37 variables, entra las que destacan: la fecha de detección del incendio (*discovered_date*), las hectáreas afectadas (*assessment_hectares*), el tipo de incendio (*size_class*), la zona afectada (*fire_origin*), el grupo causante (*general_cause_desc*), el tipo de empresa implicada (*industry_identifier_desc*), la actividad que se estaba desarrollando cuando se produjo el fuego (*activity_class*) o la causa exacta de este (*true_cause*), entre otras. Se estudiarán los datos de forma conjunta y separando por períodos de dos años, salvo el último, de la forma siguiente: 1996-1997, 1998-1999, ..., 2012-2013 y 2014. Las Tablas 1 y 2 recogen a modo de resumen el total de incendios producidos y de hectáreas quemadas por año.

Año	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*
Número de fuegos	376	456	1698	1355	783	989	1447	1188	1612	1448
Área quemada (ha)	780.76	2094.52	5104.08	6971.74	3625.97	1377.09	3156.6	2332.09	8180.93	3235.36

Tabla 1: Número total de incendios y hectáreas quemadas por año durante el período 1996-2005.

Año	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*	*2012*	*2013*	*2014*
Número de fuegos	1954	1349	1712	1710	1840	1218	1568	1193	1414
Área quemada (ha)	2305.53	1239.75	1127.13	839.38	3962.66	4843.33	18396.95	576.47	2161.25

Tabla 2: Número total de incendios y hectáreas quemadas por año durante el período 2006-2014.

Para poder trabajar estadísticamente con los datos, previamente, fue necesario realizar una depuración de los mismos, ya que se ha encontrado que en una determinada característica el mismo nivel se encontraba codificado de varias formas. Además, en ciertas variables no se dispone de todos los datos, bien porque no son observables o debido a que no ha sido posible obtener esa información, de forma que cuando esto ocurra se optará por trabajar eliminando aquellas muestras de datos faltantes en vez de usar técnicas de **imputación** para reconstruir el dato completo. Por último, decir que pese a disponer de datos temporales estos vienen recogidos en distintos formatos siendo necesario **codificar** todos en uno común.

1.2. ANÁLISIS EXPLORATORIO

En primer lugar, se realiza un análisis exploratorio de los datos, estudiando cada una de las variables cualitativas y cuantitativas por separado, con el fin de extraer conclusiones acerca de la importancia de estas con relación al número de incendios. A la hora de realizar el estudio de los datos se consideran dos escenarios, el primero teniendo en cuenta el total y el segundo separando por períodos de dos años de la forma 1996-1997, 1998-1999, ..., 2012-2013 y 2014 (los datos del año 2014 se estudian de forma individual).

Posteriormente, se analiza la relación entre varios pares de variables cualitativas escogidas bajo el criterio lógico de que estas puedan presentar relación entre ellas.

1.3. ESTUDIO DE DEPENDENCIA DE LAS VARIABLES CUALITATIVAS

Con la finalidad de esclarecer la posible dependencia entre las variables cualitativas que forman la base de datos se va a estudiar la existencia de relación entre estas dos a dos, seleccionando únicamente aquellas de las que se sospecha que puedan presentar algún tipo de dependencia.

En primer lugar se ha realizado el estudio teniendo en cuenta los datos de todos los años considerados y seguidamente, separando por los períodos de tiempo citados anteriormente, se estudia si los datos presentan resultados diferentes al ser considerados de esta forma. Se he empleado el test **ji-cuadrado** para determinar si existe dependencia entre las variables y en caso afirmativo se han obtenido los **coeficientes de contingencia** y el **V de Cramer**, respectivamente. En muchas ocasiones ha sido necesario construir nuevos grupos con las distintas clases de una variable para poder garantizar que se cumplían las condiciones necesarias para emplear este test de forma óptima, por ello se ha agrupado manteniendo un orden lógico dentro de cada nueva clase.

Se presentarán aquellos resultados que se intuyen más relevantes, ya que en algunos casos la necesidad de agrupar las clases en un número de grupos reducidos, para estar bajo las hipótesis óptimas del test ji-cuadrado, ha hecho que se pierda interpretabilidad de los resultados, ya que los grupos serían muy heterogéneos y por tanto carece de interés estudiar la dependencia entre dos variables con estas características. Además, los estudios bianuales considerados no parece que arrojen información distinta a los realizados considerando todos los datos e incluso, muchas veces no es posible aplicar el test ji-cuadrado pues se violan las suposiciones necesarias.

Se empieza considerando todos los datos de incendios de Alberta recopilados desde el año 1996 hasta el 2014 para el estudio de dependencia bivariante. Se han realizado diversas combinaciones de las variables cualitativas construyendo sus respectivas tablas de contingencia y viendo si estas se encuentran bajo las hipótesis del test ji-cuadrado. En diversos casos se ha obtenido que incluso realizando grupos más genéricos con las clases de cierta variable u obviando niveles de esta no se consigue verificar estas hipótesis, con lo que no tiene sentido implementar este test de dependencia y por tanto se obviará el estudio y exposición de estos.

Los estudios de dependencia realizados fueron los siguientes:

- estudio entre la variable que recoge las condiciones atmosféricas ("$weather_condition_over_fire$") y la clase de incendio ("$size_class$");
- estudio si existe relación entre el tipo de industria ("$industry_identifier_desc$") y la actividad causante del incendio ("$activity_class$");
- se estudia si depende el tipo de zona afectada según la empresa causante de un incendio, para esto se estudia la posible existencia de relación entre las variables "$industry_identifier_desc$" y "$fire_origin$";
- relación entre el tipo de avance que tiene un fuego ("$fire_position_on_slope$") y el tipo de incendio ("$size_class$");
- existencia de relación entre la posición de un fuego "$fire_position_on_slope$" y las condiciones atmosféricas en las cuales se produce "$weather_conditions_over_fire$";
- se estudia la relación de las patrullas de actuación con el crecimiento del fuego;
- se estudia si tiene relevancia el crecimiento del fuego con el grupo de actuación;
- se indagan posibles relaciones entre el lugar de origen del fuego, $fire_origin$, con otras variables que podrían explicar este;
- se observa una relación notable entre la causa general, "$general_cause_desc$", y el lugar de origen del fuego, "$fire_origin$" y
- se estudia qué actividad específica se estaba desarrollando cuando se produjo tal fuego en las distintas zonas, estudiando la relación entre "$true_cause$" y "$fire_origin$".

2. TÉCNICAS ESTADÍSTICAS

Una vez conocidos los datos interesa estudiar las posibles estructuras de dependencia existentes entre ellos con el fin de poder entender o conseguir explicar cuáles son las causas que provocan los incendios, predecir cómo serán en el futuro según las características que presenten o si existe alguna forma de prevenirlos entre otros.

Para conseguir llevar a cabo este objetivo se realizan los siguientes análisis:

- Estudio de dependencia entre dos variables continuas que recogen los datos acerca del perímetro de incendio la primera y del número de hectáreas quemadas la segunda, para ver si se puede afirmar, como parece lógico, que el número de hectáreas destruidas es proporcional al tamaño del incendio y en qué medida.
- Se buscará un modelo de regresión que consiga explicar el número de hectáreas quemadas a través de un número suficientemente aceptable de variables, es decir, que consiga explicar la mayor parte de la variabilidad contando con un número de variables no demasiado grande para que el modelo sea fácilmente interpretable.
- A continuación, con el objetivo de analizar la posibilidad de una dependencia en la localización se hará un estudio espacial de aquellas variables que cabe esperar que puedan presentar tal relación, implementando, entre otras, técnicas de **procesos puntuales**.
- Finalmente, para conseguir determinar si se puede afirmar la existencia de temporadas de incendios así como cuando se producirían, se realiza un análisis de **series de tiempo** sobre el número de incendios y el número de hectáreas destruidas a lo largo de estos años.

2.1. ESTUDIO DE LA RELACIÓN ENTRE EL PERÍMETRO Y LAS HECTÁREAS QUEMADAS

Se realiza un **análisis de dependencia** entre el perímetro de la zona afectada por el incendio y el número de hectáreas quemadas. Es de esperar que ambas variables presenten dependencia, ya que parece obvio la existencia de esta entre el perímetro y las unidades de área afectada. Interesa conocer cómo se puede modelar dicha relación entre ambas variables por lo que se apela a los **modelos de regresión** para este fin.

De nuevo se realizará el estudio teniendo en cuenta el total de datos y seguidamente separando estos por etapas. Además, se estudiará el grado de dependencia entre ambas variables y se intentará, en cada caso, modelizar dicha dependencia a través de un modelo de regresión cuando sea posible.

2.2. MODELOS DE REGRESIÓN PARA LA PREDICCIÓN DEL NÚMERO DE HECTÁREAS QUEMADAS

Una de las mayores preocupaciones a la hora de combatir incendios es conseguir determinar, o tener una estimación, de cuál puede llegar a ser el número de hectáreas afectadas según el valor o la clase que tomen sus variables inicialmente.

Pese a que en un incendio influyen muchas variables no controlables (cambio de la meteorología, distinta actuación de equipos de emergencia, desconocimiento de la causa del incendio, ...) se intentará, a través de un modelo de regresión, modelizar el número de hectáreas quemadas teniéndose en cuenta sólo algunas de las variables de las cuales se conoce su valor tras finalizar el fuego.

Con este fin se busca construir un **modelo de regresión lineal general** en el que intervengan varias variables adecuadamente seleccionadas. Dado que se cuenta con un gran número de posibilidades para introducir en el modelo y además parece lógico que existan interacciones entre ellas, se hace un estudio preliminar acerca de cómo explicarían ciertas variables el número de hectáreas, a través de un **modelo lineal simple**, con el objetivo de determinar las más importantes.

Dentro de los factores que se han determinado como más importantes a la hora de explicar las hectáreas que involucrará un incendio de forma individual han sido:

- la variable tipo de incendio ("$size_class$"), la cual clasifica los fuegos en los niveles *"A"*, *"B"*, *"C"*, *"D"* o *"E"* según su tamaño;
- el avance del incendio,"$fire_position_on_slope$", con los niveles "$Bottom$", "$Flat$", "$Lower1/3$", "$Middle1/3$" y "$Upper1/3$" y

- la zona de origen del fuego ("*fire_origin*") separando en 9 clases: "*B.C*"[1], "*DND*"[1], "*Indian_Reservation*"[1], "*Metis_Settlement*"[1], "*National_Park*"[1], "*Private_Land*"[1], "*Provincial_Land*"[1], "*Provincial_Park*"[1] y "*Saskatchewan*"[1].

Una vez que se han estudiado los modelos lineales simples se busca mejorar el ajuste considerando conjuntamente sus variables en un modelo de regresión lineal múltiple, ya que parece adecuado que presenten relaciones entre sí.

Tras estudiar estos modelos se plantea si se está obviando alguna variable que podría ser de interés a la hora de explicar el número de hectáreas quemadas, una vez que se tienen en cuenta las características citadas. Para ello se considerada un modelo lineal con una mayor cantidad de variables.

Aún así no se consigue que el modelo explique de forma óptima el número de hectáreas quemadas. Con la finalidad de mejorar este se estudia introducir interacciones entre las variables, seleccionando aquellas combinaciones que podrían ser lógicas dado el carácter de estas. Añadiendo interacciones al modelo parece que mejora el resultado de la estimación, pero aun así no es un buen resultado, de forma que se hace necesario recurrir a modelos más complejos donde entren en juego más variables o interacciones. De esta forma se consigue un modelo (1), con variables regresoras "*size_class*", "*fire_position_on_slope*" e "*industry_identifier_desc*" que es capaz de explicar un 71% de la variabilidad del número de hectáreas quemadas, lo que permitiría, una vez conocido el valor de las variables explicativas, determinar que extensión se puede esperar que destruya el fuego. Este modelo tiene la forma:

$$y_{ijk} = \mu_{ijk} + \alpha_i + \alpha_j + \alpha_k + \gamma_{ij} + u_{ijk} \tag{1}$$

donde α_i, α_j y α_k son el incremento de la media al pasar de la clase de incendios *"A"*, con crecimiento de tipo *"Bottom"* e indicador de industria *"Comm. Tourism"*, a las restantes de las variables "*size_class*", "*fire_position_on_slope*" y "*industry_identifier_desc*" respectivamente. Finalmente γ_{ij} recoge la interacción entre "*size_class*" y "*fire_position_on_slope*" y u_{ijk} es el error del modelo. Destacar que cuando se está recorriendo los niveles de cierta variable, con los de las restantes fijados, por ejemplo en el caso de "*size_class*" lo hace el subíndice j, las variables con otros índices asociados toman un valor fijo (en este caso α_i, α_k).

El mejor modelo de regresión que se ha conseguido para explicar el número de hectáreas quemadas es muy complejo ya que incluye tres interacciones: "*size_class*" con "*fire_position_on_slope*", "*fire_type*" con "*fire_position_on_slope*" y "*activity_class*" con "*responsible_group_desc*", además de tener también en cuenta la variable "*industry_identifier_desc*". De esta forma se tiene un modelo de difícil interpretación pero a cambio se consigue explicar cerca del 80% de la variabilidad (R^2-ajustado= 0.7925).

2.3. ESTUDIO ESPACIAL DE DISTINTAS VARIABLES

Los datos recogen la posición espacial en términos de longitud y latitud de la zona determinada de Alberta donde se ha producido cada uno de los incendios registrados. Esto permite enfocar la información de la que se dispone dentro del contexto de la estadística espacial.

Surge así el interés de conocer si existe de algún modo dependencia espacial para ciertas variables, como podría ser identificar si hay zonas propensas a incendios, si dependiendo del tipo de área el número de hectáreas quemadas es mayor o si en cambio, el tamaño de un incendio puede estar relacionado con la clase de flora que hay en el lugar.

Con este fin se estudia la posible existencia de patrones espaciales para las variables que recogen la información mencionada. De esta forma se representan los incendios ocurridos, atendiendo a dichas variables, sobre el plano de la provincia de Alberta. De nuevo se tendrá en cuenta el conjunto de datos totales para ver si hay zonas más afectadas o más propensas que otras en global y a continuación, se realizan los estudios por años por si pudiese existir algún patrón a lo largo del tiempo.

[1]Las zonas donde se ha recogido incendios se clasifican según el grupo propietario del terreno B.C.: Columbia Británica-Canadá, DND: National Defence District (distrito de defensa nacional), reservas indias, zonas pertenecientes a Metis Settlement, parques nacionales, terreno privado, terrenos o parques provinciales y zona de Saskatchewan.

Así se estudiará que ocurre con el número de incendios, el número de hectáreas quemadas, *"assessment_hectares"*, la clase de los incendios (*"size_class"*) y el tipo de flora de cada zona donde se ha producido un incendio (*"fuel_type"*).

Finalmente se analizan los resultados gráficos obtenidos y se compara la localización de las distintas variables, para ver si es posible determinar alguna relación a la hora de explicar la localización de los incendios.

Se llevaron a cabo los siguientes análisis espaciales:

- número de incendios,
- número de hectáreas afectadas,
- tamaño inicial del incendio y
- clase de flora quemada.

2.4. PROCESOS PUNTUALES (ANÁLISIS EXPLORATORIO)

Los datos espaciales se clasifican principalmente en tres tipos dependiendo de su estructura, pueden ser de **tipo geoestadístico**, **datos de áreas o regionales** o finalmente **patrones puntuales** (véase Giraldo-Henao (2011)). En el caso de la geoestadística se dirá que los datos pertenecen a esta disciplina cuando se puedan obtener mediciones de la variable de estudio en todas las localizaciones deseadas de la región, un ejemplo es la medición de la temperatura en una placa de metal. En cambio, un proceso espacial con datos de áreas o regionales estará constituido por un único representante de la variable para cada una de las áreas consideradas que formarán la región de estudio, un ejemplo de estos datos puede ser el número de habitantes de cada comunidad autónoma de España medido el año anterior. Finalmente un proceso puntual es aquel en el que la decisión al respecto de donde se hace la medición no depende del investigador, siendo la ubicación de cada medida del fenómeno a estudiar dada por naturaleza o de antemano. Los datos de incendios son un ejemplo de proceso puntual.

La importancia de los procesos puntuales radica en conocer si existe algún patrón en la disposición de los datos, si se puede considerar que estos se distribuyen de forma aleatoria en el espacio o si en cambio siguen una distribución regular o agregada. Con este fin se estudia si el promedio de eventos por unidad de área $\lambda(x)$, denominado **intensidad**, es homogéneo a lo largo de la región. El primer requisito para poder determinar que los datos tienen algún patrón observado es rechaza la hipótesis de *aleatoriedad espacial completa CSR* (ver Fuentes-Santos (2016) para más información), es decir, rechazar que la intensidad $\lambda(x)$ es homogénea a lo largo de la región. Los procesos que cumplen la propiedad CSR se denominan **procesos Homogéneos de Poisson**.

En Fuentes-Santos (2016) se proponen test gráficos para caracterizar el patrón de los datos y poder determinar si estos se distribuyen de forma totalmente aleatoria o si por el contrario siguen alguna agrupación o colocación específica. Se define, para un proceso puntual estacionario e isotrópico, la K-función con el fin de poder determinar si se puede rechazar la hipótesis de que $\lambda(x)$ se distribuye de forma homogénea. Esta función caracteriza la estructura de segundo orden del proceso encargada de reflejar cualquier tendencia de los datos a aparecer agrupados, de forma independiente o regularmente espaciados y tiene la expresión:

$$K(t) = \lambda^{-1} E[\text{eventos a una distancia menor que } t \text{ de un evento arbitrario}] \tag{2}$$

la cual vale $K(t) = \pi t^2$ cuando se tiene un proceso Homogéneo de Poisson.

Cuando no se puede suponer que el proceso sea totalmente aleatorio, la K-función homogénea puede sobreestimar la dependencia, para solucionar este problema se define la K-función inhomogénea, donde ahora la intensidad no se supone constante. La K-función inhomogénea se puede estimar a través de la fórmula (3), propuesta por Baddeley:

$$\hat{K}_{inhom}(t) = \frac{1}{|A|} \sum_{x_i \in X \cap A} \sum_{x_j \in (X \cap A) \setminus \{x_i\}} \frac{I(\|x_i - x_j\| \leq t)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j) w_{ij}}. \tag{3}$$

Se puede emplear la estimación de la K inhomogénea con la finalidad de aplicar un contraste de independencia análogo al propuesto para CSR empleando la K homogénea, donde ahora aceptar

la hipótesis nula supondría que el diseño se corresponde con un proceso Inhomogéneo de Poisson donde la intensidad se supone variable.

2.4.1 ESTUDIO DE PROCESOS PUNTUALES

Uno de los requisitos para poder aplicar la teoría mostrada en Fuentes-Santos (2016) es que se tenga isotropía en los datos espaciales considerados. Esto requiere aceptar la hipótesis de que los incendios se distribuyen de forma isotrópica en el mapa de Alberta. En este caso la hipótesis de isotropía no es correctamente asumida, ya que si se considera una determinada posición del mapa de la Figura 1, por ejemplo el punto rojo mostrado, se puede ver que si uno se desplaza hacia el suroeste y cuenta el número de incendios del cuadrante correspondiente va a ser cero, mientras que si en cambio nos desplazamos hacia el noreste se espera una cantidad considerable, lo que se opone al principio de isotropía. Pese a no disponer de esta se desarrollarán los procesos puntuales y se expondrán sus resultados, teniendo en cuenta que se puede estar produciendo un error grande.

Para poder determinar si existe algún tipo de patrón en los datos se calculan las respectivas K-funciones homogénea e inhomogénea (2) y (3) para los datos de incendios ocurridos en la provincia de Alberta desde el año 1996 hasta el año 2014. En primer lugar, se realizará un estudio considerando el total de estos y a continuación se verá que ocurre si se estudian los datos recogidos de la forma bianual ya vista en secciones anteriores.

Con el fin de llevar a cabo este objetivo se utiliza la librería *spatstat* de Adrian Baddeley et al. (2015) de R (ver R Core Team (2017)). Esta librería permite implementar test gráficos para determinar si se puede asumir que la intensidad es homogénea así como obtener una estimación de la densidad puntual, tanto para la distribución de los datos como para variables asociadas a estos las cuales se introducirán finalmente como marcas.

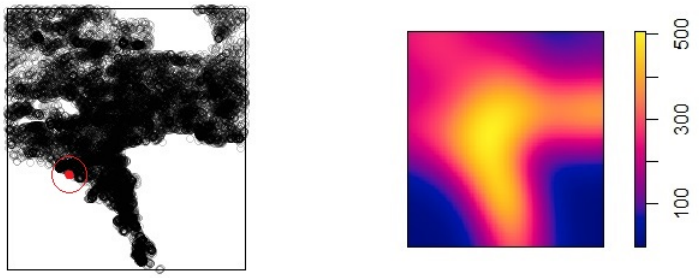


Figura 1: Mapa de los incendios documentados en la provincia de Alberta (Canadá) entre los años 1996 y 2014 tras proyectar las localizaciones en una región cuadrada a la izquierda. Estimación tipo núcleo de la intensidad a través de un kernel gaussiano isotrópico a la derecha.

Una forma de estudiar como son las propiedades de segundo orden de un proceso espacial puntual Homogéneo de Poisson es estimar el valor de la K-función (2), la cual permite medir la fuerza y el tipo de interacción entre los eventos del proceso puntual. Ripley propone un estimador insesgado de esta, el cual tiene la expresión

$$\hat{K}_{Ripley}(t) = \frac{|A|}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{|\{x_j : d(x_i, x_j) \leq t\}|}{w_{ij}}. \tag{4}$$

Dado que a veces no se puede trabajar bajo el supuesto de que el proceso es Homogéneo de Poisson es necesario obtener también una estimación de la K-función inhomogénea, para la cual se empleará la versión propuesta por Baddeley vista en (3). Calculando (4) y (3) y representando los resultados se observa que no se puede asumir que la intensidad sea constante, ya que los estimadores discrepan bastante del valor $K(t) = \pi t^2$. Para poder determinar si es posible aceptar que exista algún patrón espacial se calculan las bandas de confianza para la función K_{pois}, la cual representa el valor teórico que debería tener la K-función cuando el proceso es Homogéneo de Poisson, y se

representan estas junto a los estimadores de la K-función homogénea de Ripley (4). Este método gráfico permite determinar si existe algún patrón espacial en la disposición de los datos como podría ser de *agrupamiento o cluster*, *regular* o si en cambio los datos se distribuyen de *forma aleatoria*.

Proseguidamente se estima la intensidad de este proceso a través de métodos no paramétricos empleando un estimador tipo núcleo con un kernel gaussiano isotrópico.

Finalmente se estudia qué ocurre con la variable que recoge el número de hectáreas quemadas ("*assessment_hectares*"). Para poder introducirla en el contexto puntual se codifica como marca del número de incendios, de esta forma para cada localización en la que haya un dato este contará con el número de hectáreas quemadas por ese incendio. Para solventar posibles problemas a causa de los valores atípicos se considera oportuno trabajar tomando aquellos datos tales que su número de hectáreas quemadas es menor o igual que 0.22, dado que la mayoría de incendios no llegar a quemar 0.3 hectáreas (valor del tercer cuantil de la variable).

2.5. ESTUDIO DE SERIES TEMPORALES

Los datos empleados para elaborar este trabajo también contienen información acerca de la fecha y hora en que fueron descubiertos los incendios documentados. De acuerdo a esta, es interesante plantear un análisis temporal con la finalidad de determinar la existencia de algún patrón en el tiempo. Para este fin se procede a realizar un análisis de series de tiempo tanto del número medio de incendios en cada mes como de la variable que recoge el número de hectáreas quemadas ("assessment_hectares").

Dado que la estructura temporal de la base de datos correspondiente a cada "año de incendio", siendo el intervalo de meses en el cual se documentan los fuegos, depende del período considerado, se hace necesario la unificación de ambos formatos en uno común.

- Desde el año 1996 al 2001 el "año de incendio" va desde el día 1 de enero al día 31 de diciembre de ese mismo año.
- A partir del año 2002 el "año de incendio" se define desde el día 1 de abril de ese año al día 31 de marzo del año siguiente, de esta forma el año de incendio correspondiente al 2002 iría del 1 de abril del 2002 al 31 de marzo del año 2003.

Con el fin de poder desarrollar el estudio temporal se estructuran estas configuraciones en el segundo formato. De esta forma, se establece como "año de incendio", para cada año natural, el período comprendido desde el día 1 de abril al 31 de marzo del año siguiente. Para este propósito, se han empleado las bases de datos anuales que se encuentran en la página web del Gobierno de Alberta (véase https://www.albertaparks.ca/media/442827/nsr2005_final_letter.pdf), las cuales contienen la misma información pero separada por "año de incendio". De este modo, se organizan los datos con el formato estipulado y se juntan los resultados creando una nueva base de datos. Dado que las bases de datos individuales correspondientes a los años 2013 y 2014 no tienen registrado ningún incendio a partir de octubre, se descarta, ya que podrían faltar datos debido a que este fenómeno no se ha dado nunca en los años anteriores.

Una vez realizado este proceso se busca ajustar un **modelo ARIMA** adecuado que permita explicar tanto el comportamiento de la variable que contabiliza el número de incendios mensuales, como el número de hectáreas quemadas a través de los años considerados. Con la finalidad de testar cómo de bien va a predecir, se ajustarán mediante los datos correspondientes al período de años 1996-2011, para a continuación predecir los valores correspondientes al año 2012 comparando estos con los reales.

Se llevaron a cabo los dos siguientes análisis de series temporales:

1. número de incendios y
2. hectáreas quemadas.

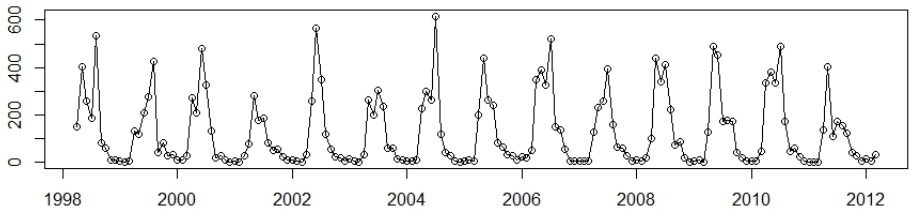


Figura 2: Serie temporal del número de incendios documentados.

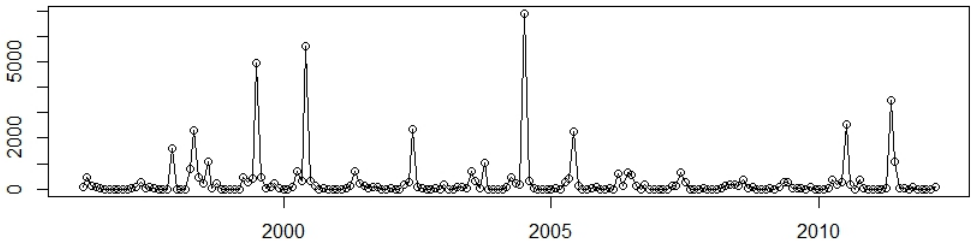


Figura 3: Serie temporal del número de hectáreas quemadas.

REFERENCIAS

Baddeley, A., Rubak, E. and Turner, R. (2015). Spatial Point Patterns: Methodology and Applications with R. London: Chapman and Hall/CRC Press, 2015. http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/

Alberta Agriculture and Forestry, http://agriculture.alberta.ca/acis/maps/.

Alberta Government, https://www.albertaparks.ca/media/442827/nsr2005_final_letter.pdf.

Alberta Wildfire, http://wildfire.alberta.ca/resources/default.aspx.

Fuentes-Santos I. (2016) Nonparametric inference for first-order characteristics of spatial and spatio-temporal point processes. Universidade de Santiago de Compostela, PhD Thesis.

Giraldo-Henao R. (2011) Estadística Espacial. Universidad Nacional de Colombia, Bogotá.

Historica Canada, http://www.thecanadianencyclopedia.ca/en/article/alberta/.

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

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

COORDENADAS XEOGRÁFICAS DE UNIDADES ESTATÍSTICAS E A SÚA ASIGNACIÓN: UN CASO TÍPICO

García Olano, J.R.[1], Iglesias Patiño, C.L.[1] e López Vizcaíno, M.E.[1]

[1] Instituto Galego de Estatística

RESUMO

Este relatorio presenta o estudo dun caso onde se comproba a calidade das coordenadas xeográficas asignadas ás entidades singulares do nomenclátor toponímico de Galicia e se valida unha técnica de cálculo dun centro dun conxunto de puntos no rural galego.

Palabras e frases chave: medidas de centralización bidimensionais, xeo-referenciación, validación externa

1. INTRODUCIÓN

Dende hai moito tempo as unidades estatísticas están referenciadas territorialmente. Así e todo, esta referencia non descendía do nivel municipal agás algunhas senlleiras excepcións como o nomenclátor de poboación (IGE, 2017 a).

Dunha xeira, cómpre refinar esta referenciación para podermos ofrecer información, inda que sexa dun xeito aproximado, dos espazos inferiores ao municipal como as entidades colectivas ou singulares ou calquera outro que non coincida exactamente cun termo municipal. Actualmente só dispomos desa fonte estatística para estudar o fenómeno do poboamento. O emprego das entidades singular e colectiva e dos conceptos de núcleo e diseminado había resolver estes problemas aínda que pode que non de xeito moi preciso, por exemplo, unha mesma aglomeración urbana interseca municipios distintos, o núcleo de Ferrol-Narón (60.894 + 32.963 = 93.857 hab) é menor có de Ourense (98.942 hab) no entanto, meirande có de Santiago de Compostela (79.807 hab). Ou amosa a presenza de zonas non-urbanas en concellos que se consideran urbanos, o núcleo do concello máis poboado de Galicia, Vigo (198.537 hab) é menor có da Coruña (213.418 hab) (IGE, 2017 a).

Doutra xeira, non debemos esquecer a captación dos datos sen a que non habería resultados que difundir. Por exemplo, canto máis preciso sexa o enderezo máis doada ha ser a localización do entrevistado e con iso habiamos abaratar os custos da recolla. Non é doado localizar unha vivenda no rural galego, adoito só se dispón do lugar (que podemos identificar coa entidade singular) e o concello. Nunha situación mellor, disponse do número, mais esta numeración adoito non é correlativa porque non segue as vías de comunicación, e da parroquia (entidade colectiva) que permite distinguir dous lugares con idéntica denominación dentro do mesmo concello.

Outro exemplo, de utilidade da entidade colectiva, é a variable nuclear de Eurostat grao de urbanización dos concellos que o IGE desagregou a este nivel e aplicou tamén ás parroquias (entidades colectivas ou singulares non incluídas nunha colectiva) para clasificalas en xa tres edicións 2001, 2011 e 2016 (IGE, 2016) e que pode ter ata consecuencias fiscais como a Orde do 9 de febreiro de 2017 da Consellería de Facenda.

Con respecto a entidade singular, podería ter a súa utilidade na xeo-localización de preto da cuarta parte da poboación de Galicia cuxo enderezo non se da en termos de vía (rúa, avenida, praza…) aínda que topa co problema de que falta a delimitación do seu perímetro e de que o nomenclátor de toponimia (Xunta de Galicia, 2017) non concorda co nomenclátor que o INE recompila des dos datos do Padrón Municipal de Habitantes (PMH) que subministran os concellos (IGE, 2017 a). En xeral, o nomenclátor toponímico é mais extenso có estatístico. Para estudar estoutra utilidade da entidade singular, o IGE realizou unha proba coa axuda do Instituto de Estudos do Territorio (IET) da Consellería de Medio Ambiente e Ordenación do Territorio.

Velaquí a estrutura do relatorio. Na seguinte sección presentamos a elección do espazo de proba, como realizamos a campaña, o instrumental empregado nela e mais unha técnica de asignación de coordenadas xeográficas a unidades estatísticas de nivel superior. Na sección de resultados, comparamos os datos recolleitos en campo cos inicialmente fornecidos polo IET e tamén os resultados da validación externa a que sometemos a devandita técnica. Na última sección presentamos unhas breves conclusións.

2. MATERIAL E MÉTODOS

Seleccionamos o concello da Baña, da provincia da Coruña, en particular o seu distrito primeiro, sección primeira, que corresponde a dúas parroquias (San Vicenzo da Baña e San Mamede de Suevos), por

- figurar na mostra que o IGE tiña que recoller para a Enquisa de Poboación Activa (EPA),
- estar situado preto de Santiago de Compostela,
- non ser excesivamente extenso,
- ter un perfil rural –núcleo/diseminado– característico da nosa comunidade autónoma,
- e mais ser factible a identificación da posición no territorio de forma sinxela, pois os límites e a denominación das parroquias e lugares atópanse indicados nas vías, mediante a súa correspondente sinalización vertical.

O IET proporcionou ao IGE unha relación de 23 entidades coas súas correspondentes coordenadas (X,Y) en unidades UTM para o fuso 29 desa sección censual da Baña. Como íamos utilizar un dispositivo de navegación terrestre, en particular un TOMTOM que non admite as coordenadas UTM, foi preciso converter estas coordenadas nun formato compatible co TOMTOM, isto é, ou ben en formato graos, minutos, segundos ou en graos decimais (este foi o elixido, ao ser menos gravosa a introdución manual un a un dos datos no dispositivo).

Para esa transformación, utilizamos a app de conversión que o Instituto Geográfico Nacional de España (IGN) pon a disposición dos interesados na súa páxina web. Decidímonos por este conversor por ser o organismo oficial estatal competente na materia e, polo tanto, consideramos, a priori, que debe ofrecer máis garantías do seu correcto deseño e funcionamento.

Confeccionamos unha sinxela táboa para facilitar ao mesmo tempo a introdución "manual" dos datos para cada entidade no dispositivo e a recolla das lecturas, unha vez acadada a posición asignada, que nos devolvía o devandito dispositivo. Para o proceso de comprobación, simulamos un percorrido típico en campo dun enquisador para a recolla de datos da EPA. Ordenamos as entidades singulares da táboa pola súa proximidade xeográfica partindo da situada máis ao norte e realizando o percorrido sobre o eixo norte-sur, sentido sur, movéndonos alternativamente e transversalmente a este eixo dirección leste-oeste, en función das posicións ordenadas das entidades.

Tras o traballo de campo procedeuse a realizar unha comparanza coas coordenadas orixinais fornecidas polo IET por regresión linear. Isto permitiu unha macrodepuración onde detectamos un erro grosero de transcrición. Asemade aproveitamos estes datos recolleitos no campo para validar unha técnica de asignación de coordenadas xeográficas a unha unidade estatística de nivel superior (neste caso, a entidade singular de poboación) a partirmos das coordenadas de unidades estatísticas de nivel inferior (os inmóbeis) que presentamos a seguir.

Distinguimos o núcleo do diseminado dunha entidade singular e tratámolos por separado. Tanto para o núcleo cando non existe vía como para o diseminado, asignamos as coordenadas correspondentes a un punto medio ponderado do núcleo ou do diseminado correspondente porque o centroide dos puntos (vector de medias aritméticas) non foi satisfactorio nalgúns casos. O ponderador dun punto (x_i, y_i) é igual a d_i^{-1} sendo $d_i = d\left((x_i, y_i); (\overline{x_{-i}}, \overline{y_{-i}})\right)$ con $(\overline{x_{-i}}, \overline{y_{-i}})$ o centroide da nube de puntos orixinal agás o propio punto (x_i, y_i). No caso de que exista vía, asignamos o veciño máis próximo se dispoñen de número ou o centroide se non dispoñen del.

3. RESULTADOS

As coordenadas das 23 entidades facilitadas polo IET non presentan diferenzas apreciables coa posición sobre o terreo dos centros das entidades determinados en campo des dun punto de vista global, a regresión linear realizada entre elas amosa coeficientes de determinación superiores a 0'99, e tampouco des dun punto de vista de cada entidade en 19 casos. Das catro restantes, en 2 entidades, o dispositivo posicionounos, aproximadamente a uns 20 metros do centro de cada entidade. Noutra entidade, adxacente a outras dúas e que se suceden sobre a mesma vía sen apenas solución de continuidade, cometeuse un erro duns 160 m. E a máis apreciable, situounos cun erro duns 200 m mais en campo aberto, isto é, arredada

das edificacións. Velaquí un mapa do Concello da Baña destacando a súa sección censual 1 onde amosamos graficamente os resultados.

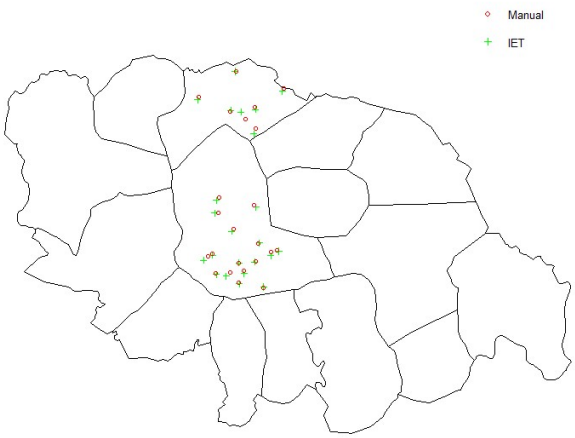


Figura 1: Comparanza entre as entidades singulares obtidas na campaña coas fornecidas polo IET

Asignamos coordenadas xeográficas ás entidades singulares do nomenclátor estatístico mediante a aplicación da técnica, descrita na sección anterior de material e métodos, aos inmobles que figuran no censo de poboación (INE, 2011).

Unha simple regresión linear entre os valores asignados, pola devandita técnica, a 20 das 23 entidades e os valores observados en campo amosa coeficientes de determinación superiores a 0'97. No entanto, a hipótese de normalidade non se sostén ao 0'05. Existen algúns puntos problemáticos, os máis deles sen moita transcendencia des dunha óptica da estatística pública. As tres entidades que faltan explícanse pola antedita diferenza entre o nomenclátor estatístico e o nomenclátor toponímico. Presentamos, a seguir, o mapa do Concello da Baña destacando a súa sección censual 1 onde amosamos graficamente os resultados.

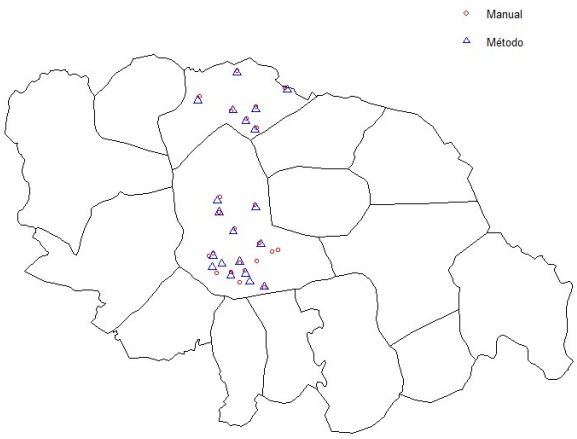


Figura 2: Comparanza das entidades singulares obtidas na campaña coas asignadas pola técnica

4. CONCLUSIÓNS

Este relatorio reflicte un claro exemplo de observación controlada que debería ter moito máis desenvolvemento na estatística pública do que ten arestora, de feito abúsase da declaración da unidade informante.

A calidade das coordenadas do IET é máis que aceptable des dun punto de vista estatístico. Des dun punto de vista xeográfico, pode precisar unha validación externa in situ, unha das razóns da carestía dos datos xeográficos. Calquera campaña ou saída a campo pode ser de utilidade para a captación destes datos e non debemos esquecer que o IGE realiza, ao ano, dúas enquisas de tamaño considerable: a Enquisa Estrutural a Fogares (IGE, 2017 b) e a EPA (IGE-INE, 2017). Cómpre prever outros exemplos de validación externa ou mellor aínda un plan para o futuro.

A técnica de asignación empregada pode funcionar ben para o rural galego. De feito, un dos devanditos puntos problemáticos, resultado desta técnica, está indirectamente relacionado co punto onde o erro era máis apreciable con respecto ás coordenadas orixinais. Podería empregarse para unha imputación de coordenadas pola "media", para asignar o centro de poboamento dun conxunto de puntos, para rotular os mapas…

REFERENCIAS

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

IGE (2017 a) Nomenclátor e Explotación do Padrón Municipal de Habitantes (a 1 de xaneiro de 2016) http://www.ige.eu/igebdt/esq.jsp?paxina=002001&ruta=nomenclator/nomenclator.jsp

IGE (2017 b) Enquisa estrutural a fogares Metodoloxía http://www.ige.eu/estatico/pdfs/s3/metodoloxias/met_EEF_2016_gl.pdf

IGE-INE (2017) Enquisa de poboación activa Metodoloxía http://www.ige.eu/estatico/pdfs/s3/metodoloxias/met_EPA_2005_Dupli_gl.pdf

IGN Servicio web de transformación de coordenadas http://www.ign.es/wcts-app/

INE (2011) Censo de población y viviendas http://www.ine.es/censos2011_datos/cen11_datos_metodologia.htm

Xunta de Galicia (2017) Toponimia oficial e Nomenclátor de Galicia http://www.xunta.gal/toponimia-e-nomenclator

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

ANÁLISE ESPAZO-TEMPORAL DOS APELIDOS TOPONÍMICOS EN GALICIA

María José Ginzo Villamayor[1] e Rosa M. Crujeiras Casais[1]

[1]Departamento de Estatística, Análise Matemática e Optimización, Universidade de Santiago de Compostela, mariajose.ginzo@usc.es e rosa.crujeiras@usc.es

RESUMO

Na análise dos apelidos da poboacion de Galicia é doado identificar apelidos que coinciden con nomes de lugares, é dicir, apelidos toponímicos. Ditos apelidos non soen atoparse no mesmo lugar da denominación, se non incialmente noutros lugares próximos e co paso do tempo, en lugares máis alonxados. Por exemplo o apelido Outomuro, é un apelido galego de orixe toponímica que ten a súa orixe no concello de Cartelle (na provincia de Ourense), presenta unha distribución moi concentrada nos concellos próximos e co paso do tempo noutros non tan próximos como é o caso de Vigo. Neste traballo tratarase de analizar a evolución espazo-temporal dos apelidos toponímicos en Galicia, mediante o uso de índices de diversidade.

Palabras e frases chave: Apelidos toponímicos, índices de diversidade.

1. INTRODUCIÓN

Os apelidos toponímicos (de nome de lugar) fan referencia a apelidos cuxa orixe se atopa no lugar do que procedían ou no que tiñan terras a persoa ou familia asociadas a este. A teoría xeolingüística indica que ditos apelidos presentan unha distribución espazo-temporal característica, de xeito que nun primeiro momento, aparecen en zonas próximas ao lugar de denominación, e co paso do tempo vanse alonxando. É dicir, presentan evolucións non necesariamente sinxelas, como ocorre en xeral nos estudos onomásticos.

En Ginzo-Villamayor e Crujeiras (2015) propónse un modelado do patrón espacial dos apelidos en Galicia, considerando o efectos da covariable número de habitantes por concellos, axustando o modelo proposto por Besag et al. (1991) (modelo BYM), adaptado a este contexto. Investigando só o patrón espacial dos apelidos non é posible concluír nada sobre outra das compoñentes de variación, a temporal, que poder ser igualmente de interese. Tamén en Ginzo-Villamayor e Crujeiras (2015), o modelo anterior esténdese facilmente ao caso espazo–temporal incluíndo o tempo. Na Figura 1 tense o axuste dos modelos empregando o método baseado na aproximación por integradores de Laplace aniñados INLA[1]), proposto por Rue et al. (2009), para os apelidos Miramontes, Galdo, Cudeiro, Farrapeira, Padreda e Outomuro. Estes nove apelidos son de orixe toponímica: Miramontes (no concello de Santiago de Compostela), Galdo (no concello de Mañón), Cudeiro (no concello de Ourense), Farrapeira (no concello de Salceda de Caselas), Padreda (no concello de Vilar de Barrio) e Outomuro (no concello de Cartelle).

Para unha análise máis polo miúdo dos apelidos representados na Figura 1, plantéase neste traballo o estudo da distribución de apelidos toponímicos. A modo de exemplo, na Figura 2 represéntase a distribución do apelido Outomuro, para os nados entre 1920 e 2011. Téñense 9 mapas, das provincias de Pontevedra e Ourense, únicas provincias onde aparece este apelido en Galicia e a súa distribución ao longo dos anos. Dun modo similar ao apelido Outomuro, tense na Figura 4 a distribución do apelido Ginzo[2], apelido tamén de orixe toponímica. Localízase principalmente na comarca da Mariña Oriental máis precisamente na Pontenova. Ginzo é a castelanización de Xinzo,

[1]Bayesian computing with INLA!, (2009) URL: http://www.r-inla.org/.

[2]Dicionario dos apelidos galegos, que se está a elaborar na sección de Onomástica do Instituto da Lingua Galega (USC) dirixido por Ana Isabel Boullón Agrelo.

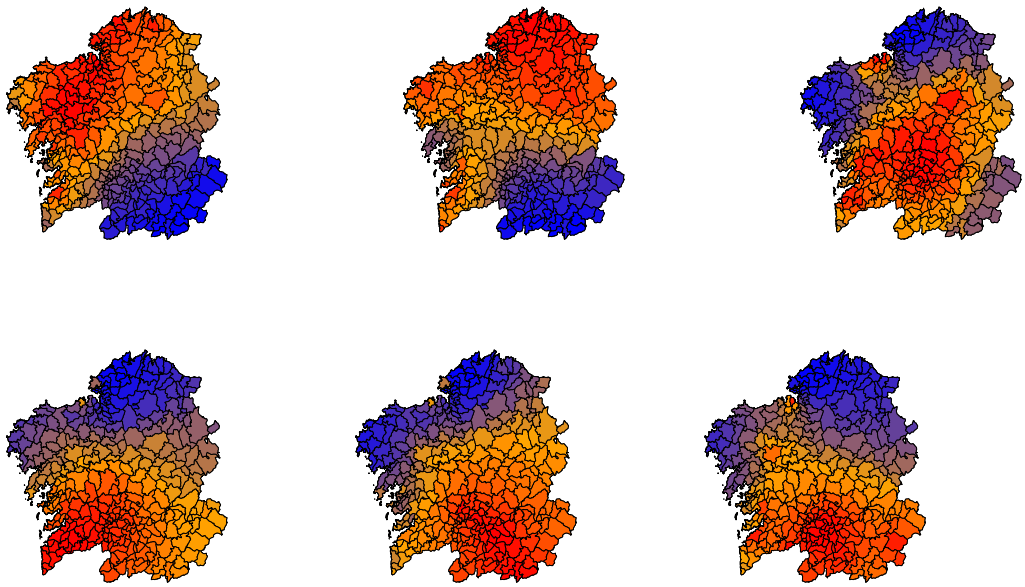


Figura 1: Dende arriba cara abaixo e de esquerda a dereita, media a posteriori para a parte estruturada do campo latente para o apelidos Miramontes (da zona noroeste de Galicia), Galdo (zona nordeste), Cudeiro (zona centro), Farrapeira (zona suroeste) e Padreda e Outomuro (zona sueste). Modelo presentado en Ginzo-Villamayor e Crujeiras (2015).

topónimo de orixe prerromana (med. Genitio) que se repite en 12 lugares de Galicia, ademais de dar nome a un concello (Xinzo de Limia). Os apelidos proceden dos que están localizados na Mariña lucense e na estrema con Asturias, dalgún dos lugares que hai nos concellos de Foz, A Pontenova ou mesmo Tapia de Casariego (Asturias).

En contraposición, na Figura 5 tense a distribución do apelido Fernández. É un apelido de tipo patronímico con sufixo "-ez", que era o procedemento usado na Idade Media para significar "fillo de", porque daquela os apelidos non se transmitían por herdanza, e cambiaban de xeración en xeración. Trátase dunha derivación do nome propio Hernando, berce e orixe de Fernández ou Hernández. É difícil desligar os dous apelidos xa que ambos, Hernández e Fernández, teñen a súa orixe no nome propio Ferrán, que significa esforzado guerreiro, derivación da palabla céltica "Har", traducida por "guerra" e "poder". Partindo de Ferrán ou Hernán, derívanse Fernando e Fernández e Hernando e Hernández. En catalán consérvase igual á súa orixe, Ferrán, nome que en castelán convértese en Fernando. É posible que o apelido Fernández fose tomado do rei Fernando. Como se pode observar a distribución é totalmente diferente a de Outomuro ou Ginzo, nos dous primeiros casos o apelido concéntrase nos concellos onde xorde ou limítrofes e vaise expandindo pouca a pouco. Lasker e Kaplan (1983) levan a cabo un estudo similar para apelidos británicos no que conclúen resultados similares, como por exemplo o caso do apelido McKinley.

1.2 DATOS

Para levar a cabo este traballo empregáronse dúas fontes de datos, por un lado o Nomenclátor[3] de Galicia (ano 2012) que conta con 37.308 rexistros. O Nomenclátor contén a toponimia das provincias, parroquias, concellos e lugares de Galicia, de acordo cos ditames da Comisión de Toponimia e coas normas ditadas pola Xunta de Galicia. A segunda fonte de datos empregada é o censo de apelidos de Galicia (correspondente ao ano 2011) que contén 20.757 apelidos diferentes.

[3]Nomenclátor URL:
http://abertos.xunta.gal/catalogo/territorio-vivienda-transporte/-/dataset/0270/nomenclator-galicia/.

2. METODOLOXÍA

2.1 OPERACIÓNS CO NOMENCLÁTOR

Neste primeiro traballo coa toponimia, de tódolos núcleos do Nomenclátor elimináronse as partículas do tipo"As", "Os", "de arriba", "de abaixo", "grande" ou "grandes", "pequeno", "Santa", "San", ... Unha vez eliminadas as partículas anteriores quedamos con aqueles "núcleos" que coinciden exactamente con apelidos. Para este estudo, non se consideraron derivacións, é dicir, Lameira $\longleftrightarrow$ Lameiro ou viceversa. O número total de coincidencias é preto de 16.000 (43.48 % do nomenclátor). Fíxose unha primeira análise gráfica sobre a distribución espacial de algún destes apelidos e a raíz desta análise, xorde a idea de investigar a súa distribución espacial. Obsérvase que apelidos como "Casaldoeiro" (coincide con un núcleo do concello de Brión), "Someso" (no concello da Coruña), "Mallou" (no concello de Santiago de Compostela) e "Allariz" (no concello de Marín) aparecen con maior frecuencia naqueles concellos próximos ao lugar de orixe.

Por outra banda hai apelidos como "Toande", "Caberta", "Marrondo", "Negrelos" ou "Corredoura", que aparecen en concellos diferentes a aqueles onde hai núcleos con ese nome e si limítrofes.

Dado que o número de apelidos dos que se sospeita a súa orixe toponímica é moi elevando, determinouse unha serie de apelidos que cubren todo o territorio galego de xeito máis ou menos uniforme. O número de apelidos selecciondos é noventa, distribuídos en cinco zonas (dezaoito por zona), a zona noroeste (Táboa 1), a zona nordeste (Táboa 2), a zona nordeste (Táboa 3), a zona nordeste (Táboa 4) e a zona nordeste (Táboa 5) de Galicia.

AMENEDO	ANCA	ANTELO	BANDIN	CAAMAÑO	CARNOTA
CUNDINS	FORNIS	GRILLE	LASTRES	LEIS	LENS
LISTE	MIÑO	MIRAMONTES	MORONO	PRESEDO	PRIMOY

Táboa 1: Apelidos da zona noroeste de Galicia.

AMIEIRO	ANDION	BASANTA	CABARCOS	FUNCASTA	GALDO
GOAS	GOIRIZ	GRUEIRA	LADRA	LIZ	LODOS
MARFUL	MOSTEIRIN	PALEO	PERTEGA	PIQUIN	QUENDE

Táboa 2: Apelidos da zona nordeste de Galicia.

BARDELAS	CARDEIRO	FISTEOS	GURGUEIRO	LAMAZARES	LAREO
MARAÑIS	NAVAZA	ORZA	PALMOU	PAMPIN	REGADIO
REGALADE	RIADIGOS	SABORIDA	SOENGAS	VENCE	ZOBRA

Táboa 3: Apelidos da zona centro de Galicia.

Para representar as 5 zonas consideradas, lévase a cabo o procedemento que se describe a continuación. Trátase de dividir Galicia en 5 zonas: noroeste, nordeste, centro, suroeste e sueste. Pola definición de noroeste, nordeste, suroeste e sueste, parece estar máis o menos claro qué concellos pertencen a cada unha desas zonas. É máis complicado definir os concellos que están no centro. Partindo de que o concello de Lalín na provincia de Pontevedra considérase o centro xeográfico de Galicia, proponse a seguinte definición, dirase que están no centro aqueles concellos que estean na "zona q". Defínese "zona q" aqueles concellos cuxa coordenada UTM-X está entre o primeiro e terceiro cuartil das coordenadas UTM-X de Galicia e ademais a coordenada UTM-Y está entre o entre o primeiro e terceiro cuartil das coordenadas UTM-Y de Galicia (véxase a Figura 3).

Considéranse cinco apelidos dos noventa toponímicos para non facer demasiado amplo este resumo. Na Táboa 6 tense a distribución dos apelidos Caamaño, Cudeiro (o terceiro na primeira fila na Figura 1), Galdo (o segundo na primeira fila na Figura 1), Miramontes (o primeiro da

AMOEDO	BANGUESES	BESADA	BOULLOSA	BOUZADA	BUA
FARRAPEIRA	GARGAMALA	LORES	LUSQUIÑOS	MAGARIÑOS	MEAÑO
MORAÑA	NAVIA	OITAVEN	OUTERELO	PALLAS	PINTOS

Táboa 4: Apelidos da zona suroeste de Galicia.

BARAZAL	CERREDELO	CORBILLON	CORZO	CUDEIRO	DONIZ
GUDIÑA	GUNTIÑAS	LEVICES	LIMERES	MEDEIROS	MONDELO
NOVOA	OUTOMURO	PADREDA	PEAGUDA	SABUZ	SALPURIDO

Táboa 5: Apelidos da zona sueste de Galicia.

Figura 1) e Mondelo para cada unha das zonas consideradas de Galicia. Ademais, cada un deles se identifica cunha zona de Galicia.

2.2 ÍNDICES PARA CUANTIFICAR A DIVERSIDADE DE APELIDOS

Considérese o índice $i = 1, \ldots, n$ para denotar unha certa rexión xeográfica. Cada rexión ten asociada unha colección S_i de apelidos. O número total de apelidos de unha certa rexión i denótase por n_i. En Ginzo-Villamayor e outros (2013) preséntase a definición de diferentes medidas de similaridade entre apelidos, baseadas no concepto de isonimia, se ben existen outras alternativas, como o índice de Hill, "número efectivo de especies" (Hill, 1973). Hill (1973) define unha familiar paramétrica de "números de diversidade" definida na ecuación (1).

$$J_t(\lambda) = \left(\sum_{k \in S_i} p_{ki}^{\lambda} \right)^{\frac{1}{1-\lambda}}, \tag{1}$$

onde p_{ki} denota a frecuencia relativa do apelido k na rexión i, no instante t e $\lambda \geq 0$.

A idea fundamental dunha familia de índices paramétricos é que o parámetro libre, neste caso λ, permite ao usuario considerar non só unha medida á vez senón tamén avaliar e comparar varios simultaneamente. De feito, para valores específicos de λ, a familia (1) correspóndese con varias medidas clásicas de diversidade, para $\lambda = 0$, $J_t(0) = S$, obtense a riqueza das especies, neste caso dos apelidos; para $\lambda = 1$ a exponencial da entropía de Shannon[4] $J_t(1) = e^{H_t}$ (Shannon, 1948), e para $\lambda = 2$ a "inversa" do índice de Simpson[5] $J_t(2) = D_t^{-1}$ (Simpson, 1949). As expresións de H_t (entropía de Shannon) e D_t (índice Simpson) veñen definidas respectivamente nas ecuacións (2) e (3):

$$H_t = -\sum_{i}^{S} p_i^t \log p_i^t, \tag{2}$$

$$D_t = \sum_{i}^{S} \left(p_i^t\right)^2. \tag{3}$$

Considerando dous estatísticos de bondade de axuste, moi utilizados, a relación de verosimilitude G e o χ^2 de Pearson, téñense as ecuacións (4) e (5):

$$G = 2\sum_{i=1}^{S} n_i \log\left(\frac{n_i S}{n}\right) \tag{4}$$

$$\chi^2 = \sum_{i=1}^{S} \frac{(n_i - \frac{n}{S})^2}{\frac{n}{S}}. \tag{5}$$

[4]Claude Elwood Shannon (1916 - 2001): matemático americano, enxeñeiro electrónico e criptógrafo, coñecido como o "o pai da teoría da información".

[5]Edward Hugh Simpson (1922 -): estatístico británico.

Apelido	NOROESTE	NORDESTE	CENTRO	SUROESTE	SUESTE
CAAMAÑO	86,37 %	2,33 %	2,29 %	8,67 %	0,34 %
CUDEIRO	6,40 %	3,20 %	84,80 %	1,60 %	4,00 %
GALDO	28,82 %	70,55 %	0,16 %	0,47 %	0,00 %
MIRAMONTES	92,65 %	1,29 %	2,39 %	3,68 %	0,00 %
MONDELO	9,06 %	21,01 %	7,61 %	6,88 %	55,43 %

Táboa 6: Distribución dos apelidos Caamaño, Cudeiro, Galdo, Miramontes e Mondelo nas 5 zonas de Galicia.

Cressie e Read (1984) xeneralizan esta idea introducindo e estudando unha familian paramétrica para o estatístico de bondade de axuste o que incorpora tanto a G como a χ^2. Coa inclusión dun parámetro libre λ, T. Read e N. Cressie, derivaron a seguinte forma paramétrica para o estatístico xeneralizado de bondade de axuste (Cressie e Read, 1984; Read e Cressie, 1988).

$$I_n(\lambda) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{S} n_i \left[\left(\frac{n_i}{n/S} \right)^{\lambda} - 1 \right]. \tag{6}$$

Variando distintos valores de λ nesta expresión, obtéñense diferentes estatísticos de bondade de axuste. A ecuación, $I_n(\lambda)$, non está definida para $\lambda = -1$ e $\lambda = 0$, no límite de $\lambda = -1$ e $\lambda = 0$ pode ser calculado.

Read e Cressie (1988) céntranse es aspectos estatísticos da familia, estatísticos de bondade de axuste na ecuación (6) (distribución asintótica e comportameto en termos de tamaño e potencia, entre outros). En particular, estudan a súa conexión coas medidas de diverxencia de información teórica, por exemplo a diverxencia de Kullback-Leibler.

3. RESULTADOS

Na Táboa 7 téñense os valores do estatístico de Cressie e Read para os valores de λ de -0,5; 0,5 e 1 e para os distintos anos de nacemento. É importante sinalar o proceso de urbanización que produciu en Galicia a partires do ano 1965. Isto hai que telo en conta nas columnas "1970-79", "1980-89", "1990-99" e "2000-11" das Táboas 7 a 9. Este proceso queda patente tamén nas Figuras 2 e 4, pois a partir do 1965, pode apreciarse o salto destes apelidos dende concellos limítrofes o do lugar de orixe a núcleos de maior poboación ou grandes cidades. Na Táboa 8 téñense os valores do estatístico de Cressie e Read para os valores de λ de 1,5; 2 e 2,5 e para os distintos anos de nacemento; así mesmo na Táboa 9 téñense os valores do estatístico para os valores de λ de 3, 5 e 10.

Como traballo futuro preséntanse dous retos: por unha banda realizar un contraste de hipóteses, para comparar as distribucións das proporcións dos apelidos toponímicos e o segundo reto é estudar o espallamento espacial destes apelidos dende o seu lugar de orixe.

AGRADECEMENTOS

María José Ginzo e Rosa M. Crujeiras agradecen o apoio dos proxectos MTM2016-76969-P e MTM2013-41383-P do Ministerio de Economía e Competitividad, a Rede IAP P7/06 de Belgian Science Policy, a rede Tecnoloxías e Análise dos Datos Lingüísticos[6], ao investigador Xulio Sousa Fernández[7], e o Instituto Galego de Estatística por ceder os datos para esta investigación.

[6]Páxina web de TecAnDali http://ilg.usc.es/tecandali/.
[7]Instituto da Lingua Galega (USC).

λ	Zona	1920-29	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	1990-99	2000-11
-0,5	NOROESTE	914,83	1365,94	3990,89	3279,87	4380,34	3050,20	5206,76	1876,52	3359,37
-0,5	NORDESTE	426,52	694,86	472,54	383,33	1201,28	680,65	467,09	377,26	338,51
-0,5	CENTRO	414,78	751,60	1498,23	2230,31	5266,78	793,00	1367,43	283,22	341,68
-0,5	SUROESTE	560,67	921,96	1244,97	2991,05	2685,81	2119,02	2255,66	4681,35	1772,07
-0,5	SUESTE	291,13	496,17	251,57	237,07	408,42	116,97	223,37	33,49	354,69
0,5	NOROESTE	1077,53	1533,20	4615,45	6135,90	5658,51	3323,52	7974,23	2807,70	5844,24
0,5	NORDESTE	528,44	893,73	453,87	360,87	1233,34	694,17	673,17	485,60	481,38
0,5	CENTRO	600,30	789,93	2058,08	2779,73	12011,34	1191,99	2139,20	332,19	523,67
0,5	SUROESTE	676,88	898,80	1253,90	4358,85	5128,30	2519,61	2892,18	7689,52	2536,58
0,5	SUESTE	288,16	484,38	285,82	209,76	464,24	117,44	346,44	31,46	459,25
1	NOROESTE	1475,80	1980,14	6245,99	11941,73	8501,24	4242,73	13595,90	4436,16	10678,56
1	NORDESTE	718,46	1299,38	503,40	394,07	1479,16	814,29	978,41	641,69	731,11
1	CENTRO	912,87	937,47	3062,42	3868,17	29058,46	1890,63	3485,80	406,97	778,27
1	SUROESTE	867,89	1044,26	1504,44	7062,03	10501,23	3341,68	4146,97	13933,80	3790,85
1	SUESTE	329,49	560,42	350,06	218,21	586,70	132,00	541,51	32,32	644,53

Táboa 7: Estadístico de Cressie e Read para os valores de λ de -0,5; 0,5 e 1 e para os distintos anos de nacemento.

REFERENCIAS

Besag, J., York, J. e Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. Annals of the institute of statistical mathematics, 43, 1–59.

Cressie, N.A.C., e Read, T.R.C. (1984) Multinomial Goodness-of-Fit Tests. Journal of the Royal Statistical Society B, 46, 440–464.

Ginzo-Villamayor, M.J.; Crujeiras, R. M. e Sousa Fernández, X. (2013) Surname patterns in Galicia. Libro de Actas do congreso. XI Congreso Galego de Estatística e Investigación de Operacións. A Coruña (España).

Ginzo-Villamayor, M.J. e Crujeiras, R. (2015). Modelización Onomástica. Libro de actas: ACTAS XII Congreso Galego de Estatística e Investigación de Operacións Lugo, 22-23-24 de outubro de 2015. ISBN 978-84-8192-522-7. (pp: 292-299). XII Congreso Galego de Estatística e Investigación de Operacións. Lugo (España).

Hill, M.O. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology, 54, 427-432.

Lasker, G.W. e Kaplan, B.A. (1983) English Place-Name Surnames Tend to Cluster Near the Place Named. Names, 31(3), 167–177.

Rue, H., Martino, S. e Chopin, N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. Journal of the Royal Statistical Society, Series B, 71, 319–392.

Shannon, C.E. (1948). A mathematical theory of information. Bell System Technical Journal, 27, 379–423.

Simpson, E.H. (1949). Measurement of diversity. Nature, 163, 688.

λ	Zona	1920-29	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	1990-99	2000-11
1,5	NOROESTE	2313,81	2861,99	9569,57	28130,94	15011,54	6066,36	27817,35	8102,30	23339,35
1,5	NORDESTE	1086,48	2202,56	597,83	459,17	1949,26	1035,49	1580,30	929,68	1280,25
1,5	CENTRO	1579,40	1214,46	5226,68	6086,44	87241,57	3505,86	6582,21	537,45	1287,30
1,5	SUROESTE	1215,35	1322,32	1975,91	13525,52	26596,92	4964,42	6817,87	30623,49	6428,02
1,5	SUESTE	407,60	709,21	466,73	239,77	814,66	158,84	956,32	34,37	1021,55
2	NOROESTE	4075,85	4562,54	16307,43	76380,05	30291,31	9577,69	66124,21	16641,58	58518,48
2	NORDESTE	1794,05	4250,22	754,46	566,41	2794,79	1413,11	2781,57	1461,13	2516,92
2	CENTRO	3029,10	1704,22	9998,10	10615,80	301592,78	7404,46	13977,49	759,94	2330,40
2	SUROESTE	1840,53	1805,71	2798,34	29626,40	78303,04	8146,93	12614,09	78094,49	12098,15
2	SUESTE	540,93	971,85	671,80	276,54	1225,84	203,39	1859,06	37,71	1794,47
2,5	NOROESTE	7895,44	7898,67	30377,25	229195,27	67754,70	16450,48	176671,70	37428,78	162209,52
2,5	NORDESTE	3181,59	9068,92	1005,46	734,88	4317,12	2051,03	5239,00	2461,36	5403,85
2,5	CENTRO	6283,42	2567,93	20932,80	20123,63	1142264,81	17293,63	32435,19	1141,80	4538,37
2,5	SUROESTE	2983,34	2635,60	4221,16	71789,40	254737,35	14546,92	25741,98	221883,12	24729,53
2,5	SUESTE	764,23	1428,18	1033,50	333,27	1973,26	275,40	3887,53	42,57	3425,19

Táboa 8: Estadístico de Cressie e Read para os valores de λ de 1,5; 2 e 2,5 e para os distintos anos de nacemento.

λ	Zona	1920-29	1930-39	1940-49	1950-59	1960-69	1970-79	1980-89	1990-99	2000-11
3	NOROESTE	16469,25	14624,64	60857,45	737740,79	163582,64	30290,08	515579,91	90192,51	483649,68
3	NORDESTE	5974,38	20837,57	1407,06	997,05	7112,48	3139,47	10399,90	4391,91	12381,00
3	CENTRO	13819,81	4115,15	46992,15	40739,31	4601854,36	43521,12	80377,34	1808,82	9364,27
3	SUROESTE	5122,53	4077,86	6711,68	187313,19	885957,21	27848,02	56855,77	681297,42	53889,85
3	SUESTE	1140,99	2229,62	1681,51	418,07	3358,98	391,86	8588,81	49,30	6977,34
5	NOROESTE	484394,45	266463,98	1585161,74	117519147,84	8611456,39	552878,81	62585210,34	4651745,45	58643669,62
5	NORDESTE	108109,84	879127,35	7941,46	4778,45	83066,34	25312,38	224894,99	65677,77	496810,01
5	CENTRO	458465,80	41951,65	1842460,98	1039902,26	1717702865,61	2677117,65	4474057,25	16863,72	246499,75
5	SUROESTE	66524,72	36054,56	62128,71	13350441,69	186443409,59	599208,26	2257633,05	94731935,44	1817820,67
5	SUESTE	8578,76	20507,17	17227,92	1448,66	41482,40	2339,25	286317,63	109,53	179542,49
10	NOROESTE	7,09E+09	1,39E+09	2,25E+10	1,09E+14	5,51E+11	3,09E+09	3,68E+13	3,20E+11	3,08E+13
10	NORDESTE	4,70E+08	2,96E+10	2,56E+06	8,66E+05	1,64E+08	1,65E+07	1,33E+09	1,86E+08	1,42E+10
10	CENTRO	7,88E+09	5,32E+07	6,36E+10	1,18E+10	1,25E+16	2,42E+11	3,21E+11	1,79E+07	2,78E+09
10	SUROESTE	1,42E+08	3,55E+07	5,50E+07	1,90E+12	3,26E+14	5,77E+09	9,56E+10	7,40E+13	4,20E+10
10	SUESTE	5,18E+06	2,14E+07	1,79E+07	1,31E+05	7,82E+07	7,24E+05	4,93E+09	2,01E+03	1,83E+09

Táboa 9: Estatístico de Cressie e Read para os valores de λ de 3, 5 e 10 e para os distintos anos de nacemento.

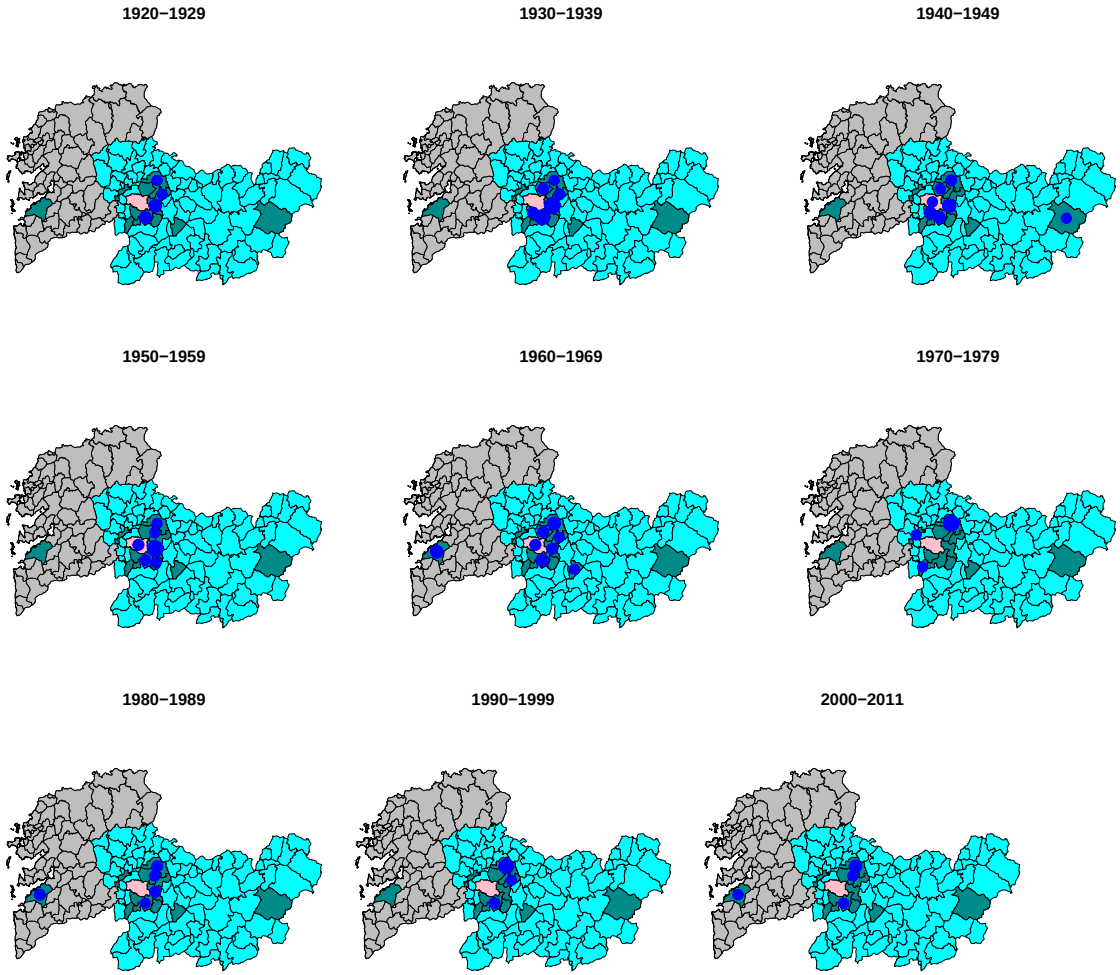


Figura 2: Mapas de distribución do apelido Outomuro cos nacidos dende 1920 ata 2011 en diferentes concellos de Galicia. Na cor gris se representan os concellos da provincia de Pontevedra e en cor cyan os da provincia de Ourense. En cyan máis oscuro señálanse todolos concellos onde con xente que leva este apelido no grupo de idade correspondente. En cor rosa se representa o concello de Cartelle, Outomuro é unha localidade deste concello. Nos concellos que ademáis do fondo cyan oscuro hai puntos azuís, estes representan, que houbo xente con ese apelido nacida no intervalo correspondente.

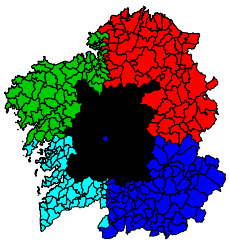


Figura 3: Mapa de Galicia dividido en 5 zonas segundo a definición de "zona q", para os concellos do centro. En cor verde se representa a zona noroeste, en cor vermello a zona nordeste, en negro a zona centro, en azul turquesa a zona suroeste e en azul escuro a zona sueste. O punto azul central representa o centro xeográfico de Galicia (no concello de Lalín).

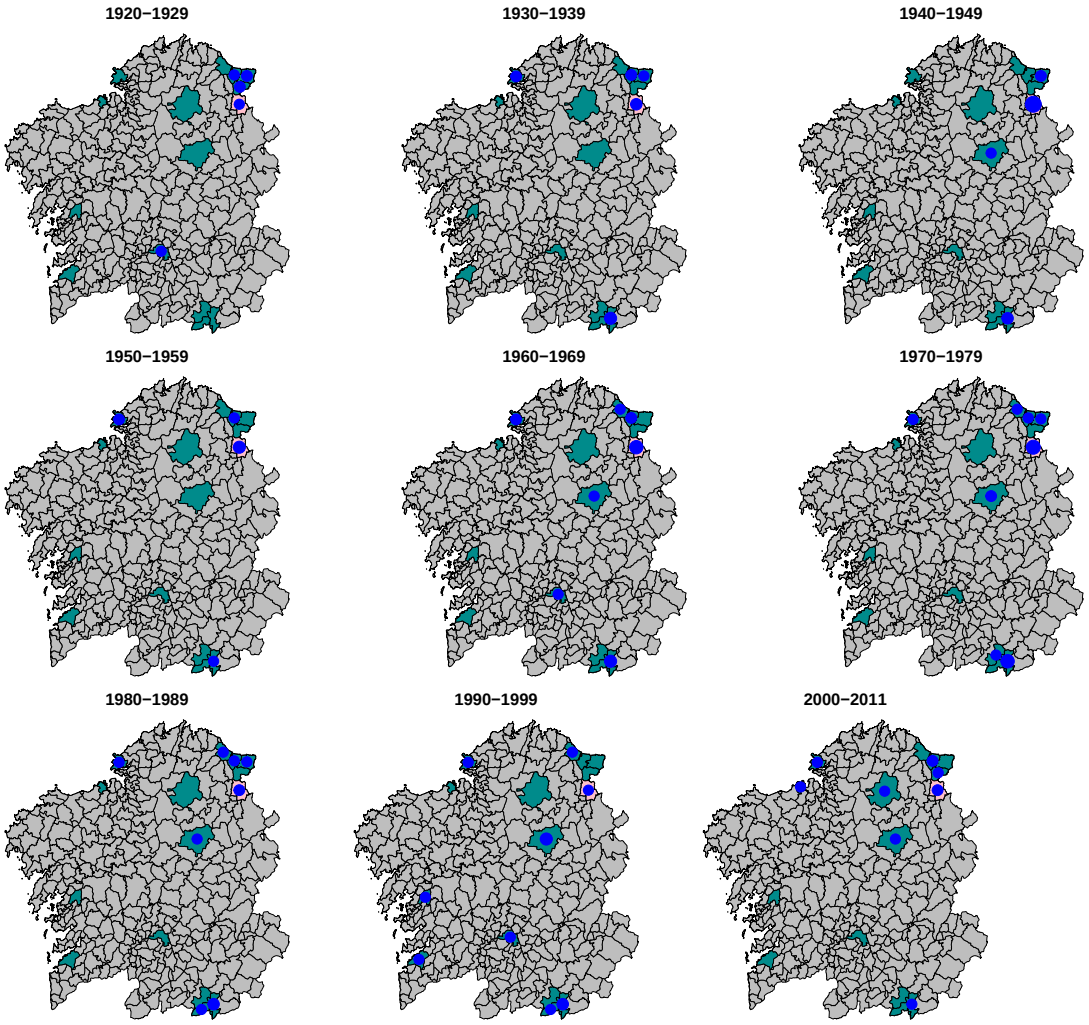


Figura 4: Mapas de distribución do apelido Ginzo cos nacidos dende 1920 ata 2011 en diferentes concellos de Galicia. Na cor gris se representan os concellos de Galicia onde non hai nacemendos de ningunha persoa co apelidos Ginzo. En cyan oscuro señálanse todolos concellos onde con xente que leva este apelido no grupo de idade correspondente. En cor rosa se representa o concello de A Pontenova, Xinzo é un lugar deste concello. Nos concellos que ademáis do fondo cyan oscuro hai puntos azuís, estes representan, que houbo xente con ese apelido nacida no intervalo correspondente.

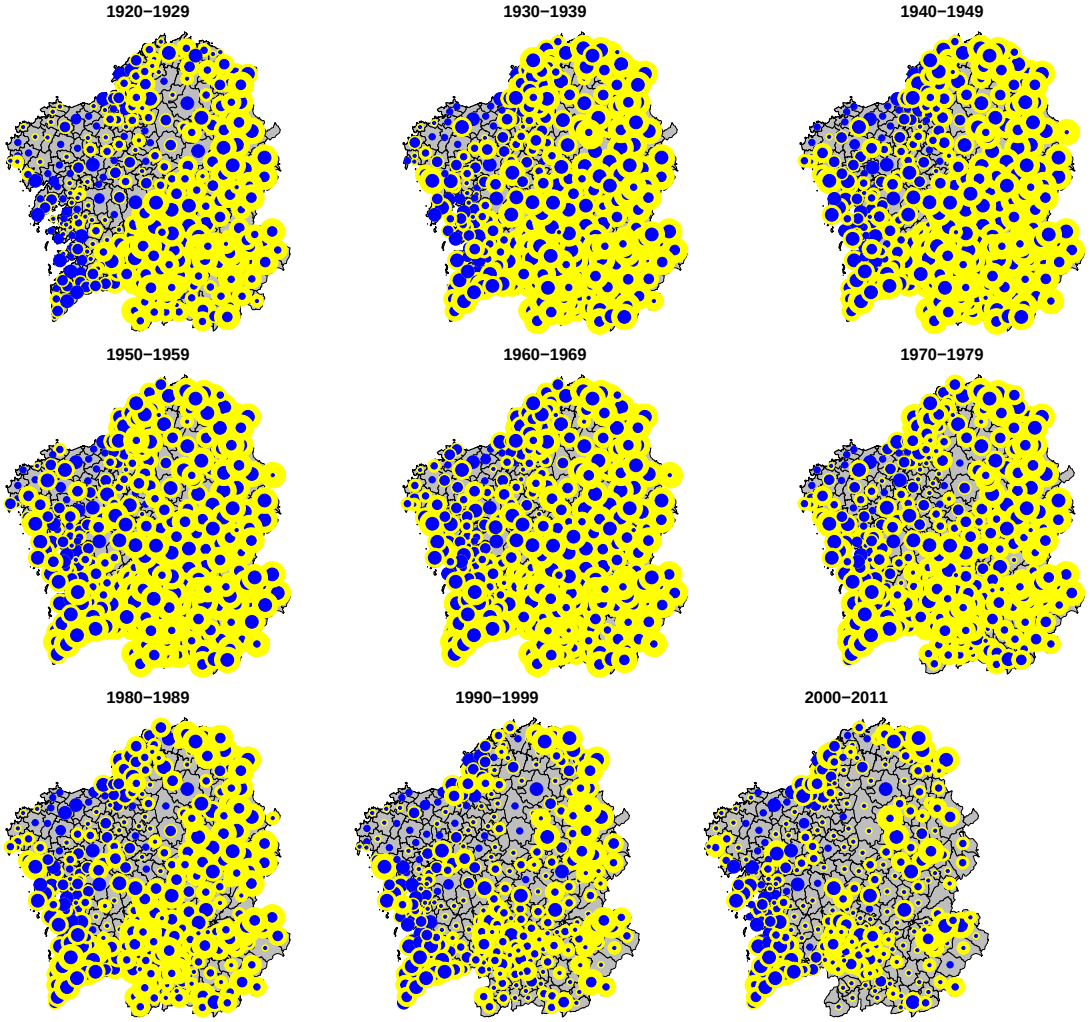


Figura 5: Mapas de distribución do apelido Fernández cos nacidos dende 1920 ata 2011 en diferentes concellos de Galicia. Na cor gris se representan os concellos de Galicia. Con puntos azuis representa o número de Fernández por concello e en cor amarelo o mesmo ponderado por poboación (dividido entre a poboación dos concellos no ano 2011).

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

UN NOVO INDICADOR DE DESEMPEÑO CON BASE NUNHA VELLA IDEA

Carlos L. Iglesias Patiño[1]

[1] Instituto Galego de Estatística

RESUMO

Este relatorio propón un novo indicador de avaliación da 'performance' global dun organismo, o seu grao de execución, inspirado nun vello resultado de Laplace e, mesmo, de Bayes. Define este grao de execución de Laplace-Bayes e estuda as súas características, en contra ou a prol, do punto de vista descritivo (poboación fixa) e tamén segundo un modelo sinxelo asumíbel. Remata con algúns resultados da súa aplicación e unhas breves conclusións.

Palabras e frases chave: indicador, superpoboación, mellora continua 'kaizen', encollemento 'shrinkage'

1. INTRODUCIÓN

Un dos problemas que observamos na planificación estatística galega é a propensión dos organismos a unha programación conservadora, isto é, a permanencia, no mellor dos casos, nun número de actividades cunha execución media aceptábel ou, no peor, nunha redución do número de actuacións para mellorar a súa execución media. É mais doado acadar unha execución plena (100 %) con poucas actividades que con moitas, no entanto, debe premiarse que se programen máis actuacións.

Habemos comprobar, unha vez máis, como a estatística pode ser un instrumento interesante na avaliación e no seguimento das políticas públicas. Neste caso concreto tratamos de calcular un indicador do desempeño dun órgano a partir das cadansúas execucións das actividades programadas por el. O indicador que se vén empregando para esta agregación no seguimento dos programas estatísticos anuais é a media aritmética simples que habemos denominar grao de execución orixinal (GEO).

O indicador alternativo que se propón é unha xeneralización do estatístico de Laplace. Primeiro habemos introducir este estatístico, despois habémolo xeneralizar para aplicalo a esta situación como indicador. Estudaranse as súas vantaxes e os seus inconvenientes e logo aplicarase a uns casos concretos. Tamén se ha estudar como estimador segundo un modelo de superpoboación.

Segundo Wilson (1927), a lei de sucesión de Laplace establece que se temos experimentado S éxitos e F fallas en $N = S + F$ experiencias "iguais", a posibilidade de éxito na experiencia $N + 1$ é $(S + 1)/(N + 2)$. Para un exemplo da súa aplicación ao nacer do sol pode consultarse Feller (1989).

A lei de sucesión de Laplace basease no teorema de Bayes. A técnica empregada no seu establecemento pode considerarse como precursora da actual estatística bayesiana onde a distribución a priori é a uniforme, non informativa.

2. DEFINICIÓN E PROPIEDADES DO NOVO INDICADOR

O estatístico de Laplace está a tratar con variábeis dicotómicas, 0 contra 1. Imos xeneralizalo permitindo que a variábel tome valores en todo o segmento unidade. As execucións son variábeis, en tanto por cento, que toman valores des do 0 %, cando non consta que se teña realizado ren, ata o 100 % cando está plenamente executada a actividade.

Isto é, teremos unha execución, en tanto por un, da actividade a, $X_a \in [0,1]$. O grao de execución de Laplace-Bayes (GELB) defínese como

$$\tilde{G}_E := \frac{\sum_{a=1}^{A} X_a + 1}{A + 2}(1 - \delta(A)),$$

onde A é o número de actividades programadas e $\delta(A) = 1$ se $A = 0$, outramente, $\delta(A) = 0$. Podemos interpretalo como que se engadiron, ás execucións do organismo, unha actividade sen executar e outra completamente executada de programar o organismo algunha actividade, se non, é nulo.

A relación entre o GELB e o GEO é a seguinte

$$\tilde{G}_E = \frac{\sum_1^A X_a + 1}{A+2} = \frac{\sum_1^A X_a}{A}\frac{A}{A+2} + \frac{1}{2}\frac{2}{A+2} = \bar{G}_E\frac{A}{A+2} + \frac{1}{2}\frac{2}{A+2}$$

Isto é, unha media ponderada entre o actual grao de execución e ½ cuns ponderadores que dependen do número de actividades programadas, A contra 2, o número de actividades reais fronte ás dúas ficticias.

Obsérvese que se un organismo realiza todas as actividades "a medias", o novo grao de execución para este organismo mediocre é $\tilde{G}_E = 1/2 = 50\ \%$. Polo tanto, se $A = 1, 2$, proporciona unha certa regresión á mediocridade; se $A > 2$, un paulatino achegamento ao GEO.

Para estudarmos como inflúe programaren unha nova actividade, decatémonos de que os graos de execución poden expresarse dos seguintes xeitos, dunha xeira

$$(A+1+2)\tilde{G}_{E,A+1} = (A+2)\,\tilde{G}_{E,A} + \tilde{G}_{E,A} + X_{A+1} - \tilde{G}_{E,A}$$

$$(A+1+2)\tilde{G}_{E,A+1} = (A+2+1)\tilde{G}_{E,A} + X_{A+1} - \tilde{G}_{E,A}$$

$$\tilde{G}_{E,A+1} = \tilde{G}_{E,A} + \frac{X_{A+1} - \tilde{G}_{E,A}}{(A+1)+2}$$

e mais para o GEO,

$$(A+1)\bar{G}_{E,A+1} = A\,\bar{G}_{E,A} + \bar{G}_{E,A} + X_{A+1} - \bar{G}_{E,A}$$

$$(A+1)\bar{G}_{E,A+1} = (A+1)\,\bar{G}_{E,A} + X_{A+1} - \bar{G}_{E,A}$$

$$\bar{G}_{E,A+1} = \bar{G}_{E,A} + \frac{X_{A+1} - \bar{G}_{E,A}}{A+1},$$

doutra xeira

$$\tilde{G}_{E,A+1} = \tilde{G}_{E,A}\frac{A+2}{A+3} + \frac{X_{A+1}}{A+3}$$

$$\bar{G}_{E,A+1} = \bar{G}_{E,A}\frac{A}{A+1} + \frac{X_{A+1}}{A+1}$$

O novo indicador resulta menos afectado por unha execución inadecuada dunha nova actividade programada porque $\frac{A}{A+1} < \frac{A+2}{A+3}$ e mais $A+3 > A+1$. Resiste á influencia desta actividade.

A seguir estudamos as eivas e as vantaxes. É precisa unha didáctica forte para que se comprenda que aínda co 100 % en cadansúa actividade o grao de execución asignado a cada organismo non sexa o 100 %, só tenda a el.

O GELB penaliza máis os que desenvolven menos actividades. Cando o organismo programa moitas actuacións, a distorción é moito menor con relación ao GEO. Se o organismo non programa ren, toma o valor cero co que se evita a actual diferenza entre execución nula (0) e non procede execución (denotada: —). No caso de que un organismo programe unha actividade ($A = 1$) que fracasar, $X_1 = 0$, o GELB non sería nulo senón $\tilde{G}_E = 1/3$. Isto había animar aos que comezan, que é mais doado que sexan mediocres ca cativos ou ruíns. E practicamente non inflúe nos moi desenvolveitos.

A fraqueza da suavización, tendencia ao ½, había se converter nunha oportunidade porque este valor pode servir como rente 'baseline' ou baliza 'benchmark'. Mais, se o órganismo programar dúas e só fixer a metade dunha, a estatística había valer $\tilde{G}_E$ =1'5/4 = 0'375; no caso de que só realizar a metade dunha de tres $\tilde{G}_E$ =1'5/5 = 0'3. Este comportamento conduciría a 0 asintoticamente. Para manterse en ½ habían realizalas todas a medias que non é un comportamento desexábel nun sistema estatístico que persiga o Código de conduta das estatísticas europeas como é o galego no encadramento do Plan galego de estatística (PGE) 2017-2021.

Para profundarmos no estudo deste estatístico, podemos introducir un modelo de superpoboación sinxelo como que $\mathrm{E}_m[X_a] = \gamma, \mathrm{D}_m^2[X_a] = \sigma^2$ onde γ é o grao de execución medio. Este modelo non é

restritivo porque as X_a son coutadas, están confinadas no segmento unidade ou no [0, 100] en percentaxe, e polo menos dentro dun organismo, a homoscedasticidade é asumíbel.

Segundo este modelo, a esperanza do GELB é

$$\mathrm{E}_m\left[\tilde{G}_E\right] = \mathrm{E}_m[\bar{G}_E]\frac{A}{A+2} + \frac{1}{2}\frac{2}{A+2} = \gamma\frac{A}{A+2} + \frac{1}{A+2}$$

e, xa que logo, o seu nesgo

$$\mathrm{B}_m\left[\tilde{G}_E\right] = \mathrm{E}_m[\bar{G}_E - \gamma] = \gamma\frac{A}{A+2} + \frac{1}{A+2} - \gamma = \frac{1-2\gamma}{A+2}$$

Se $\gamma < 1/2$, $\mathrm{B}_m\left[\tilde{G}_E\right] > 0$ sobrevalora o organismo, e se $\gamma > 1/2$, o infravalora como xa víramos enriba. No caso de que for $\gamma = 1/2$ é neutro. Cando o número de actividades, A, crece, o nesgo tende a 0.

Se alén diso, consideramos que as X_a son independentes, a variancia so modelo é

$$\mathrm{D}_m^2\left[\tilde{G}_E\right] = \mathrm{D}_m^2[\bar{G}_E]\left(\frac{A}{A+2}\right)^2 = \frac{\sigma^2}{A}\left(\frac{A}{A+2}\right)^2 = \sigma^2\frac{A}{(A+2)^2}$$

que resulta menor cá do GEO. Cando o número de actividades crece, esta variancia aproxímase á do GEO, son equivalentes asintoticamente.

O erro cadrático medio do GELB é

$$\mathrm{Q}_m^2\left[\tilde{G}_E\right] = \mathrm{D}_m^2\left[\tilde{G}_E\right] + \mathrm{B}_m^2\left[\tilde{G}_E\right] = \sigma^2\frac{A}{(A+2)^2} + \frac{(1-2\gamma)^2}{(A+2)^2}$$

que pode ser menor có do GEO

$$\mathrm{Q}_m^2\left[\tilde{G}_E\right] - \mathrm{Q}_m^2[\bar{G}_E] = \sigma^2\frac{A}{(A+2)^2} + \frac{(1-2\gamma)^2}{(A+2)^2} - \frac{\sigma^2}{A} =$$

$$= \frac{A^2\sigma^2 + A(1-2\gamma)^2 - \sigma^2(A+2)^2}{(A+2)^2A} =$$

$$= \frac{A^2\sigma^2 + A(1-2\gamma)^2 - \sigma^2A^2 - 4\sigma^2A - 4\sigma^2}{(A+2)^2A} =$$

$$= \frac{A(1-2\gamma)^2 - 4\sigma^2(A+1)}{(A+2)^2A}$$

Si $\gamma = 1/2$, $\mathrm{Q}_m^2\left[\tilde{G}_E\right] < \mathrm{Q}_m^2[\bar{G}_E]$. Se A é grande, $A \cong A+1$, o numerador é negativo cando $(1-2\gamma)^2 - 4\sigma^2 < 0$ isto é $(1-2\gamma) < 2\sigma$.

Para profundar na consideración dun modelo de superpoboación para o estudo máis analítico dunha poboación finita pode consultarse o conceito de superuniverso ou universo hipotético en Silva (1997). Des desta óptica, o GELB é un estimador afín, para o estudo destes estimadores lineares pode consultarse Lehmann&Casella (1998). Poden ser vistos como exemplos de encolla 'shrinking' do estimador orixinal: o constrinximento do tradicional, GEO, cara ao centro, ½. Tanto meirande é o tamaño do universo, canto máis pondera a estimativa empírica fronte á priorística.

3. RESULTADOS

Pasamos a amosar algún exemplos de aplicación do GELB. O primeiro refírese á evolución da execución do IGE e da Consellaría do Mar a par dos sucesivos programas do plano 2012-2016

Programa	Tipo de actividade	Exec.	2012	2013	2014	2015	2016
IGE	OO.EE. correntes	Nº	17	19	23	19	23
		GEO	98'2	93'2	90	100	98'7
		GELB	93'2	89'0	86,8	95'2	94'8
	AA.EE.	Nº	27	31	41	37	40
		GEO	98'3	99'7	90'2	94'9	98'4
		GELB	95	96'7	88'4	92'6	96'1
Consellaría do Mar	OO.EE. correntes	Nº	3	2	2	3	3
		GEO	99'3	100	100	83'3	90
		GELB	79'6	75'0	75'0	70'0	74'0
	AA.EE.	Nº	1	1	3	3	4
		GEO	100	100	100	100	75
		GELB	66'7	66'7	80	80	66'7

Táboa 1: Execución do IGE e da Consellaría do Mar no Plan 2012-2016

Escóllese esta consellaría por ser un dos departamentos de mellor desempeño. De feito, tiña unha execución mellor có do IGE, no entanto o número de actividades era reducido, a súa función principal non é realizar estatísticas como é a do IGE. No derradeiro programa programou máis actividades e houbo algúns problemas mais non repercuten excesivamente no GELB (unha redución de 4/5 a 2/3).

Os datos que se van a empregar a seguir non están publicados. Asignámoslles as actividades a cadanseu servizo coa información dispoñíbel no Servizo de Coordinación e Planificación do IGE e non se empregan os criterios do seguimento da programación de separar operacións das outras actividades estatísticas e, dentro das operacións, as novas das correntes, vulgo, tratamos as actividades a eito sen ponderármolas.

O segundo exemplo amosa a comparanza entre dous servizos do IGE no mesmo programa, 2016. O primeiro era responsábel de 27 actividades cun $\bar{G}_E = 95{,}9$ e o derradeiro servizo, 2 actividades cun $\bar{G}_E = 100$; mentres que cadansúas execucións serían $\tilde{G}_E = 92{,}7$ e $\tilde{G}_E = 75$ de empregarmos o GELB.

Para un outro servizo do IGE, como terceiro exemplo, a comparanza entre dous programas é a seguinte

Programa	**2013**	**2015**
Nº actividades	9	7
GEO	96,7	98,6
GELB	88,2	87,8

Táboa 2: Execución imputada dun servizo do IGE en dous programas do Plan 2012-2016

4. CONCLUSIÓNS

Este indicador consegue o obxectivo perseguido de fomentar a programación de novas actividades e racha o bucle de realimentación positiva da planificación acomodaticia. Permite unha comparanza diacrónica e sincrónica máis rápida e clara do desempeño 'performance' do organismo xa que non favorece a programación conservadora. Está máis na liña da mellora continua 'kaizen' porque non permite caer no triunfalismo das execucións 100 % e o único xeito de achegarse a elas é programar máis actividades.

No caso concreto da avaliación e seguimento do actual PGE 2017-2021, podería pensarse en obter o grao de execución global sen empregar ponderadores, a diferenza de como se viña facendo nos plans anteriores, xa que as actividades estatísticas están mellor demarcadas e xa non se programan os proxectos como operacións, isto implicaría que o colectivo ao que se referir o GELB había ser meirande e, polo tanto, veríase menos afectado pola súa única eiva remanecente.

Este novo grao de execución pode empregarse para a 'monitorización' doutras planificacións onde os axentes teñan certa liberdade para fixar, por antecipado, as súas metas e logo haxa que avaliar se as conqueriron. Ademais o marco de comparanza é como nas cualificacións académicas tradicionais: o ½, ou 50%, é o "aprobado" (½ en tanto por dez é 5).

REFERENCIAS

Feller, W. (1989) *Introducción a la teoría de probabilidades y sus aplicaciones*. Noriega editores LIMUSA, C. de México

IGE (des do 2012 ata o 2016) Programa estatístico Informe de seguimento http://www.ige.eu/web/mostrar_paxina.jsp?paxina=006018002002&idioma=gl

IGE Plan galego de estatística 2012-2016 (aplicación en R-Shiny a probar)

Lehmann, E.L. & Casella, G. (1998) *Theory of Point Estimation*, 2nd Ed. Springer Verlag, New York

Silva Ayçaguer, L.C. (1997) *Estadística e Investigación Científica en el Campo de la Salud: una mirada critica*. Ed. Díaz de Santos, Madrid

Wilson, E.B. (1927) «Probable inference, the Law of Succession, and Statistical Inference» *JASA*, 22, 158, pp. 209-212

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

CONCORDANCIA ENTRE AS CUALIFICACIÓNS DO ALUMNADO E DO PROFESOR NOS OBRADOIROS DE *MOODLE*

María del Carmen Iglesias Pérez[1] e Juan Vidal Puga[1]

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

RESUMO

O módulo Obradoiro da plataforma virtual de aprendizaxe *Moodle* permite ao alumnado enviar traballos e cualificar aos seus compañeiros/as, ademais da súa propia auto-avaliación. Neste traballo analizamos a concordancia entre as cualificacións realizadas polo alumnado e as cualificacións asignadas polo profesor nunha materia de primeiro curso do grao de Dirección e Xestión Pública da Universidade de Vigo.

Palabras e frases chave: Obradoiro Moodle, avaliación por pares, auto-avaliación, concordancia.

1. INTRODUCIÓN

A adaptación das materias de grado ao sistema europeo de educación superior ten motivado unha serie de cambios metodolóxicos que implican un aumento importante do tempo que o profesorado dedica á corrección de actividades. Incluso coa axuda dunha entorna de aprendizaxe virtual, o tempo medio de corrección de actividades pode estar entre 7 e 10 días (Cantabella, 2016).

A utilización das entornas virtuais permite a corrección por pares (é dicir, entre o propio alumnado). No caso de *Moodle*, esta característica está implementada no módulo Obradoiro. Esta actividade ten indubidables vantaxes no proceso de aprendizaxe, xa que permite a interacción no grupo e o traballo cooperativo, obriga ao alumnado a ser un suxeito crítico e a emitir unha cualificación do traballo dos seus pares, favorecendo a adquisición da capacidade crítica, e facendo ao alumnado máis autónomo e responsable, non só do traballo propio senn tamén do dos demais (Lara, 2009).

Neste traballo abordase a cuestión de ata que punto esta ferramenta pode liberar ao docente da tarefa de corrixir as actividades do alumnado. Con este fin, lévase a cabo unha análise de concordancia entre as cualificacións realizadas polo alumnado e as cualificacións asignadas polo profesor nos obradoiros de *Moodle* dunha materia de primeiro curso do grao de Dirección e Xestión Pública da Universidade de Vigo.

Contextualización

Estatística: Introdución á Estatística Administrativa é una materia obrigatoria de primeiro curso impartida no segundo semestre (de xaneiro a maio). Conta con dúas modalidades, unha presencial con docencia polas mañás de luns a venres e outra semipresencial coa maioría do traballo a distancia e unha reducida docencia presencial. A modalidade presencial ten unha carga lectiva de 42 horas ao longo do semestre, das que 30 son teóricas e 12 prácticas en grupos de tamaño máis reducido. Por outra parte, a modalidade semipresencial ten unha carga lectiva presencial de 12 horas e un uso intensivo da entorna virtual proporcionado por *Moodle*. En todo caso, ambas modalidades utilizan *Moodle*, e a maioría dos traballos propostos son comúns. Na modalidade semipresencial, estes traballos supoñen un 60% da cualificación final. Na modalidade presencial, a súa realización é voluntaria e supoñen ata 3 tres puntos adicionais en cada un dos dous exames parciais realizados durante o curso. Isto significa que, na modalidade presencial, os exercicios poden supoñer ata o 35% da cualificación final, se ben esta porcentaxe redúcese a medida que o alumnado obtén mellores resultados nos exames parciais.

Dentro de cada un destes traballos, o 80% da cualificación final depende do propio envío, mentres que o 20% restante obtense pola corrección dos outros traballos enviados, incluíndo a auto-avaliación. *Moodle* (versión 2.5) asigna automaticamente unha cualificación (auto-denominada *xusta*) que depende da concordancia da cualificación coas realizadas polo docente e os pares. A cada envío se lle asignan catro traballos para avaliar: tres de pares e unha auto-avaliación. Co obxecto de que a corrección do docente teña prioridade, se lle asigna un peso máximo (que é 16 no caso da versión utilizada).

O alumnado foi instruido de forma oral durante as primeiras sesións presenciais de que a cualificación do 20% correspondente ás súas correccións sería feita de forma automática polo sistema en base á concordancia ca avaliación feita polo propio docente. O alumnado dispón dunha rúbrica de avaliación para realizar as correccións de cada obradoiro e posterior acceso á avaliación numérica das súas correccións unha vez o obradoiro está pechado. O sistema non proporciona máis detalles.

2. MEDIDAS DE CONCORDANCIA

O estudo da concordancia entre variables é necesario para evaluar a aceptabilidade dun novo proceso ou metodoloxía, a validación de instrumentos ou ensaios, comparación de métodos, control estatístico de procesos, etc. (Lin et al., 2002). As técnicas de análise da concordancia dependen do tipo de variable a estudar (Pita y Pértega, 2004). No caso de variables cualitativas a técnica máis utilizada é o coeficiente kappa de Cohen. Cando as variables son cuantitativas, unha práctica habitual é utilizar o coeficiente de correlación lineal de Pearson ou a proba t de comparación de medias. Aínda que a concordancia entre dúas variables implica a existencia de correlación significativa e de media cero na diferenza entre as mesmas, estas condicións non son suficientes para determinar a existencia de concordancia. Existen técnicas estatísticas máis axeitadas, entre as que destacan o coeficiente de correlación intraclase, CCI (Fleiss, 1986), o coeficiente de concordancia de Lin, CCC (Lin et al., 2002), e tamén gráficas como a de Bland e Altman (Bland e Altman, 1986) ou a representación da discordancia acumulada (con forma similar a unha curva de supervivencia) proposta por Luiz et al. (2003).

Coeficiente de correlación intraclase, CCI

É un dos coeficientes máis recomendados para medir a concordancia entre variables numéricas. O seu cálculo basease nun modelo ANOVA de medidas repetidas, onde se realiza una descomposición da variabilidade na imputable aos individuos, ao método de medición y á variabilidade aleatoria. O CCI definese como a proporción da variabilidade total imputable aos individuos. Un aspecto a ter en conta é que a fórmula do CCI depende do modelo ANOVA considerado: dous factores con efectos mixtos, dous factores aleatorios ou un factor aleatorio. Neste traballo considerase o caso dun ANOVA con efectos mixtos, sendo a fórmula do coeficiente a seguinte

$$CCI = \frac{k \cdot SS_{entre} - SS_{total}}{(k-1)SS_{total}},$$

onde n é o número de individuos, k o número de métodos en estudo, $SS_{total} = \sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij} - \bar{X}_{..})^2$, $SS_{entre} = k\sum_{i=1}^{n}(\bar{X}_{i\cdot} - \bar{X}_{..})^2$ e X_{ij} é a medida do método j no individuo i.

A Táboa 1 presenta a valoración da concordancia segundo os valores do CCI (Fleiss, 1986).

Valor do CCI	Concordancia
≥ 0.90	Moi boa
0.71 - 0.90	Boa
0.51 - 0.70	Moderada
0.31 - 0.50	Mediocre
< 0.30	Mala

Táboa 1: Valoración da concordancia segundo o coeficiente de correlación intraclase.

Coeficiente de concordancia de Lin, CCC

Este coeficiente defínese a partires dunha estandarización do valor esperado do cadrado da diferenza entre dous métodos de medición, é dicir

$$\epsilon^2 = E(X-Y)^2 = (\mu_Y - \mu_X)^2 + \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y\rho$$

$$\rho_C = 1 - \frac{\epsilon^2}{\epsilon^2/\rho = 0} = \frac{2\sigma_X\sigma_Y}{(\mu_Y - \mu_X)^2 + \sigma_Y^2 + \sigma_X^2} \cdot \rho$$

O coeficiente de concordancia de Lin, CCC, é a versión mostral do coeficiente ρ_C.

Ten interese observar que ρ_C =exactitude $*$ precisión= $\frac{2\sigma_X\sigma_Y}{(\mu_Y-\mu_X)^2+\sigma_Y^2+\sigma_X^2} * \rho$, medindo a exactitude a cercanía entre as distribucións marxináis dos métodos X e Y, e medindo a precisión o grao de variabilidade conxunto. O coeficiente de Lin pode adaptarse ao caso de $k > 2$ métodos.

Asemade, Lin recomenda acompañar o seu coeficiente cunha gráfica de dispersión de Y en función de X que mostre a recta diagonal y=x.

Método gráfico de Bland e Altman

Este método compara gráficamente a concordancia entre dúas medidas. Consiste nun gráfico de dispersión que representa a diferenza das dúas medidas fronte ao promedio das mesmas. Ademais, inclúe tres liñas horizontales: na diferenza media, e a dúas desviacións típicas da diferenza media (a desviación típica é da variable diferenza). Así, obsérvase a magnitude da diferenza, a súa variabilidade e o comportamento de dita diferenza respecto ao promedio das medidas.

Método gráfico de Luiz et al.

Ao igual que o método gráfico de Bland e Altman é un gráfico diseñado para a comparación de dúas medidas. Calcula a diferenza entre as dúas medidas en términos absolutos, e a porcentaxe acumulada de casos nos que se supera cada unha das diferenzas absolutas. A continuación, faise a representación da porcentaxe acumulada de discordancia en función da diferenza absoluta, dando lugar a unha gráfica cunha forma similar a unha función de supervivencia. O gráfico mostra de forma sinxela a porcentaxe de individuos que superan unha diferenza determinada, sendo moi útil desde o punto de vista práctico.

3. CONCORDANCIA DOS TRABALLOS

Unha vez finalizado o semestre, completouse a realización e corrección dun total de catorce/quince obradoiros (modalidade presencial/semipresencial) cun total de 256 traballos enviados e 653 correccións (2.6 correccións por envío) das cales 158 foron auto-correccións e 495 correccións por outros pares (1.9 correccións por envío).

Todo o alumnado que realizaba un envío tiña a posibilidade de cualificarse a se mesmo/a e a tres compañeiros/as. Dos 256 envíos, todos menos 13 (o 95%) realizaron algunha corrección, sendo a opción de "calificou e se calificou" a maioritaria (61%) entre os envíos corrixidos (Figura 1).

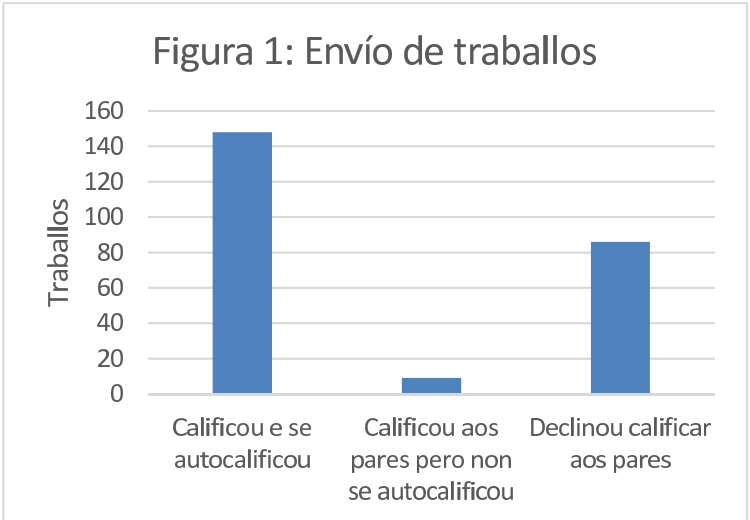


Figura 1: Envío de traballos.

A correlación entre as cualificacións asignadas polo profesor en cada unha das correccións e as realizadas por outros pares ($n = 215$) toma o valor $r = 0.84$, mentres que utilizando só as autocorreccións ($n = 145$) acádase unha correlación $r = 0.72$. Na Figura 2 pódese ver graficamente a relación entre as cualificacións do alumnado e do profesor separada en dous grupos: outros pares e autocorreccións. A bisectriz indica os puntos de concordancia perfecta.

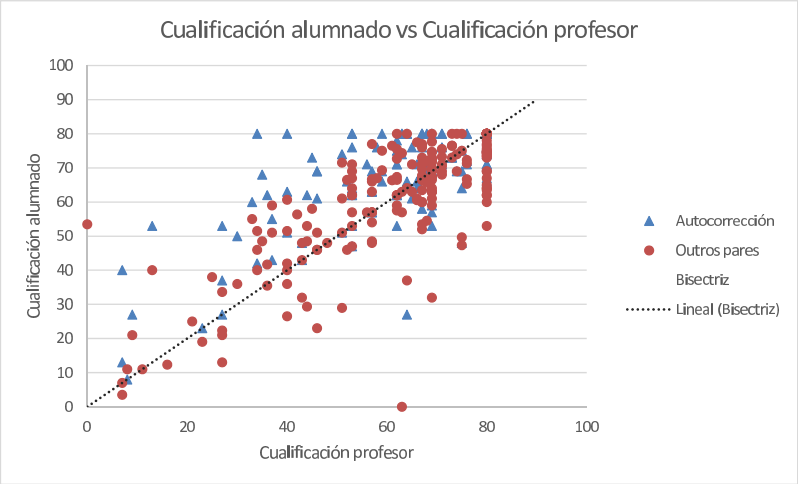


Figura 2: Relación das cualificacións do alumnado e do profesor por grupos: outros pares e autocorreccións.

Respecto á variable diferenza entre a cualificación asignada polo alumnado e polo profesor nas correccións por outros pares, non atopamos diferenzas significativas nas medias (Figura 3). Sen embargo, si que atopamos diferenzas significativas entre as autocorreccións e a cualificación do profesor. Nas autocorreccións o alumnado tende a cualificarse por enriba da cualificación do profesorado, cunha diferenza media de 5.88 puntos (sobre 80).(Figura 3).

	Proba t de mostras emparelladas						
	Diferencias emparelladas						
			95% intervalo de confianza da diferencia				
	Media	Desviación estándar	Inferior	Superior	t	gl	Sig. (bilateral)
Autoavaliación - Nota profesor	5,8828	12,2216	3,8766	7,8889	5,796	144	,000
Nota compañeiros - Nota profesor	,1047	9,6928	-1,1983	1,4076	,158	214	,874

Figura 3: Resultados da proba t de mostras emparelladas para o contraste de igualdade de medias.

A maiores do estudo preliminar de correlación e comparación de medias fixemos un estudo de concordancia entre os métodos de cualificación. A concordancia entre as autocorreccións e a cualificación do profesorado ten valores $CCI = 0.66$ e $CCC = 0.64$, indicando unha concordancia moderada (Fleiss 1986). A concordancia entre as correccións por pares e a cualificación do profesorado ten valores $CCI = 0.84$ e $CCC = 0.83$, é dicir, boa concordancia.

4. CONCORDANCIA DAS NOTAS FINAIS

Para saber ata que punto a ferramenta dos obradoiros pode liberar ao docente da tarefa de corrixir os traballos do alumnado, levamos a cabo un estudo de concordancia entre a nota final acadada polo alumnado facendo a media dos traballos corrixidos polo profesor, e a nota final

resultado da media dos traballos corrixidos polos seus pares. Estas cualificacións aparecen aquí reescaladas ao formato habitual de 0 a 10 puntos.

Acadamos unha concordancia moi boa: $CCI = 0.993$ e $CCC = 0.934$, a maiores dunha correlación moi alta, $r = 0.994$, e unha diferenza de medias entre ámbolos dous métodos non significativa: 0.07 ± 0.35 ($p = 0.24$).

Tamén estudamos o efecto, sobre a variable diferenza, da modalidade (presencial/semipresencial) e do sexo, controlando pola nota do profesorado e o número de exercicios presentados polo alumnado. Atopamos que o sexo ten un efecto significativo ($p = 0.023$), estando a diferenza media dos homes case unha décima por enriba da media das mulleres. A modalidade non tivo efecto significativo. O efecto do sexo pode verse na Figura 4, sendo máis acentuado en cualificacións altas.

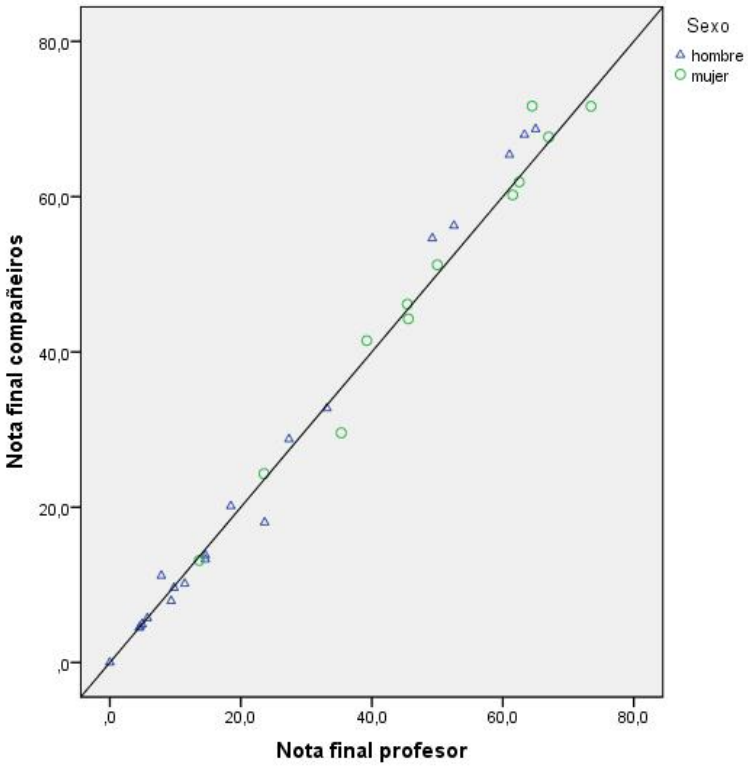


Figura 4: Relación das notas finais dos pares e do profesor distinguindo por sexo.

As Figuras 5 e 6 indagan na distribución das diferenzas. Do gráfico de Bland e Altman (Figura 5) deducimos que o 95.5% das diferenzas entre a cualificación final do alumnado e do profesorado van a estar entre 0.07 ± 0.7, é dicir entre -0.63 e 0.77 puntos. Menos de un punto de diferenza, pero pode chegar a máis de medio punto. Se pensamos que unha diferenza de ata medio punto pode ser admisible, na Figura 6 vemos que só haberá un 19% de notas cunha diferenza maior.

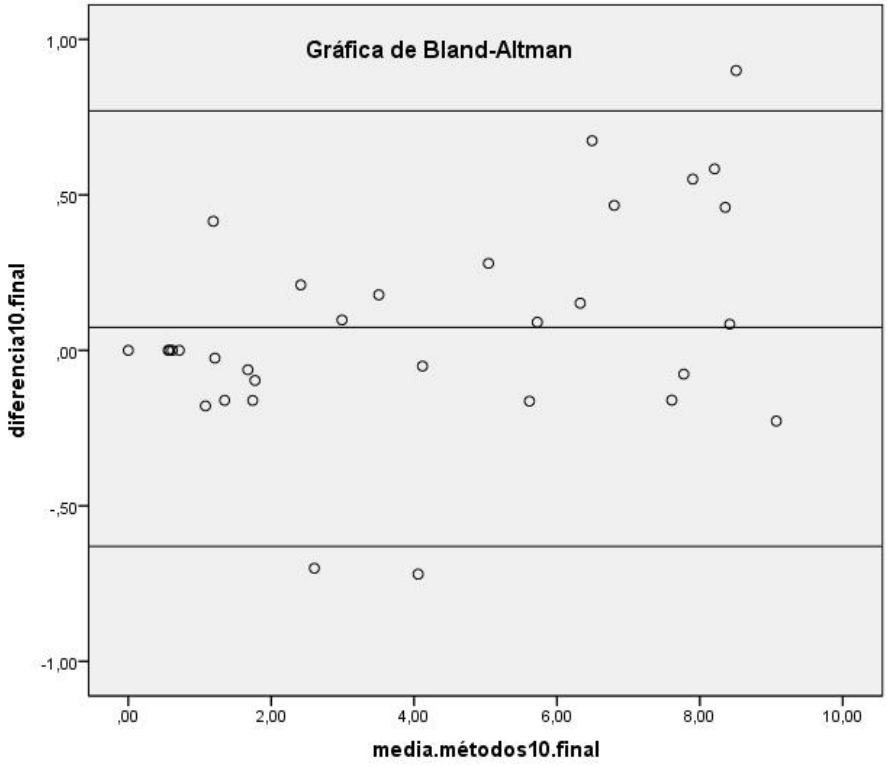


Figura 5: Gráfico de Bland e Altman: distribución da diferenza fronte á media dos métodos de cualificación

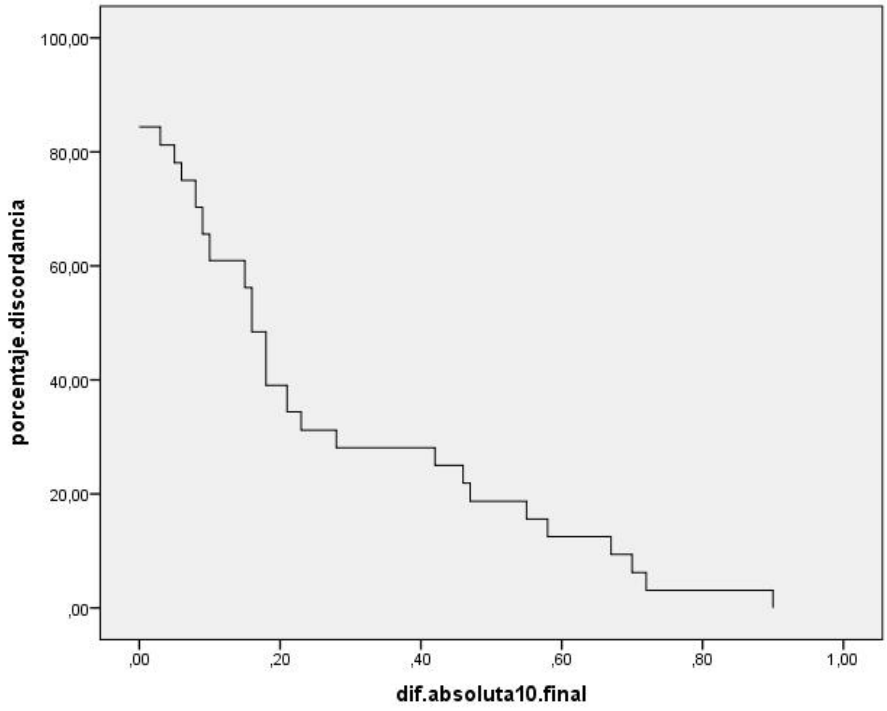


Figura 6: Complementario da distribución acumulada $(1 - F)$ das diferenzas absolutas.

5. CONCLUSIÓNS

Os principais resultados da análise realizada son:

- A concordancia entre as cualificacións do profesorado e das autoavaliacións é moderada, mentras que a concordancia da cualificación do profesorado e dos pares é boa.
- A concordancia entre a cualificación final media do profesorado e dos pares é moi boa, sendo a diferenza sempre menor que 1 punto, e no 81% dos casos menor que medio punto.

Parece polo tanto, que o sistema de corrección por pares podería ser unha boa ferramenta na avaliación continua do alumnado.

Por outra banda cabe salientar que na análise dos datos atopamos algúns valores atípicos. Estes valores salientan a importancia de construir unha boa rúbrica de avaliación, para evitar unha diferenza excesiva na corrección. Tamén a conveniencia de revisar a corrección dos traballos enviados que sexan avaliados cun cero.

REFERENCIAS

Bland. J. M., and Altman, D. (1986) Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. Lancet, 8, 307-310.

Cantabella, M., López-Ayuso, B., Muñoz, A. y Caballero, A. (2016) Una herramienta para el seguimiento del profesorado universitario en Entornos Virtuales de Aprendizaje. Revista Española de Documentación Científica, 39 (4), e153, doi:10.3989/redc.2016.4.1354

Fleiss, J.L. (1986) The design and analysis of clinical experiments. New York, Wiley.

Lara Fuillerat, J.M. (2009) *Moodle*. Manual de referencia para el profesorado. Documentación libre de GNU. En `http://pefc5.ugr.es/moodle/file.php/1/Lara-ManualMoodleProfesores-2009.pdf`

Lin, L., Hedayat, A.S., Sinha, B. and Yang, M. (2002) Statistical Methods in Assessing Agreement. Journal of the American Statistical Association, 97(457), 257-270.

Luiz, R.R., Costa, J.L., Kale, P.L. and Werneck, G.L. (2003) Assessment of agreement of a quantitative variable: a new graphical approach. Journal of Clinical Epidemiology, 56(10), 963-967.

Pita Fernández, S. y Pértegas Díaz, S. (2004) La fiabilidad de las mediciones clínicas: el análisis de concordancia para variables numéricas. En `http://www.fisterra.com/mbe/investiga/conc_numerica/conc_numerica.pdf`

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Goodness-of-fit tests for disorder detection in NGS experiments

Norman Jiménez Otero[1,2] and Jacobo de Uña Álvarez[1,2,3]

[1]Biomedical Research Centre (CINBIO), University of Vigo
[2]Statistical Inference, Decision and Operations Research (SiDOR) group, University of Vigo
[3]Department of Statistics and Operations Research, University of Vigo

ABSTRACT

Next-generation sequencing experiments (NGS) are often performed in biomedical research nowadays, leading to methodological challenges related to the high-dimensional and complex nature of the recorded data (Lander and Waterman, 1988). In this work we review some of the issues which arise in disorder detection from NGS experiments, that is, when the focus is the disorder detection for homozygosity and heterozygosity in DNA sequencing. A statistical model to cope with guanine/cytosine (GC) biases and phasing and prephasing phenomenon is proposed, and a goodness-of-fit procedure for disorder detection is derived. The method combines the proper evaluation of local p-values (one for each DNA base) with suitable corrections for multiple comparisons and the discrete nature of the p-values. The performance of the introduced procedures is investigated through simulations.

Keywords: DNA, GC-bias, multiple hypothesis testing, Poisson distribution, sequencing.

1. INTRODUCTION

The main contribution of this piece of research is to develop a theoretical framework for the detection of genetic disorders in single-end next-generation sequencing (NGS) experiments. In these experiments, the number of reads which overlap a given region of DNA bases is recorded. Two random counts are of interest: the number of reads which start at each DNA base; and the base coverage, which is defined as the number of reads which cover a given base. Following Lander and Waterman (1988), we assume that the number of reads which start at each base follows a Poisson distribution. From this basic distributional assumption and some independence conditions, we derive the distribution of the coverage, which is Poisson too. Our theoretical framework takes into account some existing biases in NGS experiments, like those arising from the random read length, the random fragment length, and the guanine/cytosine (GC) bias. All these biases affect the parameter of the Poisson distribution for the base coverage. Base-level p-values corresponding to the Neyman-Pearson hypothesis test for disorder detection are calculated. In order to take the multiple comparisons into account, we use the Benjamini-Hochberg (BH) false discovery rate (FDR) controlling procedure (Benjamini and Hochberg, 1995), and the generalization of the BH method for discrete p-values proposed by Heyse (2011). Although the multiple tests along the bases of the region are not independent, the BH method is valid because the tests corresponding to different bases are positively dependent (Benjamini and Yekutieli, 2001). Finally we investigate the performance of the proposed testing procedures through simulations.

2. PARAMETRIC MODEL FOR THE COVERAGE

We consider a single-end NGS experiment in which the target is a region of f DNA bases. The sequencing is performed through a number of fragments, each of which starts at a certain base of the region. Each fragment generates a read which may or may not contribute to the sampling information, as it will be explained now. The length of the fragments is a random variable L

supported on the positive integers $\{1, ..., k\}$, and we put $q_l = P(L = l)$, $1 \leq l \leq k$, for the probability mass of L. A fragment of length l starting at base j entails a GC-content GC_j^l, which is defined as the proportion of guanine and cytosine nucleotides along the fragment. This GC-content results in a sampling bias, in the sense that the probability for the fragment to stay in the sample is $\nu_j^l = \nu(GC_j^l)$, where ν is a smooth function, often a concave function reaching its maximum around 1/2. If the fragment remains in the sample, a read with random length R ($R \leq L$) is extracted from the fragment; this read contributes to the coverage if it completely overlaps the region of DNA bases. Therefore, any suitable model for the base coverage should account for two facts: (a) the probability for a fragment to stay in the sample depends on its length, and on the base at which it starts, through its GC-content; and (b) for a fragment contributing to the sample, the probability for the corresponding read to cover a given base depends on the distribution of R, the base at which the read starts, and the specific base for which the coverage is calculated.

As mentioned, following Lander and Waterman (1988) we assume that the number of fragments starting at base j, N_j say, follows a $P(\lambda)$ distribution (a Poisson distribution with rate λ). We also assume that N_j, $1 \leq j \leq f$ are independent random counts. Given N_j, the number of fragments staying in the sample follows a Binomial distribution, $N_j^{GC} \sim Bin(N_j, \nu_j)$, where ν_j accounts for the possible GC-bias. Actually, $\nu_j = \sum_{l=1}^{k} \nu_j^l q_l$. Under these assumptions one can easily get that, inconditionally, $N_j^{GC} \sim P(\lambda\nu_j)$. Finally, because of the independence among the base counts and the reproductive property of the Poisson distribution, it holds that the total number of fragments $N^{GC} = \sum_{j=1}^{f} N_j^{GC}$ follows a $P(\lambda \sum_{j=1}^{f} \nu_j)$ distribution. For the best of our knowledge, these model properties have not been exploited in the literature related to NGS despite of being easily obtained from primitive assumptions. See Sathirapongsasuti et al. (2011) and Li et al. (2012) for an introduction to this topic.

It has been noted above that each fragment in the NGS experiment gives rise to a read of random length R. The coverage of a given base b is defined as the number of reads among the total count N^{GC} which cover base b while completely overlapping the region. For each of the N_j^{GC} reads starting at a base j with $j \leq b$, the probability of such an event is given by $p_{b,j} = P(b-j \leq R \leq f-j+1)$, where the inequality $R \leq f - j + 1$ indicates that reads going beyond the last base of the region (f) will be discarded. When $j > b$, such a probability $p_{b,j}$ is zero. From this, and using again the independence among the base counts and the reproductive property of the Poisson distribution, it follows that the coverage of base b, C_b say, follows a $P(\lambda \sum_{j=1}^{b} \nu_j p_{b,j})$ distribution. Clearly, coverages corresponding to different bases will be statistically dependent.

3. P-VALUE CALCULATION AND FDR-CONTROLLING PROCEDURE

In Section 2 we have derived a Poisson model with rate parameter $\Lambda_0 = \lambda \sum_{j=1}^{b} \nu_j p_{b,j}$ for the coverage of the base b, C_b, in absence of disorders. It is assumed that the parameters λ, ν_j and $p_{b,j}$ are known or can be calculated with negligible noise. In this Section we introduce a local test for disorder detection, where 'local' refers to the given base b, $1 \leq b \leq f$. The null hypothesis $H_0 : \Lambda = \Lambda_0$ is tested against an alternative hypothesis $H_1 : \Lambda \neq \Lambda_0$. For the simplest disorders (deletions), the Λ parameter in the alternative hypothesis is reduced to a few values, here denoted by Λ_1, which depend on the specific disorder to be detected. Specifically, $\Lambda_1 = 0$ for the detection of homozygotic deletion; $\Lambda_1 = \Lambda_0/2$ for heterozygotic deletion. The decision is made from a single observation of the random variable C_b, which follows a $P(\Lambda_0)$ distribution under the null hypothesis. According to the Neyman-Pearson test, the null hypothesis is rejected for sufficiently small values of C_b when $\Lambda_1 < \Lambda_0$. Thus, the p-value of the observed coverage at base b, c_b say, is calculated as $P_{H_0}(C_b \leq c_b)$ for deletion disorder.

The aforementioned calculation reports as many p-values as bases in the region, that is, f p-values. In order to take the multiple comparisons for these f bases into account, we propose to use the BH FDR-controlling procedure (Benjamini and Hochberg, 1995). For a given level α, this procedure guarantees that the expected amount of disorders falsely detected among the rejected null hypotheses remains bounded by α. Despite of the existing dependence among the f coverages,

the BH method is valid due to the positive dependence among the test statistics (Benjamini and Yekutieli, 2001). In order to improve the statistical power, we do suggest to apply a correction for the discreteness of the p-values. Heyse (2011) extended the BH method for discrete p-values. There are other possible approaches, but in our preliminary simulation experiments Heyse's correction reported consistently the largest power among them, while preserving the FDR control. Therefore, this is the method we consider in the next Section. Heyse's correction is implemented in the `R` package *discreteMTP*, specifically in the function `p.discrete.adjust(p, pCDF, method = "DBH", n = length(p))`.

4. SIMULATION STUDY

In Figure 1, we depict the coverage for a simulated region of $f = 300$ DNA bases, with an experimental parameter of $\lambda = 1$. The 'case' coverage (green line) corresponds to a region with a heterozygotic deletion which affects to one base (base 150). The expected values Λ_0 (red line) were calculated by using the model introduced in Section 2. The fragment lengths and the read lengths are random variables degenerated at 78 and 100 respectively (constant read length and frangment length scenario). The existing deletion in base 150 results in a coverage lower than expected value. The potential GC-content bias is driven by the smooth parabolic curve $\nu(x) = 4(x - x^2)$, which is symmetrical around 1/2. The expected GC-content for the simulated region (a sequence of A, T, C and G nucleotides with size 300) was 1/2; since $\nu(1/2) = 1$, this means that there is almost no GC-bias in the simulated experiment.

In Figures 2 and 3 the adjusted p-values corresponding to the BH method and Heyse's correction for discreteness, respectively, are displayed. In both Figures, the only adjusted p-value falling below the $\alpha = 0.05$ threshold is that corresponding to base 150. Thus, if the researcher aims to control the expected FDR at 5% level, he will only declared a deletion at base 150, leading to a 100% of statistical power and a 0% of proportion of false discoveries (or false positives) in this case. Of course, the situation may be different for other simulated trials. In order to compare BH and Heyse's adjusted p-values, in Figure 4 we report the corresponding scatterplot (left panel) with a zoom in for the $[0, 0.05] \times [0, 0.05]$ rectangle. It is seen that Heyse's correction for discrete p-values leads to smaller adjusted p-values and, therefore, may increase the statistical power of the multi-test.

More extensive simulations were performed. More specifically, we considered scenarios with no disorder, and with homozygotic and heterozygotic deletion in a single base but with different locations inside the region: 25, 78, 150, 222 and 275, aiming to include deletions both at extreme and intermediate locations. The experimental parameter was taken as $\lambda = 1$ or $\lambda = 2$, corresponding to a smaller or larger amount of fragments (reads) in the NGS experiment. We considered both the constant and the random fragment/read length scenarios. For the random length scenario, we generated the read length distribution and the fragment length distribution following a (rounded) normal distribution with mean equal to a maximum possible length $R_{max} = 78$ and $L_{max} = 100$ respectively, and standard deviation $sd = 40$; then, the values of R (resp. L) out of the interval $[65, 78]$ (resp. $[78, 100]$) were set equal to R_{max} (resp. L_{max}). This results in a random length distribution largely concentrated at a given maximum value, with some dispersion towards smaller values, which is a realistic model in NGS experiments. We explored four GC-content bias scenarios, by dividing the region on $f = 300$ bases in three subregions of 100 bases each, and then by setting the expected GC-content for each subregion equal to some given proportions: 50%-50%-50% (no bias), 25%-50%-75%, 5%-25%-50%, and GC-content 50%-5%-25%. The GC-content bias curve used in the simulation was always $\nu(x) = 4(x - x^2)$. The results (not shown) suggest that the proposed testing procedure respects the nominal level α when there is no disorder, while it reaches a reasonable statistical power when disorders are present and the experimental λ parameter is large enough. An interesting conclusion from the simulation results is that the power of the BH FDR-controlling procedure was improved by using the correction for discrete p-values.

For more detailed illustration and discussion, in Table 1 we report the average FDR along 5000 Monte Carlo replicates for the scenario with no disorder, for the four different settings of GC-content. An experimental parameter of $\lambda = 1$, a nominal FDR of 5%, and Heyse's correction for discrete p-values were used (this applies throughout the rest of this Section). From this Table 1 it is seen that the proposed method is conservative, with a FDR ranging between 2.1% and 2.9% depending on the situation. The power and FDR attained in a setting with an heterocygotic deletion which affects to a single base are reported in Tables 2 and 3 respectively. Five different locations of the deleted base were considered. In these Tables 2 and 3, the fragment length and the read length were randomly generated. It is seen that the statistical power (defined as the proportion of trials from which the method was able to detect the deleted base) is generally large, with some exceptions when for extreme base locations or GC-contents (as expected). The power was improved when $\lambda = 2$ (results not shown). Regarding the FDR, it is seen again that the proposed procedure respects the nominal level in all the scenarios. Finally, in Table 4 we report the proportion of trials for which the method declared a deletion disorder, including the trials in which the base with a real deletion remained undetected. As expected, figures in Table 4 are close to the statistical power in Table 2.

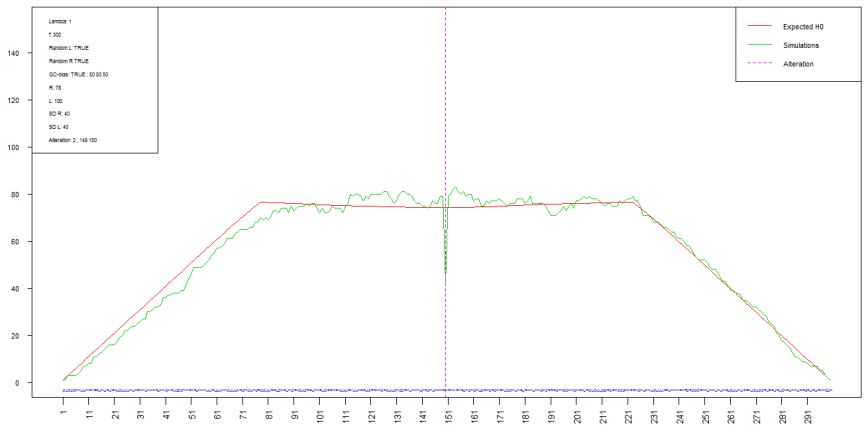


Figure 1: Coverage (green line) for a simulated region of $f = 300$ DNA bases with a heterozygotic deletion which affects to base 150 (vertical line). Expected values under the null hypothesis Λ_0 are depicted (red line). Blue letters correspond to the sequence of the region (arround 50% of G and C).

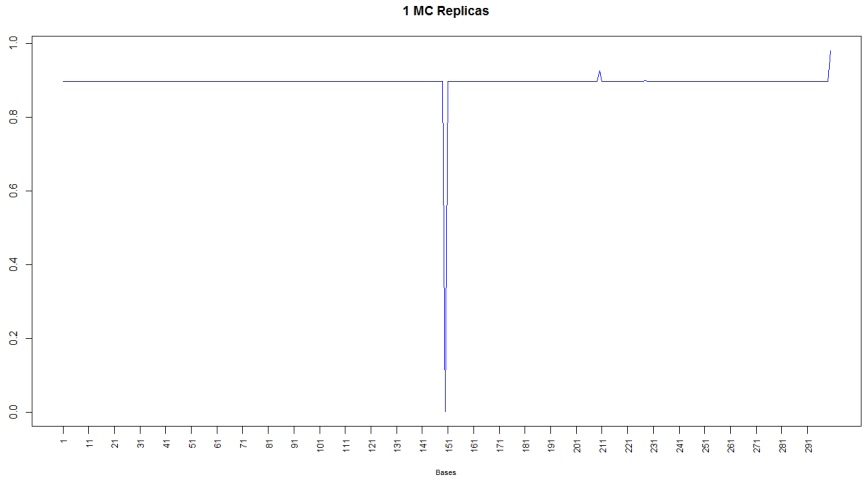


Figure 2: BH adjusted p-values for the detection of deletion computed from the coverages in Figure 1.

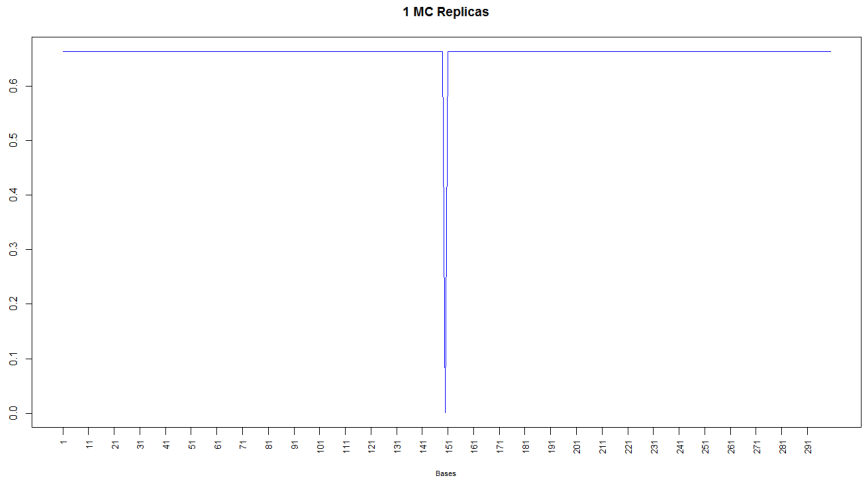


Figure 3: Heyse's adjusted p-values for the detection of deletion computed from the coverages in Figure 1.

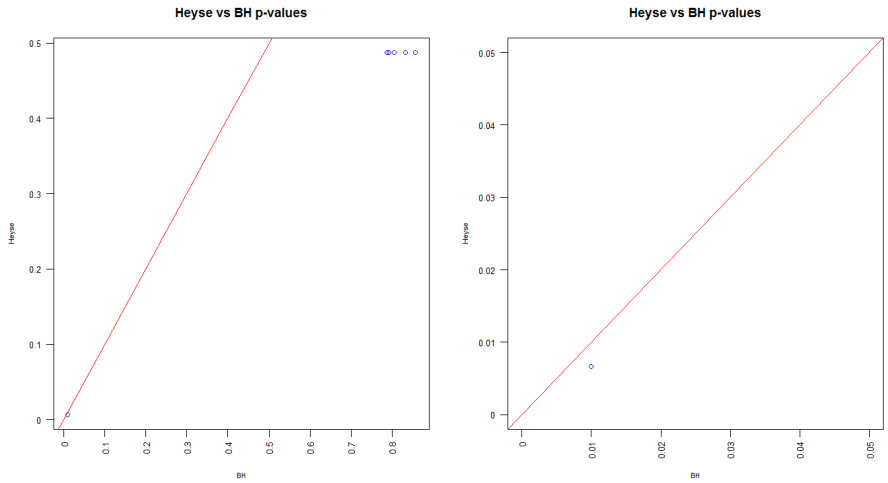


Figure 4: Left: Heyse vs. Benjamini-Hochberg adjusted p-values for the detection of deletion computed from the coverages in Figure 1. Right: Zoom in for the $[0, 0.05] \times [0, 0.05]$ rectangle

Length/GC	50/50/50	25/50/75	5/25/50	50/5/25
Constant	0.0254	0.021	0.026	0.025
Random	0.0252	0.0228	0.0228	0.0286

Table 1: Average FDR along 5000 Monte Carlo replicates. No deletion (see text for details).

Base/GC	50/50/50	25/50/75	5/25/50	50/5/25
25	0.2118	0.0926	0.003	0.1436
78	0.9138	0.8818	0.2916	0.8262
150	0.9286	0.9188	0.8578	0.243
222	0.923	0.8548	0.9376	0.4764
275	0.1852	0.1244	0.1868	0.0628

Table 2: Average power along 5000 Monte Carlo replicates. Heterocygotic deletion at a single base with five different locations (see text for details).

Base/GC	50/50/50	25/50/75	5/25/50	50/5/25
25	0.0251	0.0224	0.0233	0.0247
78	0.0220	0.0269	0.0229	0.0264
150	0.0245	0.0218	0.0195	0.0227
222	0.0233	0.0293	0.0246	0.0230
275	0.0248	0.0234	0.0275	0.0255

Table 3: Average FDR along 5000 Monte Carlo replicates. Heterocygotic deletion at a single base with five different locations (see text for details).

Base/GC	50/50/50	25/50/75	5/25/50	50/5/25
25	0.2188	0.1010	0.0234	0.1510
78	0.9140	0.8818	0.2960	0.8268
150	0.9286	0.9190	0.8578	0.2456
222	0.9230	0.8348	0.9378	0.4786
275	0.1916	0.1324	0.1934	0.0712

Table 4: Average proportion of trials along 5000 replicates in which the method was able to detect the presence of deletion at some base. Heterocygotic deletion at a single base with five different locations (see text for details).

5. CONCLUSIONS

Disorder detection from single-end NGS experiments is important in medical research, while it is challenging from a statistical point of view. In this work we have introduced a statistical model for the base coverage in such a scenario. The proposed model takes several sampling biases into account, including the possibly random fragment length and read length, and the GC-content bias. A method for disorder detection at base level which corrects for multiple comparisons and discrete p-values has been introduced. Simulations performed so far suggest that the introduced methods perform well.

Acknowledgements. First authors acknowledges financial support by CINBIO, 'Centro Singular de Investigación de Galicia 2016 - 2019. FEDER. Unha maneira de facer Europa. Promover o desenvolvemento tecnolóxico, a innovación e unha investigación de calidade. Second author acknowledges financial support by the Grant MTM2014-55966-P of the Spanish Ministerio de Economía y Competitividad.

REFERENCIAS

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* Vol. 57, No. 1, 289-300

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* Vol. 29, No. 4, 1165-1188

Heyse, J. F.(2011) A false discovery rate procedure for categorical data. *In Recent Advancements in Biostatistics, World Scientific Publishing Co.* Vol. 4, ch. 3, 43-58.

Lander ES, Waterman MS.(1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2(3): 231-239.

Li J, Lupat R, Amarasinghe KC, et al. (2012). CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28(10) 1307-1313.

Sathirapongsasuti, J. F., Lee, H., Horst, B. A. J., Brunner, G., Cochran, A. J., Binder, S. et al. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 27(19) 2648-54.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

A XESTIÓN DE RISCOS EN PROXECTOS DE EÓLICA MARIÑA MEDIANTE SIMULACIÓN DE EVENTOS DISCRETOS E SIMULACIÓN DE MONTECARLO

Adolfo Lamas Rodriguez[1], David Chas Alvarez[2] e José Antonio Muiña Dono [2]

[1] Universidade da Coruña, Navantia, UMI Navantia-UDC
[2] UMI Navantia-UDC

RESUMO

Con este traballo realizaremos un estudo sobre a xestión de riscos nun proxecto de eólica mariña usando a simulación de Montecarlo (MCS) para a análise de riscos e a simulación mediante eventos discretos (DES) para avaliar os posibles plans de mitigación que diminúan o efectos destes riscos.

Por este motivo crearanse dous modelos diferentes, o primeiro (Modelo A), usará as técnicas de MCS a partir dunha planificación real en Microsoft Project importada a Excel e utilizando o software @Risk, e un segundo modelo DES (Modelo B), realizado con ExtendSim, no que se avaliarán os efectos destas desviacións e as posibles medidas correctivas ao largo da duración do proxecto.

Palabras e frases chave: Xestión de riscos, Plans de mitigación, Simulación de Montecarlo, Simulación con eventos discretos

1. INTRODUCIÓN

O mercado da enerxía eólica mariña encontrase nunha etapa de expansión en Europa, duplicando a súa capacidade instalada nos últimos 3 anos, con unha capacidade total de 12.631 MW, e prevéndose para 2020 unha capacidade instalada total de 24,6 MW (Ewea, 2017), sendo a enerxía eólica mariña a máis dinámica en termos de capacidade instalada e desenvolvemento da tecnoloxía desde o 2010 (Carmen & Varela-v, 2017).

Ao mesmo tempo que o mercado medra, tamén o fai a complexidade das novas instalacións, realizándose estas cada vez a unha maior profundidade e distancia á costa.

Por outra parte, especialmente neste sector é necesario unha redución de custos que nos leven a obter cada vez a un menor prezo esta fonte de enerxía limpa. De feito, hoxe en día a fonte de enerxía máis barata en Europa é a enerxía eólica en terra pero a offshore demostrou que pode estar no mesmo rango de prezos (Ewea 2017a).

Son moitos os esforzos que se están facendo na eólica mariña para acadar esa redución de custos e o desenvolvemento tecnolóxico ó que ademais debemos sumar unhas exixencias de mercado cada vez maiores en canto á redución do prazo, que se aproximan á fabricación de un jacket por semana.

Ante un panorama tan esixente como o actual, onde o prazo é o presuposto do proxecto deben ser moi axustados, é onde cobra especial importancia o uso daquelas ferramentas capaces de predicir o comportamento do noso sistema ante todas as eventualidades que poidan xurdir.

Neste sentido, as técnicas de simulación de procesos industrias cobran especial importancia, sendo unha ferramenta imprescindible para acadar o correcto funcionamento da nosa empresa e pódennos axudar en distintas etapas do noso proxecto.

A simulación capacítanos para representar e estudiar un proceso real moi complexo e o entorno que o rodea, permitíndonos realizar sobre el as transformacións que desexemos coa intención de poder predicir a resposta futura do noso sistema.

A principal vantaxe que presenta o uso da simulación e a posibilidade de avaliar a resposta nun modelo artificial doutro sistema real, posibilitando a realización dunha grande cantidade de cambios no mesmo e

predicindo que sucedería se tomásemos diferentes medidas encamiñadas principalmente á redución de desperdicios.

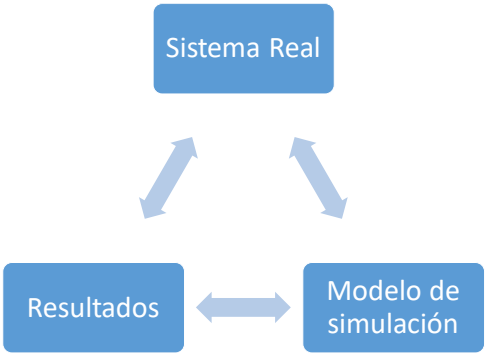


Figura 1: Simulación dun proceso industrial.

Estes modelos de simulación creados pódense clasificar:

- Segundo o instante de tempo que representan:
  - Estáticos: representan un sistema nun instante de tempo determinado.
  - Dinámicos: representan a evolución do sistema ao largo do tempo.
- Segundo a aleatoriedade das variables de estado:
  - Deterministas: Non existe variabilidade algunha.
  - Estocásticos: a lo menos unha variable de estado presenta aleatoriedade.
- Segundo o modo de evolución das variables de estado:
  - Continuo: as variables evolucionan de maneira continua ao largo do tempo.
  - Discretos: as variables de estado varían só en instantes determinados de tempo.

Por outra parte, cada vez é máis frecuente e necesario, que as empresas xestionen e traten de paliar a gran cantidade de riscos aos que se expoñen, para que estes afecten da menor forma posible aos seus obxectivos. A este proceso coñécese como Xestión de Riscos e trata de avaliar e paliar as incertezas asociadas a esas ameazas.

É o estudo de estas ferramentas, a simulación e principalmente a xestión de riscos, o que nos levou a realización deste traballo que describiremos nos seguintes apartados.

2. ESTADO DO ARTE

O proceso de Xestión de Riscos é unha acción complexa e sistemática na que debe participar toda a organización. Aínda que hoxe en día é posible acceder a unha gran cantidade de información, tamén e certo que os riscos aumentan e a influencia que estes teñen sobre o sistema é cada vez maior, por iso a necesidade dunha adecuada xestión de riscos e vital para manter o noso investimento a salvo.

Por tanto, podemos concluír, que non basta só con simular o noso propio proceso, se non que é necesario saber que podería ocorrer debido a dependencia con outros procesos e a factores que se escapan do noso control. Neste sentido é habitual encontrar nos últimos anos, numerosas publicacións sobre a Xestión de Riscos propoñendo os autores como principais ferramentas a simulación de Montecarlo (MCS) e a simulación mediante eventos discretos (DES).

Pese a que o seu uso é estendido e ámbolos dous permiten representar modelos estocásticos, a principal diferencia radica na variación do propio sistema co tempo, xa que no caso de Montecarlo estamos ante un caso estático e a simulación mediante eventos discretos é dinámica.

Polo seu uso estendido e a súa grande aplicabilidade neste caso, a continuación, procederemos a realizar un pequeno repaso sobre o estado do arte da simulación de riscos aplicando estas dúas técnicas.

(Electrical, Grids, Voltage, & Management, n.d.) trata de predicir a cantidade de interrupción ao cliente nunha empresa dedica ao subministro de produtos electrónicos usando a simulación de Montecarlo. Para isto, eles identifican as etapas e os riscos máis importante a partir de datos históricos, validando finalmente os resultados acadados na simulación cos obtidos no proceso real.

En (Nucd, 2010) estúdiase mediante o método de simulación de Montecarlo, a cantidade de impactos anuais que podería sufrir un almacén típico de produtos químicos e que poderían causar danos de distinta índole, ademais de capacitalos para a avaliación da influencia de distintas protección.

(Hojjati, 2015) utiliza a simulación de Montecarlo para avaliar os riscos en distintas etapas dun proxecto de IT usando o software Primavera Risk Analysis, concluíndo que as propostas inicias do prazo

de entrega e o custo poden ser insuficientes para acadar os obxectivos debido aos riscos e a variabilidade no proxecto que non se tivera en conta.

Resulta moi interesante o artigo (Rezaie, 2007), pois nel fálase da importancia da dependencia entre os distintos ricos avaliados. Os autores destacan que nunha simulación normal de Montecarlo estas dependencias entre variables non son avaliadas (por exemplo, a redución do prazo de entrega e o aumento de custe por ese sobre esforzo) levando a resultados confusos. Por este motivo, propoñen un método alternativo, consistente, na declaración en cada iteración dunha variable libre e a definición das outras como variables dependentes da primeira e das outras declaradas.

Namentres, tamén e posible atopar diferentes autores que usan as técnicas de simulación de eventos discretos para analizar os riscos, o seu impacto sobre o modelo e o estudo dos posibles plans de mitigación.

En (Cube et al. 2016), combínase o uso de técnicas de Montecarlo coas de simulación de eventos discretos para a análise económica na cuantificación de riscos, xerando datos con Montecarlo e representando a evolución dos modelos e os custos cos eventos discretos.

(Shah et al. 2017) propoñen unha metodoloxía para a cuantificación de riscos orientada a procesos, modelando e simulando distintos escenarios de fabricación e clasificándoos segundo conveña.

(Singh & Schmitt 2009), propoñen a utilización combinada das técnicas de Montecarlo e a simulación mediante as técnicas DES nunha cadea de subministro. Para isto, os autores utilizan Montecarlo para definir unha función que representa a probabilidade e a duración das interrupción, propoñendo o uso das ferramentas DES para a avaliación do impacto deses problemas no sistema e a viabilidade dos plans de mitigación propostos.

Este breve repaso do estado do arte, lévanos a propoñer un método de avaliación e xestión de riscos aplicada a un proceso de fabricación de eólica mariña, propoñendo as técnicas de Montecarlo para avaliar os riscos e a simulación de eventos discretos para avaliar os posibles plans de mitigación.

3. DESCRIPCIÓN DO PROCESO CONSTRUCTIVO

Tal e como se comentou no punto anterior, aplicaremos o noso estudo a un proceso real de fabricación dunha jacket (estrutura sobre á que se asentan os aeroxeradores no mar), mais concretamente, centrarémonos na fabricación dos pilotes, que poderían definirse como elementos construtivos capaces de servir de fixación e transmitir a cargas xeradas pola jacket e o aeroxerador ao chan.

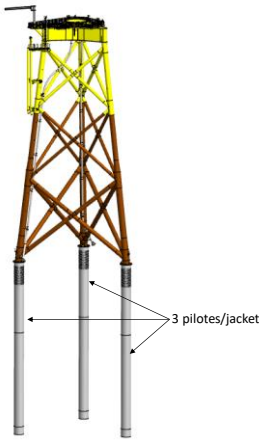


Figura 2: Jacket.

Basearemos o noso estudio nun proxecto actual de fabricación de 126 pilotes nun prazo de 13 meses, o cal é unha perfecta mostra das fortes esixencias actuais do mercado.

Para a realización do noso traballo precisaremos aprovisionarmos da nosa materia prima, neste caso o aceiro, o cal segundo o acordo alcanzado, chegará ao noso estaleiro segundo se amosa na Táboa 1.

Fito de chegada de aceiro	Días dende o inicio do proxecto
1 (Pilotes do 1 ao 72)	7
2 (Pilotes do 73 ao 126)	76

Táboa 1: Chegada do aceiro ao estaleiro.

O proceso de fabricación de pilotes móstrase na Figura 3, e coma se pode apreciar é un proxecto de certa envergadura que conta con distintas etapas, as cales teñen que realizarse en paralelo ou en serie, e que polo tanto se poderían asemellar a calquera outro proceso doutra industria, polo que a viabilidade do noso estudio podería extrapolarse a outras industrias moi diferentes.

1	Corte e biselado das chapas
2	Curvado das chapas
3	Soldadura lonxitudinal
4	Ensamblado de seccións
5	Soldadura circular
6	Soldadura de reforzos
7	Testado non destrutivo
8	Marcado de seccións
9	Inspección Final

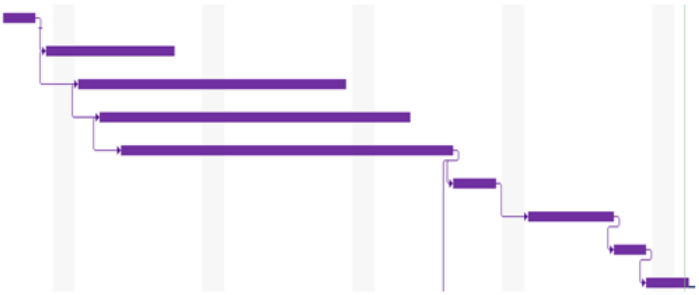


Figura 3: Diagrama de fluxo do proceso de fabricación de pilotes.

Para a realización da planificación base deste proxecto e que tomaremos como referencia, asúmese unha xornada de traballo de cinco días semanais, deixando a xornada dos sábado como marxe ante os posibles imprevistos que poidan aparecer. Por outra banda, a capacidade de fabricación do proxecto varía ao largo do tempo, por unha parte, dispoñemos de 15 liñas de fabricación de pilotes para os 60 primeiros, 12 liñas para os seguintes 48 pilotes e ate 18 para os últimos 18 pilotes.



Figura 4: Parte do proceso construtivo dos pilotes.

Unha vez o proceso de fabricación dos pilotes finalizou, estes deben ser almacenados tal e como se amosa na Figura 5, ate o cumprimento dos distintos fitos de entrega do noso cliente de acordo ao contrato establecido, e que se amosa na Táboa 2.



Figura 5: Almacén final de pilotes.

Fito de entrega dos pilotes	Días dende o comezo do proxecto
Do 1 ao 18	198
Do 19 ao 36	228
Do 37 ao 54	261
Do 55 ao 72	293
Do 73 ao 90	326
Do 91 ao 108	361
Do 109 ao 126	409

Táboa 2: Fechas de entrega dos pilotes ao cliente.

4. MODELOS DE SIMULACIÓN

O noso modelo de simulación debe representar a realidade do proceso de fabricación de pilotes de acordo ao explicado no apartado anterior.

Para a validación do noso modelo decidiuse partir dun modelo determinista, polo tanto carente de variabilidade algunha, que debería representar o que acontece na planificación teórica dispoñible ate o momento, outorgada pola compañía.

Unha vez o noso modelo foi validado, procedeuse a facer un estudo das posibles causas que farían varias as previsión establecidas para as seguintes etapas, mediante a axuda de expertos e os datos históricos de duración das actividades outorgados polo equipo de proxectos e baseados na duración das tarefas en proxectos de semellante envergaduras.

Estes datos históricos foron recollidos dun diagrama Gantt coa súa táboa de tarefas en formato .mpp (Microsoft Project), e os principais campos de información utilizados a hora de definir as distribucións empíricas de duración de tarefas e chegada de material foron:

- Data real e teórica de chegada de materias primas.
- Data real e planificada de comezo.
- Duración real e planificada.
- Data real e planificada de finalización.

Cos datos mencionados anteriormente, elaborouse unha base de datos sobre os retrasos no abastecemento de materia prima e a duración de cada tarefa durante o proxecto. A partir de esta base de datos, procedeuse a estimar as distribución empíricas para cada proceso e que foron utilizas para o estudo posterior.

Unha vez o modelo foi validado e as posibles causas de variabilidade e o seu efecto asignado, procedeuse á elaboración dos modelos estocásticos, tanto para a MCS, ó que denominamos como modelo A, como para DES, ó que denominamos como modelo B.

4.1. Simulación de Montecarlo (Modelo A)

O modelo de simulación realizado para a análise de riscos mediante a simulación de Montecarlo foi, en primeira instancia, realizado con Microsoft Project e inclúe ate 1331 voces que representan o proceso íntegro da fabricación dos 126 pilotes.

Concluída esta tarefa decidiuse utilizar o complemento de Excel @Risk, dado o grande número de autores que o utilizan, importando os datos do proxecto que se desenvolveu en Project a Excel e introducindo de forma intuitiva e sinxela a variabilidade que comentabamos no punto anterior.

Como dato extraído do modelo resúltanos interesante a data de finalización das actividades con Montecarlo para saber en que casos e con que probabilidade eramos capaces de acadar os obxectivos establecidos co noso cliente.

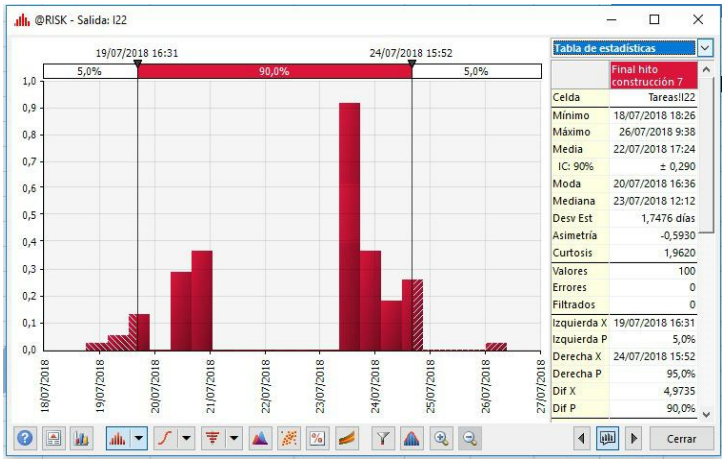


Figura 6: Resultados acadados con @Risk.

Como se pode ver na figura anterior, o software utilizado é capaz de amosar unha serie de parámetros dos resultados obtidos, como a mediana, a desviación estándard ou a moda, os cales poderíamos incluso adaptar a unha función estatística.

Resultou unha vantaxe interesante tamén que se respectase o calendario laboral introducido no modelo inicial de Microsoft Project, xa que como se pode ver na Figura 6, respéctanse tanto os festivos coma os fines de semana.

4.2. Simulación de eventos discretos (Modelo B)

Como xa se comentou nos apartados anteriores, un segundo modelo, denominado, Modelo B, foi realizado con ExtendSim para avaliar mediante simulación os plans de mitigacións propostos.

No devandito modelo representarase a estratexia construtiva dos pilotes completa tal e coma se fixera no caso anterior, e para facilitar a modificación futura de parámetros decidiuse que o modelo fose paramétrico para que puidese intercambios os datos de partida cun ficheiro en Excel.

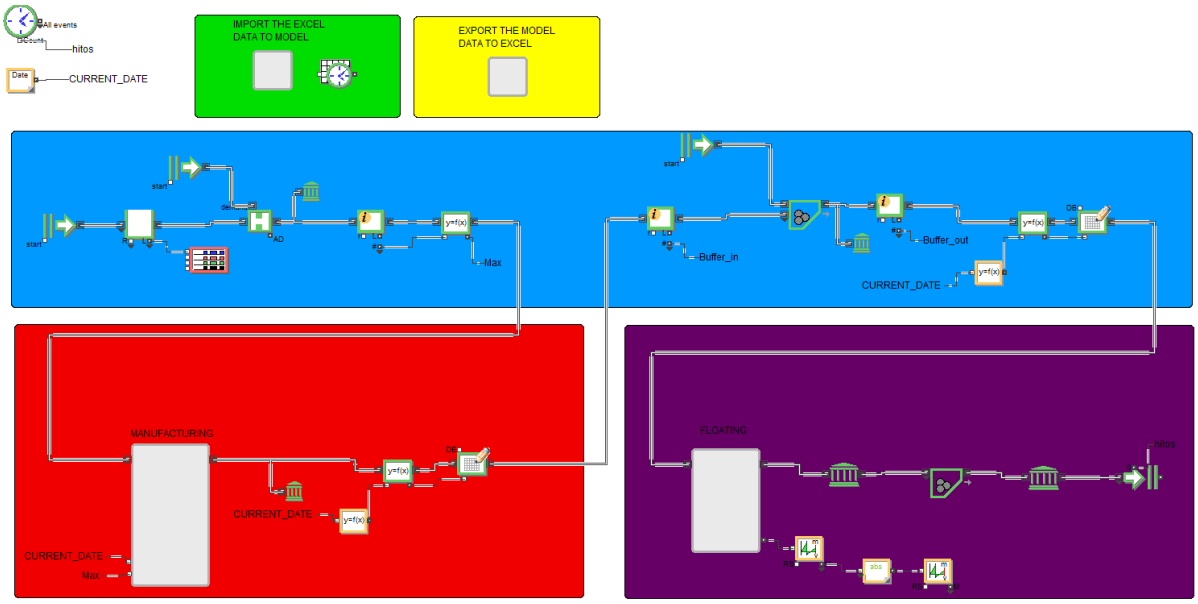


Figura 7: Modelo realizado con ExtendSim.

Na Figura 7 móstrase o modelo realizado con ExtendSim en grandes bloques xerárquicos que non son posibles describir detalladamente neste traballo, polo que se optou por utilizar un código de cores para representar as principais etapas e o fluxo de traballo, tal e como se explica a continuación:

- Os bloques de cor verde e amarelo son os encargados do intercambio de datos co ficheiro de Excel, importando datos antes do comezo da simulación e exportando os resultados tras a finalización da mesma, para a súa mellor interpretación.
- O bloque de cor azul controla as restricións do modelo, (chegada dos materiais, capacidade da planta, fitos do cliente...)
- O bloque de cor vermello representa a estratexia construtiva de fabricación de pilotes
- En último lugar, o bloque de cor púrpura é o encargado de executar a análise de riscos, das variables de saída que se determinen.

Debemos destacar que neste caso, represéntanse unicamente as condicións do taller (celas de traballo, operarios, etc.) e importamos os datos coa chegada dos materias, sendo agora o propio modelo o que xerando datos de forma similar a como se faría utilizando o Microsoft Project, segundo o avance dos procesos no taller.

5. RESULTADOS

Os resultados deste traballo divídense en dous apartados, o primeiro presenta a análise de riscos debido o retraso na chegada do aceiro, esta análise foi obtida mediante a simulación de Montecarlo (Modelo A), de forma posterior, usando a simulación de eventos discretos, obtívose un plan de mitigación para o risco cuantificado coa simulación de Montecarlos (MCS).

5.1. Análise de riscos coa simulación de Montecarlo (MSC)

Mediante a simulación de Montecarlo realizouse unha cuantificación do risco de romper os fitos de entrega fixados polo cliente. Esta análise fíxose baixo 3 hipóteses diferentes, das cales se representan os resultados nas táboas 3, 4 e 5

- Chagada do aceiro na data prevista para ámbolos dous fitos de entrega (táboa 3).
- 1 mes de retraso na entrega do primeiro fito de aceiro (táboa 4).
- 2 meses de retraso na entrega do primeiro fito de aceiro (táboa 5).

Fabricación			Entrega de pilotes	Retraso (días)	Sábados dispoñibles
Descrición	Mínimo	Máximo			
Fito 1 de fabricación (IC=90%)	122	124	198	-	16
Fito 2 de fabricación (IC=90%)	165	167	228	-	22
Fito 3 de fabricación (IC=90%)	208	212	261	-	28
Fito 4 de fabricación (IC=90%)	247	249	293	-	34
Fito 5 de fabricación (IC=90%)	327	332	326	6	35
Fito 6 de fabricación (IC=90%)	368	371	361	10	41
Fito 7 de fabricación (IC=90%)	410	414	409	5	47

Táboa 3: Resultados da MSC coa chegada do aceiro na data prevista.

Fabricación			Entrega de pilotes	Retraso (días)	Sábados dispoñibles
Descrición	Mínimo	Máximo			
Fito 1 de fabricación (IC=90%)	152	154	198	-	12
Fito 2 de fabricación (IC=90%)	194	196	228	-	18
Fito 3 de fabricación (IC=90%)	238	241	261	-	24
Fito 4 de fabricación (IC=90%)	277	279	293	-	30
Fito 5 de fabricación (IC=90%)	356	361	326	35	35
Fito 6 de fabricación (IC=90%)	398	400	361	39	41
Fito 7 de fabricación (IC=90%)	441	443	409	34	47

Táboa 4: Resultados da MSC con 1 mes de retraso na chegada do aceiro.

Fabricación			Entrega de pilotes	Retraso (días)	Sábados dispoñibles
Descrición	Mínimo	Máximo			
Fito 1 de fabricación (IC=90%)	183	182	198	-	7
Fito 2 de fabricación (IC=90%)	223	225	228	-	14
Fito 3 de fabricación (IC=90%)	269	273	261	12	20
Fito 4 de fabricación (IC=90%)	306	307	293	14	25
Fito 5 de fabricación (IC=90%)	387	394	326	58	35
Fito 6 de fabricación (IC=90%)	427	433	361	72	41
Fito 7 de fabricación (IC=90%)	472	473	409	66	47

Táboa 5: Resultados da MSC con 2 meses de retraso na chegada do aceiro.

Coma se pode ver nas táboas 3 e 4 os retrasos poden ser mitigados coa utilización dos sábados dispoñibles dende a entrega do aceiro ata a data de entrega dos pilotes finalizados, pola contra, na última

das hipóteses, táboa 5, o retraso dos últimos 3 fitos non pode ser mitigado utilizando un sábados dispoñibles como plan de mitigación.

5.2. Deseño de plans de mitigación

Co obxectivo de desenvolver un plan de mitigación que permita satisfacer os fitos do cliente no caso da terceira hipótese (2 meses de retraso na entrega do primeiro fito de aceiro), utilizouse o modelo de simulación de eventos discretos, tomando coma variable o número de liñas de fabricación en función do tempo, as cales poden alcanzar un número máximo de 18 liñas en paralelo.

O modelo de simulación de eventos discretos, ofrece dúas solucións para evitar o incumprimento dos fitos do cliente.

A primeira das solucións pasa por incrementar o número de liñas de fabricación a 18 para os primeiros 72 pilote, pasando a 12 liñas de fabricación do pilote 73 ao 108, e para finalizar volver a 18 liñas para os 18 últimos pilotes. Quedando como resultado da simulación a táboa 6.

Fabricación			Entrega de pilotes	Retraso (días)	Sábados dispoñibles
Descrición	Mínimo	Máximo			
Fito 1 de fabricación (IC=90%)	140	143	198	-	7
Fito 2 de fabricación (IC=90%)	182	184	228	-	14
Fito 3 de fabricación (IC=90%)	222	225	261	-	20
Fito 4 de fabricación (IC=90%)	264	265	293	-	25
Fito 5 de fabricación (IC=90%)	346	346	326	20	35
Fito 6 de fabricación (IC=90%)	387	388	361	27	41
Fito 7 de fabricación (IC=90%)	427	429	409	20	47

Táboa 6: Primeira solución proposta pola simulación de eventos discretos.

Coma se pode ver na táboa 6, os fitos non chegan a cumprirse con esta modificación das liñas se se utiliza tan so unha 5 xornadas de traballo a semana, pero como se pode ver na última columna da táboa, disponse de sábados suficientes para paliar os retrasos xerados e cumprir os fitos do cliente.

A segunda solución que ofrece o modelo de eventos discretos, pasa por manter na totalidade do proxecto 18 liñas de fabricación de pilotes, obtendo os resultados mostrados na táboa 7, na que se pode comprobar o cumprimento da totalidade dos fitos do cliente sen a necesidade de empregar 6 xornadas de traballo por semana.

Fabricación			Entrega de pilotes	Retraso (días)	Sábados dispoñibles
Descrición	Mínimo	Máximo			
Fito 1 de fabricación (IC=90%)	141	143	198	-	7
Fito 2 de fabricación (IC=90%)	182	184	228	-	14
Fito 3 de fabricación (IC=90%)	223	226	261	-	20
Fito 4 de fabricación (IC=90%)	264	265	293	-	25
Fito 5 de fabricación (IC=90%)	304	305	326	-	35
Fito 6 de fabricación (IC=90%)	346	347	361	-	41
Fito 7 de fabricación (IC=90%)	387	390	409	-	47

Táboa 7: Segunda solución proposta pola simulación de eventos discretos.

6. CONCLUSIÓNS

A realización deste traballo permitiu analizar os riscos reais e a súa influencia en distintas etapas dun proceso real, xustificando a necesidade da realización de este tipo de estudos xa que a variabilidade

estatística que afecta a tódolos procesos poden alterar tanto de maneira positiva como negativa os resultados acadados.

Neste traballo en primeiro lugar utilizouse un software para a simulación de Montecarlo, que destacou pola súa sinxela comunicación con Microsoft Project, un dos softwares de planificación de proxectos máis utilizados na actualidade tanto por pequenas coma por grandes empresas. Con este programa a simulación de Montecarlo e sinxela de utilizar ate para persoas sen coñecementos específicos e permite identificar a resposta do sistema ante a existencia deses riscos, tanto positivos como negativos.

Sen embargo, a natureza da simulación de Montecarlo, por ser estática, non permite ver a evolución do sistema no tempo, por iso resulta interesante aplicar as técnicas DES para a avaliación de riscos e ver como responde o sistema co tempo, ademais de facilitar de forma notable a realización de cambios e a re asignación de recursos de acordo as necesidades existentes.

REFERENCIAS

Carmen, S., & Varela-v,P (2017). Estimation of the potential effects of offshore wind on the Spanish economy. Renewable Energy, 111, 815-824.

Ewea. (2017). Reviving wind markets and delivering on our 2030 objetives. Retrieved July 14, 2017, from http://www.ewea.org/statistics/offshore-statistics/

Ewea. (2017b). "The European Offshore Wind Industry." Key Trends and Statistics 2016. www.ewea.org. http://www.ewea.org/statistics/offshore-statistics/

Liaqat A. Shah, Alain Etienne, Ali Siadat & François Vernadat. (2017). "Process-Oriented Risk Assessment Methodology for Manufacturing Process Evaluation." International Journal of Production Research 55:15 (June): 4516–29.

Cube, Philipp Von, Lasse Härtel, Robert Schmitt, Christophe Ponsard, Philippe Massonet, Renaud De Landtsheer, Gustavo Ospina, Stephan Printz, and Sabina Jeschke. (2016). "Monetary Quantification of Supply Risks of Manufacturing Enterprises - Discrete Event Simulation Based Approach." Procedia CIRP 57: 164–70.

Hojjati, S & Noudehi, N (2015).The use of Monte Carlo simulation in quantitative risk assessment of IT projects, Int. J. Advanced Networking and Applications, 1, 2616-2621.

Shiromani A.V. Goerdin & Johan J. Smit & Ravish P.Y. Mehairjan (2015). Monte Carlo simulation applied to support risk-based decision making in electricity distribution networks, PowerTech, 2015 IEEE Eindhoven.

Singh, Mahender, and Amanda J Schmitt. (2009). "Quantifying Supply Chain Disruption Risk Using Monte Carlo and Discrete-Event Simulation." In Proceedings of the 2009 Winter Simulation Conference, 1237–48. Hilton Austin Hotel Austin, TX, USA.

Rezaize, K & Amalnik, M.S. & Gereie, A & Ostadi, B & Shakhseniaee, M (2007). Using extended Monte Carlo simulation method for the improvement of risk management : Consideration of relationships between uncertainties, Applied Mathematics and Computation, 190, 1492-1501.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Nonparametric covariate significance tests in mixture cure models

Ana López Cheda[1], Ricardo Cao[1] and M. Amalia Jácome[2]

[1]Grupo MODES, Dep. Matemáticas, INIBIC, Facultade de Informática, Universidade da Coruña
[2]Grupo MODES, Dep. Matemáticas, INIBIC, Facultade de Ciencias, Universidade da Coruña

ABSTRACT

A covariate significance test on the probability of cure is proposed, based on a completely nonparametric method for the estimation of mixture cure models. Its efficiency is evaluated in a Monte Carlo simulation study, in which the distribution of the test is approximated by bootstrap. Finally, the proposed methodology is applied to a database of sarcomas from the University Hospital of Santiago (CHUS).

Keywords: survival analysis; censored data; bootstrap.

1. INTRODUCTION

A common assumption in survival analysis is that all of the subjects will eventually experience the event of interest if they are followed for long enough. However, this assumption is becoming more unlikely to be true, for the remarkable progress in cancer treatments, which led to longer patient survival and increased the probability of cure. As a consecuence, often classical survival analysis is not realistic and a new approach is necessary. Consequently, there has been an increasing interest in cure models over the past few years.

Cure model techniques allow modelling and predicting the prognosis of a patient considering, as a novelty, the possibility that the individual is cured. The goal in cure models is to estimate the cure probability of the population (incidence) and the survival function of the susceptible individuals (latency). The existing methods are limited to parametric and semiparametric models, which require some assumptions on the underlying distribution of the data. Nonparametric methods provide an alternative that does not require any assumption. This work considers the completely nonparametric approach with covariates by López-Cheda et al. (2017a).

2. MAIN RESULTS

Let Y be the time to occurrence of the event (for example, death), $\boldsymbol{X}$ be a set of covariates and ν a binary variable where $\nu = 0$ (or $Y < \infty$) indicates that the individual belongs to the susceptible group and $\nu = 1$ (or $Y = \infty$) indicates that the subject is cured. The conditional probability of not being cured is $p(\mathbf{x}) = P(\nu = 0|\mathbf{X} = \mathbf{x})$ and the conditional survival function of the uncure patients is $S_0(t|\mathbf{x}) = P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 0)$. Then, the mixture cure model can be written as:

$$S(t|\mathbf{x}) = 1 - p(\mathbf{x}) + p(\mathbf{x})S_0(t|\mathbf{x}), \tag{1}$$

where $1-p(\mathbf{x})$ is the incidence and $S_0(t|\mathbf{x})$ is the latency. We assume that each individual is subject to random right censoring and that the censoring time, C, is independent of Y given the covariates $\mathbf{X}$. Let $T = \min(Y, C)$ be the observed time and $\delta = I(Y \leq C)$ the uncensoring indicator. Observe that $\delta = 0$ for all the cured patients. This also happens for uncured patients with censored lifetime. The sample is denoted by $\{(\mathbf{X}_i, T_i, \delta_i), i = 1, \ldots, n\}$, independent and identically distributed (iid) copies of the random vector $(\mathbf{X}, T, \delta)$.

Without loss of generality, let X be a univariate continuous covariate. To estimate the conditional survival function with covariates, the generalized Kaplan-Meier estimator by Beran (1981) is considered:

$$\hat{S}_h(t|x) = \prod_{T_{(i)} \leq t} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right), \tag{2}$$

where

$$B_{h(i)}(x) = \frac{K_h(x - X_{(i)})}{\sum_{j=1}^{n} K_h(x - X_{(j)})}$$

are the Nadaraya-Watson (NW) weights with $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, the rescaled kernel, with bandwidth $h \to 0$, $T_{(i)}$ is the i-th ordered statistic for the T-sample and $X_{(i)}$ and $\delta_{(i)}$ are the corresponding i-th concomitants in the X and the δ samples. Xu and Peng (2014) introduced the following kernel incidence estimator:

$$1 - \hat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right) = \hat{S}_h(T^1_{\max}|x), \tag{3}$$

where $T^1_{\max} = \max_{i:\delta_i=1}(T_i)$ is the largest uncensored failure time. López-Cheda et al. (2017a) study deeply the incidence estimator in (3), and using (1), they proposed the following nonparametric latency estimator:

$$\hat{S}_{0,b}(t|x) = \frac{\hat{S}_b(t|x) - (1 - \hat{p}_b(x))}{\hat{p}_b(x)}, \tag{4}$$

where $\hat{S}_b(t|x)$ is the Beran estimator of $S(t|x)$ in (2) and $1 - \hat{p}_b(x)$ is the estimator by Xu and Peng (2014) in (3). The nonparametric latency estimator in (4) has been extensively studied in López-Cheda et al. (2017b).

In cancer studies it is interesting to test if a covariate has some influence on the cure rate or on the survival time of the susceptible patients. Let $\mathbf{W} = (\mathbf{X}, \mathbf{Z}) = (X_1, ..., X_q, Z_1, .., Z_p)$ be the explanatory covariates. A covariate significance test, based on the significance test by Delgado and González-Manteiga (2001), is proposed to study if the cure probability, as a function of the covariate vector $\mathbf{W}$, only depends on $\mathbf{X}$, but not on $\mathbf{Z}$:

$$H_0 : E(\nu|\mathbf{X}, \mathbf{Z}) = 1 - p(\mathbf{X}) \qquad H_1 : E(\nu|\mathbf{X}, \mathbf{Z}) = 1 - p(\mathbf{X}, \mathbf{Z}),$$

where the function $p(\mathbf{X}, \mathbf{Z})$ depends on $\mathbf{Z}$ (i.e. is a function that depends not only on $\mathbf{X}$ but also on $\mathbf{Z}$).

Since ν is not observed, let η be the variable satisfying $E(\nu|\mathbf{X}) = E(\eta|\mathbf{X}) = 1 - p(\mathbf{X})$ as follows:

$$\eta = \frac{\nu(1 - I(\delta = 0, T \leq \tau(\mathbf{x})))}{1 - G(\tau(\mathbf{x})|\mathbf{X})},$$

where $\tau(\mathbf{x})$ is a time beyond which a subject is considered cured. Without loss of generality, in practice, we estimate $\tau(\mathbf{X})$ in the following way for a continuous univariate covariate X: using a bandwidth h_τ, we consider a subset of individuals j with $|x_j - x_i| < h_\tau$, and $\tau(x_i)$ will be estimated as the largest T_j with $\delta_j = 1$ in the subset. If there is no $\delta_j = 1$ in the subset, then $\tau(x_i)$ is equal to the available $\tau(x_l)$ for the nearest x_l to x_i. If there are several nearest values to determine x_l, then we consider that $\tau(x_i)$ is the mean of those. Preliminary studies suggested that a good bandwidth choice is:

$$h_\tau = \left(X_{[n]} - X_{[1]}\right)\ 0.25\ n^{-1/9},$$

where $X_{[n]}$ and $X_{[1]}$ are the maximum and minimum values of the covariate X, respectively.

Moreover, it is assumed that C does not depend on the covariates, $\mathbf{X}$, and then $G(\tau(\mathbf{X}_i)|\mathbf{X})$ is estimated by the product limit estimator, $\hat{G}(\tau(\mathbf{X}_i))$. This gives the following predictions for the η_i:

- $\delta_i = 1$ (failure) $\Rightarrow$ uncured ($\nu_i = 0$) $\Rightarrow \hat{\eta}_i = 0$.
- $\delta_i = 0$ (censored) and $T_i \leq \tau(\mathbf{X}_i) \Rightarrow \hat{\eta}_i = \frac{\nu_i(1-1)}{1-\hat{G}(\tau(\mathbf{X}_i))} = 0$.
- $\delta_i = 0$ (censored) and $T_i > \tau(\mathbf{X}_i) \Rightarrow$ cured ($\nu_i = 1$) $\Rightarrow \hat{\eta}_i = \frac{1}{1-\hat{G}(\tau(\mathbf{X}_i))}$.

For $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$ the three cases in Table 1 can be considered. Only the first two for different types of data will be studied here, and the third one will be addressed in a future work.

	Dimension $\mathbf{W}$	Dimension $\mathbf{X}$	Dimension $\mathbf{Z}$
Case 1: $H_0 : 1 - p(z) = 1 - p$	1	0	1
Case 2: $H_0 : 1 - p(x, \mathbf{z}) = 1 - p(x)$	$1 + p$	1	p
Case 3: $H_0 : 1 - p(\mathbf{x}, \mathbf{z}) = 1 - p(\mathbf{x})$	$q + p$	q	p

Table 1: Different cases of covariate significant testing.

Case 1:
First, we focus on the case with only one covariate, $\mathbf{W} = Z$. We study if the cure rate can be considered as a constant value versus if it depends on the covariate:

$$H_0 : E(\nu|Z) = 1 - p \qquad\qquad H_1 : E(\nu|Z) = 1 - p(Z),$$

where $p(Z)$ is not a constant value under the alternative hypothesis. Our test will be based on the following observations: $\{(Z_i, \hat{\eta}_i), i = 1, \ldots, n\}$. For Case 1, we consider four different scenarios, depending on the type of data to analyse: continuous, binary, ordinal qualitative and non-ordinal qualitative. Note that the binary and the ordinal qualitative scenarios are particular cases of the continuous scenario.

Following Delgado and González-Manteiga (2001), the statistics we will propose when Z is a quantitative or ordinal qualitative variable is based on the following process:

$$T_n(z) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\eta}_i - \left(\frac{1}{n} \sum_{j=1}^{n} \hat{\eta}_j \right) \right) I(Z_i \leq z). \tag{5}$$

Possible test statistics are the Cramér-von Mises (CvM) test, $C_n = \sum_{i=1}^{n} T_n^2(Z_i)$, or the Kolmogorov-Smirnov (KS) test, $K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(Z_i)|$. The test statistic null distribution is approximated by bootstrap, using an independent naive resampling.

In the case of a non-ordinal qualitative covariate Z, there is no agreed way to order the values of Z from lowest to highest. This makes impossible to compute the indicator function in the test statistic (5). We propose to consider all the possible combinations of the values of Z and compute $T_n(z)$ (and also C_n and K_n) for each "ordered" combination. For example, if Z can take 3 values, there will be $3! = 6$ possible values of $T_n(z)$ (from which we obtain C_n and K_n), each corresponding to every sorting of those values. Finally, we compute the maximum of C_n and K_n and compare it with the critical point obtained by bootstrap likewise.

Case 2:
In Case 2, $\mathbf{W} = (X, \mathbf{Z})$ has $m + 1$ dimension, with a one-dimensional covariate X and an m-dimensional covariate $\mathbf{Z}$. We study if the cure probability only depends on the covariate X, that is:

$$H_0 : E(\nu|X, \mathbf{Z}) = 1 - p(X), \qquad\qquad H_1 : E(\nu|X, \mathbf{Z}) = 1 - p(X, \mathbf{Z}),$$

where $p(X, \mathbf{Z})$ depends on $\mathbf{Z}$ under the alternative hypothesis. To do this we use the observations $\{(X_i, \mathbf{Z}_i, \hat{\eta}_i), i = 1, \ldots, n\}$.

***X* continuous:**
The statistic is defined as:

$$T_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_h(X_i)\left(\hat{\eta}_i - \hat{m}_h(X_i)\right) I\left(\mathbf{W}_i \leq \mathbf{w}\right), \tag{6}$$

where $\hat{f}_h$ is the estimated density function of the covariate X which depends on the bandwidth h, and $\hat{m}_h$ is the nonparametric estimator of the regression function $m(x) = E\left(\hat{\eta}|X = x\right)$, obtained by the Nadaraya-Watson kernel method. Similarly to Case 1, we consider the Cramér-von Mises, $C_n = \sum_{i=1}^{n} T_n^2(\mathbf{W}_i)$ and the Kolmogorov-Smirnov, $K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(\mathbf{W}_i)|$ statistics. The test distribution under H_0 is approximated by bootstrap.

***X* discrete or ordinal qualitative:**
For a discrete or ordinal qualitative variable X, the estimated density, $\hat{f}_h(X_i)$, and the estimated regression function, $\hat{m}_h(X_i)$, in the test statistic in (6) are replaced by

$$\hat{\Pi}(X_i) = \frac{1}{n}\sum_{j=1}^{n} I(X_j = X_i)$$

and

$$\hat{m}(X_i) = \frac{\sum_{j=1}^{n} I(X_j = X_i)\hat{\eta}_j}{\sum_{j=1}^{n} I(X_j = X_i)} = \frac{\frac{1}{n}\sum_{j=1}^{n} I(X_j = X_i)\hat{\eta}_j}{\hat{\Pi}(X_i)},$$

respectively.

***X* non-ordinal qualitative:**
For a qualitative variable in $\mathbf{W}$ with no intrinsic order in its values, the indicator function $I\left(\mathbf{W}_i \leq \mathbf{w}\right)$ in the test statistic is computed for all the possible "ordered" permutations of the values of $\mathbf{W}$.

For example, for the case of $\mathbf{W} = (X, Z)$ qualitative, we consider that the values of the variables are $X_i \in \{a_1, \ldots, a_{l_x}\}$ and $Z_i \in \{b_1, \ldots, b_{l_z}\}$. The general test statistic is defined as:

$$T_n(x, z) = \frac{1}{n}\sum_{i=1}^{n} \hat{\Pi}(X_i)\left(\hat{\eta}_i - (1 - \hat{m}(X_i))\right) I\left(X_i \leq x\right) I\left(Z_i \leq z\right). \tag{7}$$

Since both variables are qualitative (which are neither numeric, nor ordinal) then the order of these values can be important (mainly for CvM). Therefore, we obtain all the possible ordered values of $\mathbf{W}$ (for instance, if $m = 2$ and each variable can take 3 different values, then we will have $(3!)(3!) = 36$ different orders of these values), we sort them and we consider the maximum values of the tests C_n and K_n evaluated in $\mathbf{W}$. Finally, we compare them with their corresponding critical point obtained by bootstrap likewise.

Specifically, let us denote by $S_n(a_j, b_k)$:

$$S_n(a_j, b_k) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \hat{\Pi}(X_i)\left(\hat{\eta}_i - (1 - \hat{m}(X_i))\right) I\left(X_i = a_j\right) I\left(Z_i = b_k\right).$$

In order to compute the statistic $T_n(x, z)$ in (7), we take into account that $T_n(a_1, b_1) = S_n(a_1, b_1)$. In general:

$$T_n(a_j, b_k) \quad = \quad \sum_{r=1}^{j}\sum_{s=1}^{k} S_n(a_r, b_s),\ j = 1, \ldots, l_x;\ k = 1, \ldots, l_z,$$

for each "ordered" combination of the values of X and Z: $a_1 < a_2 < \cdots < a_{l_x}$; $b_1 < b_2 < \cdots < b_{l_z}$.

3. SIMULATION RESULTS

For the sake of simplicity, in this section we only considered the most representative situations of Cases 1 and 2.

Case 1:
Under the null hypothesis, $H_0: E(\nu|Z) = 1-p$, we consider four different constant values for the incidence: $1-p = 0.7$, $1-p = 0.5$, $1-p = 0.3$ and $1-p = 0.2$. Under the alternative hypothesis, $H_1: E(\nu|Z) = 1-p(Z)$, and when Z is a continuous covariate, we consider Model 1 and Model 2 defined below. If Z is discrete, binary or qualitative, we only consider Model 1 in order to generate the data.

Model 1 The incidence is $1-p(z)$, where

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1+\exp(\beta_0 + \beta_1 z)}, \tag{8}$$

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$, and the latency is a function which almost fulfills the proportional hazard model and that it has been truncated

$$S_0(t|z) = \begin{cases} \frac{\exp(-\lambda(z)t)-\exp(-\lambda(z)\tau_0)}{1-\exp(-\lambda(z)\tau_0)},\ t \le \tau_0 \\ 0,\ t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda(z) = \exp((z+20)/40)$. The percentage of censored data is 54% and of cured data is 47%.

Model 2 The probability of uncure is:

$$p(z) = \frac{\exp\left(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3\right)}{1+\exp\left(\beta_0 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3\right)}, \tag{9}$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$ and $\beta_3 = 0.0020$ and the survival function of the susceptible population is

$$S_0(t|z) = \frac{1}{2}\left(\exp(-\alpha(z)t^5) + \exp(-100t^5)\right),$$

with

$$\alpha(z) = \frac{1}{5}\exp((z+20)/40).$$

In this model, the percentages of censored and cured data are 62% and 53%, respectively.

***Z* continuous:**
We considered Models 1 and 2 for a continuous covariate $Z \sim U(-20, 20)$. The results in Table 2 (Table 3) were obtained under the null (alternative) hypothesis. It is noteworthy that, under the null hypothesis, the size of the test is very similar to the significance level, α=0.05, for the different constant values for p. Furthermore, under the alternative hypothesis, the power of the test is very close (or even equal) to 1.

***Z* qualitative:**
A qualitative covariate Z with three possible values $\{b_1, b_2, b_3\}$ is considered. According to the probability mass function given by $\Pi_z(b_i) = P(Z = b_i)$, we study two situations: $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (1/3, 1/3, 1/3)$ and $(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3)) = (3/5, 1/5, 1/5)$. The observations were simulated from Model 1, finding the values that yield $p(b_1) = 0.5$, $p(b_2) = 0.2$, and $p(b_3) = 0.7$.

The results in Table 4, which were obtained under the null hypothesis, are very similar to the significance level, $\alpha = 0.05$. Regarding the alternative hypothesis (see Table 5), the power of the test is higher for large sample sizes and when $\Pi_z(b_1) = \Pi_z(b_2) = \Pi_z(b_3) = 1/3$.

		Model 1		Model 2	
n	p	CvM	KS	CvM	KS
50	0.3	0.0484	0.0600	0.0550	0.0636
	0.5	0.0500	0.0564	0.0490	0.0582
	0.7	0.0446	0.0472	0.0510	0.0504
	0.8	0.0418	0.0360	0.0400	0.0412
100	0.3	0.0608	0.0636	0.0544	0.0584
	0.5	0.0494	0.0562	0.0488	0.0586
	0.7	0.0444	0.0520	0.0474	0.0468
	0.8	0.0414	0.0396	0.0484	0.0470
200	0.3	0.0490	0.0530	0.0528	0.0534
	0.5	0.0552	0.0616	0.0540	0.0572
	0.7	0.0498	0.0508	0.0516	0.0514
	0.8	0.0446	0.0432	0.0466	0.0480

Table 2: Size of the test for Case 1 with Z continuous with distribution $U(-20, 20)$ under the null hypothesis.

	Model 1		Model 2	
n	CvM	KS	CvM	KS
50	0.9890	0.9862	0.4200	0.4148
100	0.9994	0.9992	0.7330	0.7402
200	1	1	0.9670	0.9746

Table 3: Power of the test for Case 1 with Z continuous with distribution $U(-20, 20)$ under the alternative hypothesis.

n	p	CvM	KS
50	0.2	0.0512	0.0526
	0.5	0.0494	0.0520
100	0.2	0.0544	0.0538
	0.5	0.0488	0.0532
200	0.2	0.0486	0.0500
	0.5	0.0456	0.0516

Table 4: Size of the test for Case 1 with Z qualitative under the null hypothesis with $\Pi_z(b_1) = \Pi_z(b_2) = \Pi_z(b_3) = 1/3$. Note: the values of $\Pi_z(b_i)$, for $i = 1, 2, 3$, do not have any influence on the results.

n	$(\Pi_z(b_1), \Pi_z(b_2), \Pi_z(b_3))$	$p(b_1)$	$p(b_2)$	$p(b_3)$	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.3402	0.3408
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.1680	0.1606
100	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.5588	0.5600
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.2748	0.2690
200	$(1/3, 1/3, 1/3)$	0.5	0.2	0.7	0.8028	0.7994
	$(3/5, 1/5, 1/5)$	0.5	0.2	0.7	0.4552	0.4448

Table 5: Power of the test for Case 1 with Z qualitative under the alternative hypothesis.

Case 2:
In Case 2, $\mathbf{W} = (X, \mathbf{Z})$ has dimension $m+1$, with a one-dimensional covariate X and a m-dimensional covariate $\mathbf{Z}$. For the sake of simplicity, in this simulation study we suppose that Z is also one-dimensional. We consider two different models: Model 1 and Model 2, when either X or Z are continuous. If there are no continuous covariates involved, we only consider Model 1.

Model 1 Under the null hypothesis, $H_0 : E(\nu|X, Z) = 1 - p(X)$, the probability of uncure $p(x)$ is that in (8). Under the alternative, the incidence is $1 - p(x, z)$, where

$$p(x,z) = \frac{\exp(\beta_0 + \beta_1 x(1+\beta_2 z))}{1+\exp(\beta_0 + \beta_1 x(1+\beta_2 z))}, \tag{10}$$

with $\beta_0 = 0.476$, $\beta_1 = 0.358$ and $\beta_2 = 0.225$, and the latency is:

$$S_0(t|x,z) = \begin{cases} \frac{\exp(-\lambda(x,z)t) - \exp(-\lambda(x,z)\tau_0)}{1-\exp(-\lambda(x,z)\tau_0)},\ t \leq \tau_0 \\ 0,\ t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda\,(x,z) = \exp\left((x+z+20)/40\right)$.

Model 2 The probability of uncure, $p(x)$, under $H_0 : E(\nu|X, Z) = 1 - p(X)$, is that in (9). Under the alternative, the probability of uncure is:

$$p(x,z) = \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 (1+\beta_4 z)\right)}{1+\exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 (1+\beta_4 z)\right)},$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$, $\beta_3 = 0.0020$ and $\beta_4 = 0.5$, and the survival function of the susceptible population is

$$S_0(t|x,z) = \frac{1}{2}\left(\exp(-\alpha(x,z)t^5) + \exp(-100t^5)\right),$$

with

$$\alpha(x,z) = \frac{1}{5}\exp((x+z+20)/40).$$

***X* and *Z* continuous:**
We consider two continuous covariates X, Z with distribution $U(-20, 20)$. Since we do not have a bandwidth selection method for h in (6), we consider the grid of bandwidths $h = Cn^{-1/3}$, where $C = 10, 20, 40$ and 60. Our numerical experience shows that $g = 2h$ is a good choice for the pilot bandwidth. Table 6 shows that, under the null hypothesis, the results are very similar to the significance level, α=0.05, except for very large bandwidths. Furthermore, under the alternative hypothesis, the power of the test is considerably high (see Table 7).

***X* and *Z* qualitative:**
Let both X and Z be qualitative variables, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. The cure probabilities $1 - p(a_i, b_j)$, with $i, j = 1, 2, 3$ are computed from the function $p(x, z)$ in (10). In Tables 8 and 9 we can see the different scenarios under H_0 and H_1, respectively. We work with two different situations depending on the corresponding probability mass functions: in the first one, both for X and Z are $(1/3, 1/3, 1/3)$, whereas in the second one, both are $(3/5, 1/5, 1/5)$.

Table 10 shows the results under the null hypothesis. In Scenario 1, the results are close to the significance level ($\alpha = 0.05$), except for CvM test, with $n = 50, 100$ and $(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3)) = (1/3, 1/3, 1/3)$. In the other 3 scenarios, the results are very competent regardless the probability mass function of X. The results under the alternative hypothesis are shown in Table 11. Note that in the 4 scenarios, the power is higher if we consider that the probability mass function of X is $(3/5, 1/5, 1/5)$.

		Model 1		Model 2	
n	h	CvM	KS	CvM	KS
50	2.71	0.0504	0.0524	0.0494	0.0500
	5.43	0.0428	0.0436	0.0464	0.0444
	10.86	0.0372	0.0486	0.0448	0.0438
	16.28	0.0442	0.0782	0.0692	0.0734
100	2.15	0.0494	0.0514	0.0526	0.0520
	4.31	0.0476	0.0556	0.0544	0.0542
	8.62	0.0580	0.0708	0.0514	0.0490
	12.93	0.0590	0.0988	0.0700	0.0782
200	1.71	0.0436	0.0412	0.0510	0.0496
	3.42	0.0420	0.0448	0.0558	0.0542
	6.84	0.0522	0.0606	0.0508	0.0494
	10.26	0.0590	0.0948	0.0614	0.0708

Table 6: Size of the test for Case 2 with X and Z continuous with distribution $U(-20, 20)$, under the null hypothesis.

		Model 1		Model 2	
n	h	CvM	KS	CvM	KS
50	2.71	0.2016	0.2182	0.2632	0.2466
	5.43	0.2596	0.2698	0.3244	0.3074
	10.86	0.2610	0.2696	0.3210	0.3150
	16.28	0.2278	0.2442	0.3084	0.3152
100	2.15	0.3906	0.4736	0.6228	0.5662
	4.31	0.5132	0.5556	0.6918	0.6128
	8.62	0.5160	0.5564	0.6864	0.6320
	12.93	0.4718	0.5206	0.6746	0.6438
200	1.71	0.7428	0.8554	0.9730	0.9364
	3.42	0.8492	0.9156	0.9832	0.9486
	6.84	0.8582	0.9114	0.9830	0.9568
	10.26	0.8358	0.8914	0.9852	0.9576

Table 7: Power of the test for Case 2 with X and Z continuous with distribution $U(-20, 20)$, under the alternative hypothesis.

	(b'_1, b'_2, b'_3)	(b'_1, b'_2, b'_3)
Scenario 1	$(-3.5434, -3.4427, -3.6537)$	$(-2.2879, -1.7995, -2.0064)$
$a'_1 = -6.5585$	0.5000	0.3400
$a'_2 = -1.1678$	0.5943	0.5566
$a'_3 = 0.9109$	0.6304	0.6581
Scenario 2	$(-3.5434, -3.4350, -3.6537)$	$(-2.2879, -2.0006, -1.995)$
$a'_1 = -6.5585$	0.5000	0.3400
$a'_2 = -4.5690$	0.5261	0.3957
$a'_3 = 0.9109$	0.6304	0.6583
Scenario 3	$(-3.5434, -3.4350, -4.0960)$	$(-2.2879, -2.0006, -2.6671)$
$a'_1 = -6.5585$	0.5000	0.3400
$a'_2 = -4.5690$	0.5261	0.3957
$a'_3 = 4.2229$	0.6444	0.7466
Scenario 4	$(-3.5434, -3.4350, -3.3932)$	$(-2.2879, -2.0006, -1.8585)$
$a'_1 = -6.5585$	0.5000	0.3400
$a'_2 = -4.5690$	0.5261	0.3957
$a'_3 = -3.2466$	0.5501	0.4501

Table 8: Uncure probabilities considered under H_0, $p(a_i, b_j) = p(a_i)$ for Case 2 when X and Z are qualitative, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively, when the probability mass functions for both X and Z are $(1/3, 1/3, 1/3)$ (first column), and $(3/5, 1/5, 1/5)$ (second column). Here, a'_i and b'_j denote the values at with the function p was evaluated to obtain $p(a_i, b_j)$.

	$b'_1 = 0.6157$	$b'_2 = -3.5434$	$b'_3 = -7.7026$
Scenario 1 (Least favorable situation)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -1.1678$	0.5	0.5966	0.6862
$a'_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 2 (Intermediate situation)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.2	0.5360	0.8423
$a'_3 = 0.9109$	0.7	0.6323	0.5590
Scenario 3 (Favorable situation a)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.2	0.5360	0.8423
$a'_3 = 4.2229$	0.9	0.6862	0.3470
Scenario 4 (Favorable situation b)			
$a'_1 = -6.5585$	0.1	0.5000	0.9000
$a'_2 = -4.5690$	0.2	0.5360	0.8423
$a'_3 = -3.2466$	0.3	0.5598	0.7905

Table 9: Uncure probabilities, $p(a_i, b_j)$, with $i, j = 1, 2, 3$, considered under H_1 in the four scenarios for Case 2 when X and Z are qualitative, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively. Here, a'_i and b'_j denote the values at with the function p was evaluated to obtain $p(a_i, b_j)$.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.0392	0.0490	0.0452	0.0548	0.0480	0.0558	0.0432	0.0526
	$(3/5, 1/5, 1/5)$	0.0496	0.0528	0.0490	0.0510	0.0524	0.0524	0.0530	0.0512
100	$(1/3, 1/3, 1/3)$	0.0392	0.0512	0.0500	0.0552	0.0472	0.0562	0.0452	0.0546
	$(3/5, 1/5, 1/5)$	0.0524	0.0502	0.0454	0.0496	0.0520	0.0546	0.0514	0.0512
200	$(1/3, 1/3, 1/3)$	0.0478	0.0552	0.0496	0.0588	0.0470	0.0496	0.0486	0.0526
	$(3/5, 1/5, 1/5)$	0.0496	0.0520	0.0540	0.0544	0.0532	0.0514	0.0506	0.0490

Table 10: Size of the test under H_0 for Case 2 with X and Z qualitative, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively, for the different scenarios. The probability mass function of Z equals that of X.

		Scenario 1		Scenario 2		Scenario 3		Scenario 4	
n	$(\Pi_x(a_1), \Pi_x(a_2), \Pi_x(a_3))$	CvM	KS	CvM	KS	CvM	KS	CvM	KS
50	$(1/3, 1/3, 1/3)$	0.1658	0.1802	0.2968	0.3224	0.2410	0.2780	0.4586	0.5164
	$(3/5, 1/5, 1/5)$	0.4874	0.4818	0.5054	0.5104	0.5056	0.5126	0.5482	0.5740
100	$(1/3, 1/3, 1/3)$	0.2932	0.2888	0.5358	0.5478	0.4556	0.5206	0.7584	0.7896
	$(3/5, 1/5, 1/5)$	0.7324	0.7220	0.7536	0.7612	0.7500	0.7616	0.7774	0.8060
200	$(1/3, 1/3, 1/3)$	0.5128	0.4854	0.8232	0.8324	0.7350	0.8266	0.9578	0.9640
	$(3/5, 1/5, 1/5)$	0.9218	0.9142	0.9340	0.9382	0.9354	0.9408	0.9412	0.9516

Table 11: Power of the test under H_1 for Case 2 with X and Z qualitative, with values $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively, for the different scenarios. The probability mass function of Z equals that of X.

4. APPLICATION TO REAL DATA

We study a dataset related to patients with sarcomas, provided by Angel Díaz-Lagares, postdoc researcher in Cancer Epigenomics from Translational Medical Oncology (OMT) group, Health Research Institute of Santiago (IDIS) and the University Hospital of Santiago (CHUS). The database consists of 261 observations with 32 covariates related to clinical data, which are detailed in Table 12. The event of interest is the death due to the appearance of sarcomas in different parts of the body. A total of 195 observations are censored, which corresponds to 74.71% of the data.

The results show that only two covariates are significant for the cure rate: "mpnst neurofibromatosis heredity" and "discontiguous lesion count".

REFERENCES

Beran, R. (1981) Nonparametric regression with randomly censored survival data. Technical report. Berkeley, University of California.

Delgado, M. and González-Manteiga, W. (2001) Significance testing in nonparametric regression based on bootstrap. The Annals of Statistics, 29, 1469–1507.

López-Cheda, A. and Cao, R. and 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. and Jácome, M. A. and Cao, R. (2017b) Nonparametric latency estimation for mixture cure models. TEST, 26, 353–376.

Xu, J. and Peng, Y. (2014) Nonparametric cure rate estimation with covariates. Canadian Journal of Statistics, 42, 1–17.

Name	Type	Values
days to birth	Continuous	Between 7530 and 32873
gender	Binary	male, female
race	Qualitative	asian, black or african american, white
ethnicity	Binary	hispanic or latino, not hispanic or latino
other diagnosis	Binary	yes, no
radiation therapy	Binary	yes, no
pharmaceutical adjuvant postop.	Binary	yes, no
histological type	Qualitative	dedifferentiated liposarcoma, leiomyosarcoma (LMS), undifferentiated pleomorphic sarcoma (UPS), synovial sarcoma, malignant peripheral nerve sheath tumors (MPNST), myxofibrosarcoma, desmoid tumor
leiomyosarcoma histologic subtype	Qualitative	poorly differentiated, conventional, well-differentiated
mpnst neurofibromatosis heredity	Binary	familial, sporadic
tumor depth	Binary	Values: deep, superficial
year of initial pathologic diagnosis	Continuous	Between 1994 and 2013
age at initial pathologic diagnosis	Continuous	Between 20 and 90
margin status	Binary	negative, positive
residual tumor	Qualitative	R0, R1, R2, RX
tumor total necrosis percent	Qualitative	no necrosis, focal, extensive, moderate
mitotic count	Continuous	Between 0 and 102
tumor multifocal	Binary	yes, no
discontiguous lesion count	Continuous	Between 0 and 9
radiologic tumor burden	Continuous	Between 2 and 800
pathologic tumor burden	Continuous	Between 1.6 and 800
local disease recurrence	Binary	yes, no
metastatic diagnosis	Binary	yes, no
metastatic site at diagnosis	Qualitative	bone, liver, lung, lung and liver, other
new tumor after initial treatment	Binary	yes, no
radiologic tumor length	Continuous	Between 1.8 and 30
radiologic tumor width	Continuous	Between 1.1 and 27.8
radiologic tumor depth	Continuous	Between 0.9 and 25
pathologic tumor length	Continuous	Between 1.2 and 39.5
pathologic tumor width	Continuous	Between 0 and 30
pathologic tumor depth	Continuous	Between 0 and 18
location	Qualitative	viscera, retroperitoneum, gynecological, upper extremity, lower extremity, head and neck

Table 12: Covariates with clinical information in the sarcomas dataset.

COMPARACIÓN DE MODELOS DE REGRESIÓN PARA LA DETERMINACIÓN DE LA VIDA ÚTIL DE PESCADO CONGELADO

López García, M.M.[1]; Ramil Novo, L.A.[2]; Romero Rodríguez, M.A.[1] y Vázquez Odériz, M.L.[1]

[1] Áreas de Nutrición e Bromatoloxía e de Tecnoloxía de Alimentos, Grupo GASFQ, Dpto. de Química Analítica, Nutrición e Bromatoloxía, Facultade de Ciencias, Universidade de Santiago de Compostela
[2] Grupo MODESTYA, Área de Estatística e Investigación Operativa, Dpto. de Estatística, Análise Matemática e Optimización, Facultade de Ciencias, Universidade de Santiago de Compostela

RESUMEN

En la Guía Visual para la Evaluación Sensorial de la Calidad del Pescado Congelado (Romero Rodríguez et al., 2014) se recogen los esquemas de puntuación correspondientes al Método del Índice de Calidad para diversos productos pesqueros de interés comercial a nivel español.

En este trabajo se examina cómo estimar la vida útil, x_0, cuando el producto considerado es descongelado a temperaturas de refrigeración para su consumo. En primer lugar, se analiza la relación entre el índice de calidad QI (Y) y los días en refrigeración (X) haciendo uso de dichos esquemas. Un análisis de regresión exhaustivo permite concluir que el modelo de regresión potencial ($Y=aX^b$) es generalmente el que mejor describe la relación entre Y y X en base al criterio de menor cuadrado medio residual. En segundo lugar, se examina la precisión de la estimación de la vida útil mediante un intervalo de calibración para x_0 (Lavagnini y Magno, 2007). Para determinar los límites se consideran el método de inversión y el método de Wald, que han sido implementados recientemente en el programa estadístico R (Greenwell y Schubert Kabban, 2014). Los intervalos de calibración obtenidos con el modelo potencial, cuyo cálculo requiere el uso del método de mínimos cuadros no lineales y de software estadístico, se comparan con los obtenidos con dos modelos lineales cuyos límites admiten fórmulas explícitas. Por un lado, se considera el modelo lineal simple, $Y=a+bX$, de uso común en la literatura científica previa y, por otro lado, un modelo lineal en términos de la raíz del número de días en refrigeración, $Y=a+bX^{1/2}$, modelo motivado por resultados obtenidos con el modelo potencial. Finalmente, se muestra que tanto el modelo potencial como el basado en la raíz $X^{1/2}$ también proporcionan buenos resultados para estimar la vida útil de diversas especies pesqueras frescas, lo que permite anticipar la utilidad de la metodología propuesta en otros contextos.

Palabras y frases clave: regresión; error cuadrático medio; intervalos de calibración.

AGRADECIMIENTOS

Este estudio se llevó a cabo gracias al proyecto de investigación desarrollado en colaboración con la Fundación CETMAR-Centro Tecnológico del Mar, Vigo (Pontevedra), de título “Establecimiento de umbrales de calidad (sensoriales y químicos) para especies objetivo pesquero no contempladas en el marco legislativo” (Referencia 10TAL010CT, Xunta de Galicia, D.O.G. nº. 246 de 24/12/2010).

REFERENCIAS

Greenwell, B.M. y Schubert Kabban, C.M. (2014). investr: An R Package for inverse estimation. *The R Journal*, Vol. 6/1, 90-100.

Lavagnini, I. y Magno, F. (2007). A statistical overview on univariate calibration, inverse regression, and detection limits: application to gas chromatography/mass spectrometry technique. *Mass Spectrometry Reviews*, 26, 1-18.

Romero Rodríguez, M.A., Vázquez-Odériz, L., López García, M.M., Ramil-Novo, L.A., Maroto Leal, J., Estévez Calvar, N., Sieiro Piñeiro, P., Abollo Rodríguez, E., Rodríguez Bernaldo de Quirós, A., Sendón García, R., Paseiro Losada, P., Otero Pazos, P., Cruz Freire, J.M., Barbosa Pereira, L., Vecino Bello, X. y Moldes Menduiña, A.B. (2014). *Guía visual para la evaluación sensorial de la calidad del pescado congelado*. Centro Tecnológico del Mar – Fundación CETMAR.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

ESTRATEXIA DE XEORREFERENCIACIÓN DA POBOACIÓN DE GALICIA EMPREGANDO TÉCNICAS ESTATÍSTICAS

Mª Esther López Vizcaíno[1] e Carlos L. Iglesias Patiño[1]

[1] Instituto Galego de Estatística

RESUMO

A práctica totalidade das estatísticas teñen unha base territorial e, polo tanto, xeográfica e a integración entre a información estatística e xeográfica é imparable, xa que permite sinerxías capaces de responder a preguntas que ata hai pouco non tiñan resposta. O obxectivo xeral deste relatorio é presentar un exemplo da integración entre ambos os dous tipos de información, estatística e xeográfica, que mostra a importancia e o potencial da xeorreferenciación da información. O obxectivo concreto é xeolocalizar a poboación de Galicia dispoñible no Padrón Municipal de Habitantes (PMH), empregando para iso diversas técnicas estatísticas.

Palabras e frases chave: xeorreferenciación, splines, análise de compoñentes principais, distancia

1. INTRODUCIÓN

Hai douscentos anos o matemático Domingo Fontán fixou a posición xeográfica da Torre do Reloxo da Catedral de Compostela para poder levantar a súa Carta Geométrica de Galicia, isto é, puxo a Galicia no mapa. Agora estamos tentando colocar os seus habitantes nel.

A práctica totalidade das estatísticas teñen unha base territorial e, polo tanto, xeográfica e a integración entre a información estatística e xeográfica é imparable, xa que permite sinerxías capaces de responder a preguntas que ata hai pouco non tiñan resposta.

Historicamente, o tratamento dado aos enderezos postais conducía á asignación de divisións territoriais, como é o caso da poboación en concellos, parroquias ou seccións censuais. O efecto final desta asignación é que os resultados estatísticos só poden obterse respecto destas zonas ou os seus agregados, producíndose unha perda de información non desexable. A solución a este problema pasa necesariamente pola asignación dunhas coordenadas aos enderezos postais existentes nas diferentes fontes de información. Así, a posterior agregación espacial non estará limitada a priori e respectaranse as identidades aritméticas (determinadas sumas deben coincidir cos totais xa dispoñibles) o que constitúe unha boa práctica nunha estatística pública. Con todo isto, a difusión veríase enriquecida.

Polo tanto, o obxectivo final da xeorreferenciación é indentificar mediante coordenadas xeográficas onde se sitúan as empresas, vivendas, persoas, etc. As utilidades desta representación son múltiples: na poboación residente por zonas territoriais, na planificación de servizos ou equipamentos, na planificación do transporte, no cálculo da poboación en zonas de influencia de escolas, residencias, etc.

Tendo en conta todo o anterior, o obxectivo deste relatorio é presentar un exemplo da integración entre ambos os dous tipos de información, estatística e xeográfica, que mostra a importancia e o potencial da xeorreferenciación da información. O obxectivo concreto é xeolocalizar a poboación de Galicia dispoñible no Padrón Municipal de Habitantes (PMH), empregando para iso diversas técnicas estatísticas.

2. FONTES E MÉTODOS

A información que se xeorreferenciará é a poboación de Galicia dispoñible no PMH de Galicia a 1 de xaneiro de cada ano. No PMH están dispoñibles o nome, os apelidos e o enderezo postal de todos os habitantes que teñen residencia na Comunidade Autónoma de Galicia. O obxectivo deste traballo é asignar unhas coordenadas xeográficas aos enderezos existentes no PMH formados pola vía e o número do portal. É dicir, as persoas que viven no mesmo portal terán as mesmas coordenadas xeográficas.

A fonte principal que se empregará para xeorreferenciar será a base de datos proporcionada polo Instituto Nacional de Estadística (INE) co nome "Modelo de Direcciones de la Administración General del Estado" (MDAGE). Nesta base de datos:

- Están identificados todos os inmobles de uso residencial.
- O inmoble está identificado sempre do mesmo xeito, a partir dunha información estruturada e actualizada.
- Todos os enderezos están localizados espacialmente (xeorreferenciados), coas súas coordenadas UTM X, Y no fuso 30
- Preséntase como modelo de referencia para a harmonización, integración, intercambio e consulta de enderezos en toda España

Como Galicia está no fuso 29, o primeiro que se fixo foi transformar as coordenadas UTM do fuso 30 ao 29. O traballo que se terá que resolver a partir deste punto é como asignar as persoas do PMH aos inmobles do MDAGE e conseguir así obter unha coordenada xeográfica da persoa, para cada un dos habitantes de Galicia.

As variables que se empregan para xeorreferenciar son:

- Cvia: código da vía de comunicación construída para a circulación.
- Tvia: tipo de vía
- Nvia: nome da vía
- Num: número de portal
- Cec: código de entidade colectiva
- Ces: código de entidade singular
- Cnuc: código de núcleo ou diseminado
- Dis: código de distrito censual
- Secc: código de sección censual

Neste punto hai que resaltar que a base de datos MDAGE non é completa, existen inmobles que non teñen unhas coordenadas UTM válidas ou, simplemente, que non están xeorreferenciados. De todos os xeitos, esta será a fonte de información principal, aínda que despois se complete coas seguintes fontes de información auxiliares:

- Cartociudad: proxecto colaborativo liderado polo Instituto Geográfico Nacional (IGN) de produción e publicación mediante servizos web de datos espaciais de cobertura nacional. Contén información da rede viaria continua, coas rúas con portais, e estradas con puntos kilométricos. Dispón dun servizo de procesamento web baseado en cálculos programados que operan sobre a información xeorreferenciada. Este servizo empregarase para xeorreferenciar enderezos non dispoñibles na MDAGE.

- Dirección General del Catastro: dispón de servizos web que permiten a descarga da cartografía catastral en formato vectorial e información alfanumérica, así como servizos web para a descarga de enderezos e coordenadas. Este servizo empregarase para xeorreferenciar enderezos non dispoñibles na MDAGE nin en Cartociudad.

- Google: para aqueles enderezos que non se poidan xeorreferenciar coa MDAGE, Cartociudad ou a Dirección General de Catastro, empregarase este servizo para buscar coordenadas xeográficas entregándolle un enderezo postal.

O procedemento comeza por xeorreferenciar as persoas que viven en rúas que teñen dispoñible o código de vía, tipo de vía, nome de vía e número, que corresponde principalmente ás zonas urbanas (Paso 1.1). Neste caso crúzase a base de datos MDAGE co PMH empregando as variables anteriores e asígnaselle a cada persoa as coordenadas correspondentes. Neste proceso xorde o problema de que hai rúas que teñen código de vía pero o número do portal do PMH non está dispoñible no MDAGE. Neste caso hai que imputar as coordenadas X, Y correspondentes a este novo número de vía. Optouse por aplicar a regresión con splines ás coordenadas da vía especificada, e terase en conta se os números do

portal son pares ou impares. A vantaxe de empregar os splines é que poden representarse como modelos mixtos, e polo tanto benefícianse da teoría e do software existente sobre este tipo de modelos (Ruppert et al., 2003, Ugarte et al., 2009, Wood, 2006). O modelo que se aplicou, para cada vía, foi o seguinte:

$$\mathbf{Y_i=f(X_i)+aI_i+e_i} \qquad \mathbf{(1)}$$

onde

Y é a coordenada xeográfica UTM Y no Fuso 29,
X é a coordenada xeográfica UTM X no Fuso 29
I é unha variable indicadora que toma o valor 1 se o número é par e o valor 0 se o número é impar.
e sigue unha variable aleatoria $N(0,\sigma^2)$
f é unha función suave
i=1, ..., nº total de portais dispoñibles da vía

Para o axuste deste modelo empréganse os splines penalizados.

Unha vez establecido este primeiro modelo necesítase un segundo modelo que nos permita calcular a coordenada xeográfica X dun novo portal que se vai imputar. A formulación para cada unha das vías neste segundo modelo é:

$$\mathbf{X_i=g(N_i)+bI_i+e_i'} \qquad \mathbf{(2)}$$

onde

X é a coordenada xeográfica UTM X no Fuso 29 ,
N e o número do portal do inmoble
I é unha variable indicadora que toma o valor 1 se o número é par e o valor 0 se o número é impar.
g é unha función suave
e' sigue unha variable aleatoria $N(0,\sigma'^2)$
i=1, ..., nº total de portais dispoñibles da vía

Para o axuste deste segundo modelo tamén se empregarán os splines penalizados.

Polo tanto, o procedemento para obter as coordenadas X, Y para un número de portal non dispoñible no MDAGE nunha vía específica é o seguinte:

- Coas coordenadas xeográficas X, Y que nos achega a MDAGE para unha vía axustamos o modelo 1.
- Coa coordenada xeográfica X e o número de portal que nos achega MDAGE para unha vía, axustamos o modelo 2.
- Para predicir a coordenada xeográfica X, empregamos o número que queremos predicir e o modelo 2.
- Unha vez que se obtén a coordenada xeográfica X do número que imos predicir, aplicamos o modelo 1 para obter a coordenada xeográfica Y.

No caso de que non haxa suficientes números de portais dispoñibles na vía para aplicar a regresión, optouse por asignarlle a coordenada do número de portal máis próximo.Para as persoas que quedan sen xeorreferenciar no paso anterior empregarase a información que proporciona Cartociudad (Paso 1.2).

Para as persoas que aínda quedan sen xeorreferenciar tras o paso anterior empregarase a información que proporciona a Dirección General de Catastro, en concreto a información da cartografía urbana dixital en formato shape. Esta cartografía ofrece as coordenadas das parcelas onde se ubican os inmobles. O problema que ofrecen estas bases de datos é que non existe un nexo de unión entre os ficheiros do PMH e os ficheiros do Catastro. Como nos dous casos se dispón dos enderezos dos inmobles, o que se fará será unir por enderezos tratando de vincular as vías que nos proporciona o PMH coas vías que proporciona a Dirección General de Catastro nos seus arquivos. Para isto empregarase un procedemento baseado en unir táboas mediante variables que non teñen unha coincidencia exacta mediante o paquete de R ***fuzzyjoin***. Este paquete utiliza un procedemento baseado en distancias entre cadeas de carácteres, en concreto utilizouse a distancia de Levenshtein. Esta distancia entre dúas cadeas vén medida polo número de carácteres que hai que eliminar, incluír ou substituír para ir dunha cadea a outra. Canto menor é a distancia de Levenshtein entre dúas cadeas de carácteres, máis semellantes son. Tamén empregaremos o mapa de seccións censuais, de tal maneira que a vía do Catastro ten que estar na mesma sección censual que a vía dispoñible no PMH (Paso 1.3).

Unha vez feita a correspondencia entre o código de vía do PMH e o código de vía do Catastro, é posible que no Catastro non estean todos os números de portal dispoñibles no PMH. Neste caso poden ocorrer dúas situacións diferentes (Paso 1.4):

1. Algúns números de portais que achega Catastro son nulos

   Se os números de portal que achega Catastro teñen ceros non se pode saber se se trata de números pares ou impares. Para averiguar isto empregaremos a análise de compoñentes principais (ACP). Aplicarase a ACP ás coordenadas X, Y que nos achega o Catastro para a vía. Pódese comprobar que a primeira compoñente principal correspóndese coa dirección da vía e a segunda, ortogonal á primeira, cos pares e impares. Unha vez averiguado cales son os puntos pares e impares, xa podemos calcular o centroide das coordenadas dos pares e dos impares e asignarllo ao número de portal que se quere xeorreferenciar, dependendo de que sexa par ou impar.

2. Ningún número de portal é nulo

   Nesta situación para determinar as coordenadas do número novo aplicaremos a regresión con splines que se utilizou no primeiro punto á vía especificada.

   Por último, é necesario xeorreferenciar as persoas que viven en entidades que non teñen dispoñible no PMH a clase de vía, código ou nome de vía, que se corresponde principalmente coa zona rural. Neste caso a xeorreferenciación realízase a nivel de núcleo/diseminado e número do portal (do inmoble). Para aquelas persoas que teñen un número de portal no PMH e este número está dispoñible no MDAGE, asígnaselle a coordenada dispoñible no MDAGE. Para aquelas persoas que viven en inmobles que teñen un número de portal no PMH e este número non está dispoñible no MDAGE asígnaselle a media ponderada das coordenadas do núcleo/diseminado ao que pertencen segundo a metodoloxía presentada no relatorio García, Iglesias e López (2017) (Paso 2.1).

Para as persoas que quedan sen xeorreferenciar no paso anterior asignóuselles o centroide da entidade singular á que pertencen. Para isto seguiuse o seguinte procedemento:

- Para aquelas entidades singulares das que xa se tiña persoas xeorreferenciadas, asignóuselle o punto medio das persoas que xa estaban xeorreferenciadas nestas entidades singulares (Paso 2.2).
- Para aquelas entidades singulares das que non se tiña persoas xeorreferenciadas, asignóuselle o centroide da entidade singular do mapa destas entidades que nos proporcionou o Instituto de Estudos do Territorio (IET), elaborado a partir da información do Censo de Vivendas do ano 2011 e doutros datos dispoñibles nese instituto (Paso 2.3).
- Para as persoas das restantes entidades singulares, asignóuselle o centroide da parroquia á que pertencen (Paso 2.4).

3. RESULTADOS

O procedemento anterior aplicouse a todos os concellos de Galicia e ao PMH do ano 2016. Das 2.718.525 persoas dispoñibles no PMH, xeorreferenciáronse 2.693.562 persoas (Táboa 1). No paso 1.1 e 2.1 un total de 190.763 (7%) non foi posible xeolocalizalas. No paso 1.1 quedaron sen xeorreferenciar 43.557 persoas e no paso 2.1 quedaron sen xeorreferenciar 147.206 persoas.

Urbano (con vía)				
Paso	**Persoas**	**Con xeorreferenciación**	**Sen xeorreferenciación**	**% sen xeorreferenciación**
1.1	1.977.813	1.934.256	43.557	2,20
1.2	43.557	13.169	30.388	1,54
1.3	30.388	1.802	28.586	1,45
1.4	28.586	1.646	26.940	1,36
1.5	26.940	450	26.490	1,34
1.6	26.490	1.606	24.884	1,26
Rural (sin vía)				
2.1	740.712	593.506	147.206	19,87
2.2	147.206	4.678	142.528	19,24
2.3	142.528	129.562	12.966	1,75
2.4	12.966	12.887	79	0,01
Total	**2.718.525**	**2.693.562**	**24.963**	**0,01**

Táboa 1. Persoas xeorreferenciadas en cada un dos pasos do procedemento

No paso 1.3, chamando aos servizos de Cartociudad púidose xeorreferenciar 13.169 persoas. Quedan, polo tanto, sen xeorreferenciar 30.388 persoas. Para estas, acudiuse á información xeorreferenciada que publica a Dirección General de Catastro completada coa información de Google. Conseguiuse xeorreferenciar a 5.504 persoas.

No caso das 147.206 persoas sen xeorreferencir do paso 2.1, a 4.678 asignóuselle o punto central das persoas xa referenciadas no paso 2.1, a 129.562 asignóuselle o centroide da entidade singular proporcionada polo IET, a 12.887 asignóuselle o centroide da parroquia e 79 non se conseguiron clasificar.

No Gráfico 1, preséntase unha imaxe das persoas xeorreferenciadas en Galicia seguindo o procedemento descrito.

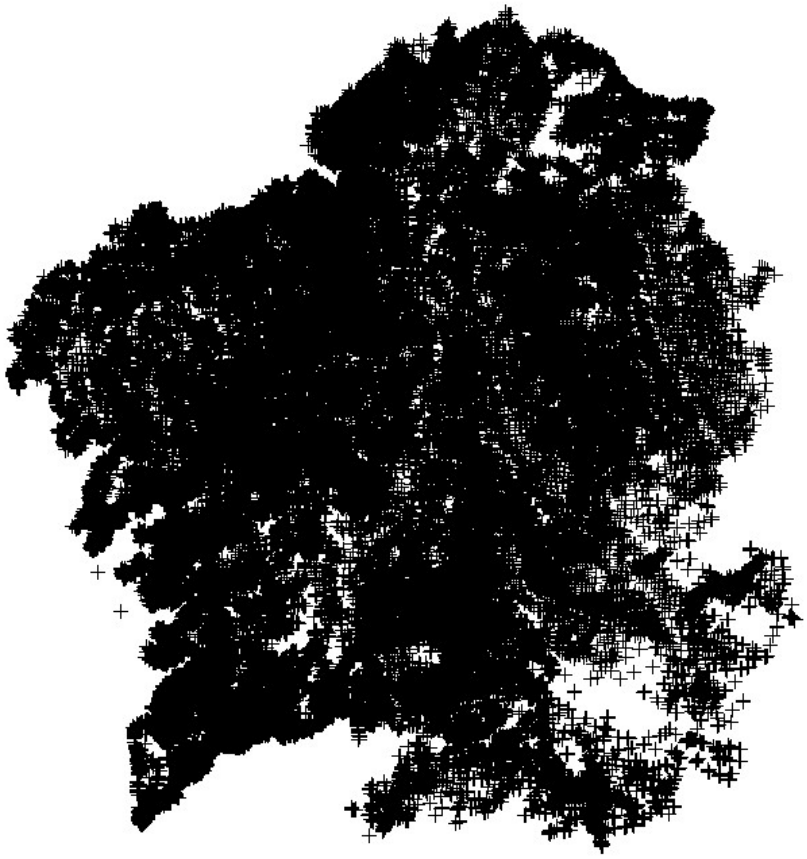


Gráfico 1. Representación da poboación xeorreferenciada

REFERENCIAS

INE (2013). Censos de población y viviendas 2011. http://www.ine.es/censos2011_datos/cen11_datos_inicio.htm

INE (2016). Modelo de Direcciones de la Administración general del Estado.

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/.

Ruppert, D., Wand, M.P., Carrol, R.J. (2003*). Semiparametric Regression*. Cambridge University Press, New York.

Ugarte, M, Goicoa T., Militino, A y Durbán, M (2009). Spline smoothing in small área trend estimation and forecasting. *Computational Statistics and Data Analysis*., 53: 3616-3629.

Wood, S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

A testing method for linear spatial trends

Andrea Meilán-Vila[1], Mario Francisco-Fernández[1] and Rosa M. Crujeiras[2]

[1] Departamento de Matemáticas, Universidade da Coruña
[2] Departamento de Estadística, Análisis Matemático y Optimización, Universidade de Santiago de Compostela

ABSTRACT

The aim of this work is to propose and analyze the behavior of a test statistic for assessing a parametric trend surface, that is, a regression model with spatially correlated errors. Specifically, linear trend surfaces will be tested. The finite sample performance of the test is addressed by a simulation study.

Keywords: Model checking; Spatial trend; Smoothing; Least squares; Nonparametric bootstrap.

1. INTRODUCTION

This work is framed in the context of spatial stochastic processes, which consists of collections of random variables indexed on a certain domain of $\mathbb{R}^d$, with a well-defined joint distribution. The observed data tend to exhibit an important feature: close observations tend to be more similar than observations which are far apart. Therefore, such observations cannot be treated as independent and the dependence structure should be taken into account in any descriptive or inferential procedure. In particular, from the perspective of regression models (trend surfaces), the dependence structure should be considered and properly introduced into the model. Estimation procedures in this context are usually based on iterative least squares methods (with a prior estimation of the dependence structure, but as a kind of nuisance to be incorporated in the model) or on maximum likelihood approaches. Nevertheless, the problem of testing parametric models has not been (up to our knowledge) considered in the statistical literature.

In Härdle and Mammen (1993), a testing procedure to check if a regression function belongs to a class of parametric models by measuring the discrepancy between nonparametric and parametric regression fits is proposed. They consider independent observations and the Nadaraya-Watson estimator. The same type of study is performed in Alcalá et al. (1999), also under independent observations, but using local polynomials. Kaiser et al. (2012) have developed goodness of fit statistics that can be used to formally assess Markov random field models for spatial data, when the model distributions are discrete or continuous and potentially parametric.

A new proposal for testing a parametric trend surface will be given in this work. We consider that the data are given by a spatial stochastic process:

$$\{Z(\mathbf{s})|\mathbf{s} \in D\}, \tag{1}$$

consisting of collections of random variables indexed in a particular domain $D \subset \mathbb{R}^d$ (observation region) with a well defined joint distribution. Considering n locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ on the region D, the set of random variables corresponding with those locations will be represented by $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$, while $\{z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)\}$ will denote a realization of this set.

In this context, we will consider the following spatial regression model:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \varepsilon(\mathbf{s}) = X(\mathbf{s})'\beta + \varepsilon(\mathbf{s}), \tag{2}$$

where $X(\cdot) \in \mathbb{R}^q$ denote the spatial regressors. The function $m(\cdot)$ denotes the deterministic (not random) part and represents large-scale changes, while $\varepsilon(\cdot)$ denotes the random component and

shows the local behavior or small-scale evolution. The error process $\varepsilon(\cdot)$ is assumed to be zero-mean, so that the whole trend is captured by $m(\cdot)$. Therefore, $m(\cdot)$ would give the first order structure (mean) and $\varepsilon(\cdot)$ explains the second-order part (as the mean is zero, what contributes is the part of covariance).

For these processes, the first moment is the expectation, while the three second order moments considered are the variance or second order moment of $Z(\mathbf{s})$ with respect to $m(\mathbf{s})$, $\sigma^2(\mathbf{s}) = \mathrm{Var}[Z(\mathbf{s})] = \mathbb{E}\{[Z(\mathbf{s}) - m(\mathbf{s})]^2\}$, the covariance of two random variables $Z(\mathbf{s})$ and $Z(\mathbf{s}')$ (which is just function of $\mathbf{s}$ and $\mathbf{s}'$), $C(\mathbf{s}, \mathbf{s}') = \mathrm{Cov}[Z(\mathbf{s}), Z(\mathbf{s}')] = \mathbb{E}\{[Z(\mathbf{s}) - m(\mathbf{s})][Z(\mathbf{s}') - m(\mathbf{s}')]\}$, and the variogram which is defined as the variance of the difference process $2\gamma(\mathbf{s}, \mathbf{s}') = \mathrm{Var}[Z(\mathbf{s}) - Z(\mathbf{s}')]$. If a spatial process is second-order stationary then $\mathbb{E}[Z(\mathbf{s})] = m$, $\forall \mathbf{s} \in D$ and $C(\mathbf{s}, \mathbf{s}') = C(\mathbf{s} - \mathbf{s}')$, $\forall \mathbf{s}, \mathbf{s}' \in D$, and the function $C(\cdot)$ is called covariogram or covariance function. The form of stationarity implied by the intrinsic hypothesis is essentially second-order stationarity, not for the process $Z(\mathbf{s})$, but for the difference, $Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})$: $\mathbb{E}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})) = 0$, $\forall \mathbf{u} \in \mathbb{R}^d$, $\forall \mathbf{s}, \mathbf{s}+\mathbf{u} \in D$ and $\mathrm{Var}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})) = 2\gamma(\mathbf{u}), \forall \mathbf{u} \in \mathbb{R}^d, \forall \mathbf{s}, \mathbf{s}+\mathbf{u} \in D$ (the function $2\gamma(\cdot)$ has been called variogram). In this case, we are going to assume that the error process $\varepsilon(\cdot)$ is second-order stationary (that is to say, the covariance function is just a function of the difference vector between the locations where the observations are taken). Note that if the process $Z(\mathbf{s})$ is second-order stationary then $C(u) = \sigma^2 - \gamma(u)$.

The goal of this work is to test if the trend function belongs to a parametric family:

$$H_0 : m(\cdot) \in \mathcal{M}_{\boldsymbol{\beta}} = \{m_{\boldsymbol{\beta}}(\cdot), \boldsymbol{\beta} \in \mathcal{B}\}, \qquad \text{vs.} \qquad H_a : m(\cdot) \notin \mathcal{M}_{\boldsymbol{\beta}}. \tag{3}$$

Fixed design will be supposed and the trend component $m_{\boldsymbol{\beta}}(\cdot)$ will be linear. To tackle the test, one of the more usual approaches is to compare a smooth version of a parametric fit $\hat{m}_{\hat{\boldsymbol{\beta}}}(\cdot)$ with a nonparametric estimator $\hat{m}(\cdot)$ of $m(\cdot)$ and thereafter to reject H_0 if the distance $d_0(\hat{m}_{\hat{\boldsymbol{\beta}}}(\cdot), \hat{m}(\cdot))$ exceeds a critical value. As d_0 a weighted L_2-norm can be used such as that in Härdle and Mammen (1993), or a Cramer-von-Mises distance such as that in Stute and Manteiga (1996).

2. TREND ESTIMATION

The trend surface estimation can be performed parametrically and non-parametrically. In the parametric context, with the decomposition given in (2), it is not clear which structure (first order or second order) must be estimated first. In practice, iterative procedures are a solution, as in Diggle and Ribeiro (2007). Denoting $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))'$, $\hat{\mathbf{Z}} = (\hat{Z}(\mathbf{s}_1), \cdots, \hat{Z}(\mathbf{s}_n))$, $\mathbf{m} = (m(\mathbf{s}_1), \ldots, m(\mathbf{s}_n))'$, $\hat{\mathbf{m}} = (\hat{m}(\mathbf{s}_1), \ldots, \hat{m}(\mathbf{s}_n))'$ and $X = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))'$, the steps are:

(1) Unweighted least squares estimation of $\boldsymbol{\beta}$, ignoring the dependence structure of the errors:

$$\boldsymbol{\beta}_{OLS} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Z} - X\boldsymbol{\beta})'(\mathbf{Z} - X\boldsymbol{\beta}). \tag{4}$$

(2) Estimation of the variance-covariance matrix of the errors based on the residuals $\tilde{\varepsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - X(\mathbf{s}_i)\boldsymbol{\beta}_{OLS}$, $i = 1, \ldots, n$, obtained from the estimator of $\boldsymbol{\beta}$ found in the step (1):

$$\hat{\Sigma}(i,j) = C_{\widehat{\boldsymbol{\theta}_{LS}}}(\mathbf{s}_i - \mathbf{s}_j) = \hat{\sigma}^2 - \gamma_{\widehat{\boldsymbol{\theta}_{LS}}}(\mathbf{s}_i - \mathbf{s}_j), \quad i,j = 1 \ldots, n,$$

where $\gamma_{\widehat{\boldsymbol{\theta}_{LS}}}(\cdot)$ is the parametric least squares estimator of the variogram and $\hat{\sigma}^2$ is an estimator of the variance.

(3) Weighted least squares estimation of $\boldsymbol{\beta}$, taking the dependence structure of the errors into account:

$$\boldsymbol{\beta}_{IGLS} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Z} - X\boldsymbol{\beta})'\hat{\Sigma}^{-1}(\mathbf{Z} - X\boldsymbol{\beta}). \tag{5}$$

Note that the parametric estimator of the trend $\boldsymbol{\beta}_{IGLS}$ given in (5) will be denoted by $\hat{\boldsymbol{\beta}}$ from here on out. As for the nonparametric estimation, we consider the Nadaraya-Watson estimator, which is defined by

$$\hat{m}^{NW}(\mathbf{s}, \mathbf{H}) = \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{s}_i - \mathbf{s}) Z(\mathbf{s}_i)}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{s}_i - \mathbf{s})},$$

where $K_{\mathbf{H}}(\mathbf{s}) = |\mathbf{H}|^{-1}K(\mathbf{H}^{-1}\mathbf{s})$ is used to assign weights, $\mathbf{H}$ is a $d \times d$ symmetric, positive definite matrix depending on the sample size n and K is the kernel function. Given $\mathbf{s}$, the bandwidth $\mathbf{H}$ controls the shape and the size of the local neighborhood used for estimating $m(\cdot)$.

3. THE PROPOSED TEST

Let $\hat{m}^{NW}(\cdot)$ be the nonparametric Nadaraya-Watson estimator of $m(\cdot)$ and $\hat{\boldsymbol{\beta}}$ a $\sqrt{n}$-consistent estimator of the "true" parameter $\boldsymbol{\beta}_0$ under the null hypothesis. The test statistic is given by

$$T_n = n|\mathbf{H}|^{1/2} \int (\hat{m}^{NW}(\mathbf{s}, \mathbf{H}) - \hat{m}^{NW}_{\hat{\boldsymbol{\beta}}}(\mathbf{s}, \mathbf{H}))^2 w(\mathbf{s}) d\mathbf{s}, \tag{6}$$

where $w(\cdot)$ is a weight function and $\hat{m}^{NW}_{\hat{\boldsymbol{\beta}}}(\cdot)$ is a smooth version of $m_{\hat{\boldsymbol{\beta}}}(\cdot)$ which is defined by

$$\hat{m}^{NW}_{\hat{\boldsymbol{\beta}}}(\mathbf{s}, \mathbf{H}) = \frac{\sum_i K_{\mathbf{H}}(\mathbf{s}_i - \mathbf{s})\, m_{\hat{\boldsymbol{\beta}}}(\mathbf{s}_i)}{\sum_i K_{\mathbf{H}}(\mathbf{s}_i - \mathbf{s})},$$

where $m_{\hat{\boldsymbol{\beta}}}(\mathbf{s}_i) = X(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}$ for linear trends. Fixed design will be suppose and the trend component $m_{\boldsymbol{\beta}}(\cdot)$ will be linear.

Alternative definitions of the test statistic T_n are possible: the integral in (6) may be replaced by a sum. For a comparison of T_n with other goodness-of-fit tests, the asymptotic power of T_n if $m(\cdot)$ lies in the alternative should be computed. We could consider alternatives as $m(\mathbf{s}) = m_{\boldsymbol{\beta}_0}(\mathbf{s}) + c_n g(\mathbf{s})$ for certain sequences c_n and $g(\cdot)$. Note that, it is most appropriate choose c_n such that the asymptotic power of T_n is bounded away from 1 and from the level.

4. BOOTSTRAP CALIBRATION

Once a suitable test statistic is available, a crucial task is the calibration of critical values for a given level α, namely t_α. Usually, the estimation of these critical values c_α such that $\mathbb{P}_{H_0}(T_n \geq t_\alpha) = \alpha$ can be done by means of the asymptotic distribution. The use of asymptotic theory for calibration the test poses some problems such as the need to estimate some nuisance functions and a slow convergence rate to the limit distribution. Under these circumstances, calibration can be done by means of resampling procedures, such as bootstrap.

In particular, we are going to use a nonparametric bootstrap procedure, see Fernández-Casal et al. (2014). The procedure consists in generating a bootstrap sample $\{Z^*(\mathbf{s}_i), i = 1, \ldots, n\}$ and then computing a bootstrap statistic T_n^* like T_n by the squared deviation between the smooth version of the parametric fit $\hat{m}^{NW}_{\hat{\boldsymbol{\beta}}^*}(\cdot)$ and the nonparametric fit $\hat{m}^{*NW}(\cdot)$. Once the bootstrap statistic is computed, the distribution of T_n^* can be approximated by Monte Carlo. From this Monte Carlo approximation define the $(1-\alpha)$ quantile t_α^* and reject the parametric hypothesis if $T_n > t_\alpha^*$. The specific steps for the algorithm are the following:

1. Obtain the trend parameter estimator $\hat{\boldsymbol{\beta}}$.

2. Estimate the covariance matrix of the errors $\hat{\Sigma}$ based on the residuals $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}(\mathbf{s}_1), \ldots, \hat{\varepsilon}(\mathbf{s}_n))'$, where $\hat{\varepsilon}(\mathbf{s}_i) = Z(\mathbf{s}_i) - X(\mathbf{s}_i)\hat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$, and find the matrix L, such that $\hat{\Sigma} = LL'$, using Cholesky decomposition.

3. Compute the independent residuals, $\mathbf{e} = (e(\mathbf{s}_1), \ldots, e(\mathbf{s}_n))'$, given by $e(\mathbf{s}_i) = L^{-1}\hat{\varepsilon}(\mathbf{s}_i)$.

5. These independent variables are centered and, from them, we obtain an independent bootstrap sample of size n, denoted by $\mathbf{e}^* = (e^*(\mathbf{s}_1), \ldots, e^*(\mathbf{s}_n))$.

6. Finally, the bootstrap errors $\boldsymbol{\varepsilon}^* = (\varepsilon^*(\mathbf{s}_1), \ldots, \varepsilon^*(\mathbf{s}_n))$ are $\varepsilon^*(\mathbf{s}_i) = Le^*(\mathbf{s}_i)$, and the bootstrap samples are $Z^*(\mathbf{s}_i) = X(\mathbf{s}_i)'\hat{\boldsymbol{\beta}} + \varepsilon^*(\mathbf{s}_i)$.

5. SIMULATIONS

In order to study the performance of our procedure for goodness-of-fit test for spatial trends, we will carry out a simulation study considering the following spatial regression model, where $\mathbf{s}_i = (s_{i1}, s_{i2})$:

$$Z(\mathbf{s}_i) = \beta_0 + \beta_1 s_{i1} + \beta_2 s_{i2} + c s_{i1}^3 + \varepsilon(\mathbf{s}_i), \quad 1 \leq i \leq n. \tag{7}$$

Data samples with size $n = 100$ are generated from an isotropic Gaussian spatial process observed at regularly spaced locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in the unit square. Different values of the parameters β_0, β_1 and β_2 are chosen: $\boldsymbol{\beta} = (2, 1, 1)$ and $\boldsymbol{\beta} = (1, 0, 1)$. We have also considered an exponential covariogram to model the dependence structure:

$$C(u) = \begin{cases} c_e\{\exp(-u/a_e)\} & \text{if } u \neq 0, \\ c_0 + c_e & \text{if } u = 0, \end{cases}$$

with different values of their parameters: $\boldsymbol{\theta} = (c_0, c_e, a_e) = (0, 0.4, 0.6)$ and $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.8, 0.8)$. No nugget effect is considered in the simulation scenarios. The nonparametric bootstrap has been performed $B = 500$ times for each sample. We have replicated each simulation 1000 times and recorded the proportion of rejections. The bandwidth is chosen in each case by cross-validation criteria, see Bowman and Azzalini (1997). We are going to suppose that $c = 0$ in model (7), that is to say, we are going to check the correct calibration of the test. We are going to consider that $\alpha = 0.05$.

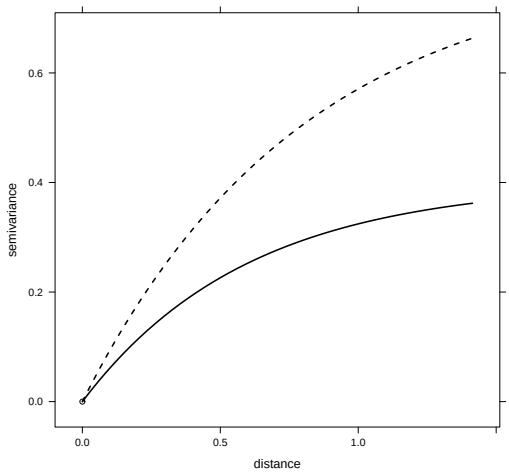


Figure 1: Exponential variograms models with $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.4, 0.6)$ (full line) and $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.8, 0.8)$ (dashed line).

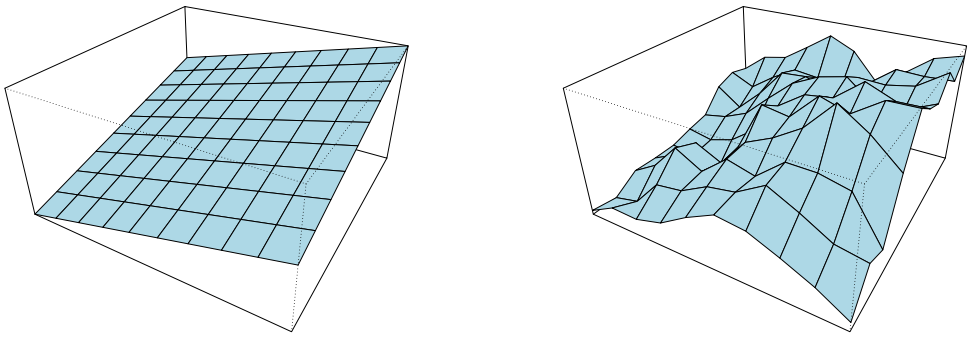


Figure 2: Spatial process (7) without the error term (left) and with the error term (right). Sample size: 100. Simulation results for a regular Gaussian spatial process in the unit square whose trend parameters are $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$. The dependence structure is explained by a exponential covariogram with parameters $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.4, 0.6)$.

Figure 1 shows the variogram function for the different scenarios considered in the simulation. The dashed line, $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.8, 0.8)$, corresponds to a variogram which presents a stronger structure of dependence compared with of the full line, $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.4, 0.6)$. Figure 2 are surfaces of a simulated process, as given in (7), considering that the trend parameters are $\beta_0 = 2$, $\beta_1 = 1$ and $\beta_2 = 1$, and the dependence structure is explained by a exponential covariogram with parameters $\boldsymbol{\theta} = (c_0, c_e, a_e)' = (0, 0.4, 0.6)$. The left panel corresponds to the process without the error (plane), while the right panel is with the error term.

Table 1 summarizes the results for the different scenarios. It can be seen that the proportion of rejections is similar to the theoretical level $\alpha = 0.05$ in all cases, although the proportion of rejections is less than the expected value.

$\boldsymbol{\beta}$	c_e	a_e	n	
$(2, 1, 1)$	0.4	0.6	100	0.041
$(2, 1, 1)$	0.8	0.8	100	0.049
$(1, 0, 1)$	0.8	0.8	100	0.044

Table 1: Proportion of rejections of the null hypothesis.

6. FURTHER WORK

In this work, we have proposed a test statistic for assesing a parametric trend surface, and we have used a boostrap procedure for calibrating this test. Although the asymptotic distribution of the test statistic is not sufficiently precise to compute the critical values of the test for small sample sizes, we will provide an expression of it. The distribution of T_n, under certain assumptions, can be approximated by a Gaussian distribution (under the null hypothesis and also under local alternatives converging to the null hypotesis at a rate c_n). Regarding the boostrap procedure, we will check the power of the test considering different alternatives.

REFERENCES

Alcalá, J. T., Cristóbal, J. A., and González-Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials.*Statistics & Probability Letters*, **42**(1), 39-46.

Bowman, A. W., and Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations* (Vol. 18). OUP Oxford.

Diggle, P., and Ribeiro, P.J. (2007). *Model-based Geostatistics.* Springer, New York.

Fernández-Casal, R., Castillo-Páez, S., and Francisco-Fernández, M. (2014). Nonparametric geostatistical risk mapping. *Stochastic Environmental Research and Risk Assessment*, 1-10.

Härdle, W., and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, 1926-1947.

Kaiser, M. S., Lahiri, S. N., and Nordman, D. J. (2012). Goodness of fit tests for a class of Markov random field models. *The Annals of Statistics*, 104-130.

Stute, W., and Manteiga, W. G. (1996). NN goodness-of-fit tests for linear models. *Journal of statistical planning and inference*, **53**(1), 75-92.

ACKNOWLEDGMENTS

This research has received financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF). This research has been partially supported by MINECO grant MTM2014-52876-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.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

ESTUDIO DE UN PROBLEMA DE PLANIFICACIÓN DE TAREAS EN UNA EMPRESA DE AYUDA A DOMICILIO

Isabel Méndez-Fernández[1], Luisa Carpente[2], Julián Costa-Bouzas[2], Ignacio García-Jurado[2], Silvia Lorenzo-Freire[2], y Guido Novoa-Flores[2]

[1]CITIC, Universidade da Coruña
[2]Grupo MODES, Departamento de Matemáticas, Universidade da Coruña

RESUMEN

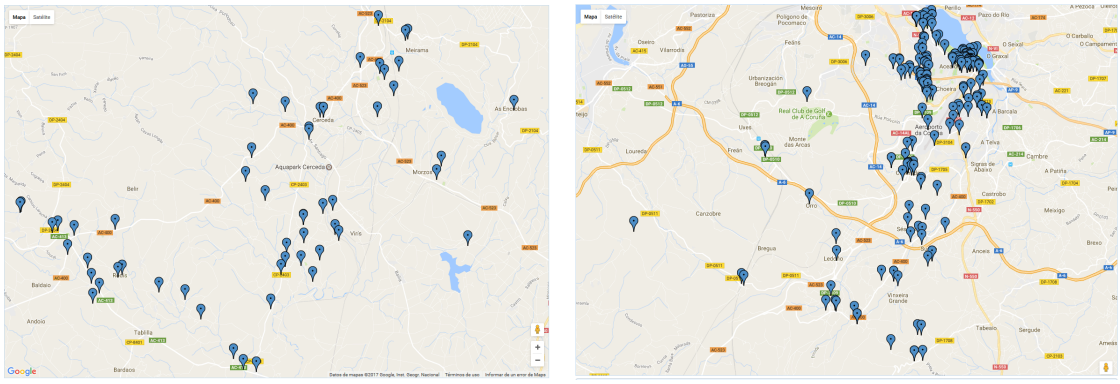
Este trabajo se centra en el estudio de un problema de planificación de tareas en una empresa de ayuda a domicilio situada en la provincia de A Coruña. Para ello, hemos diseñado un problema de programación lineal que recoge todas las características del problema. Debido al gran tamaño de los problemas reales a los que nos enfrentamos, haremos uso de algoritmos basados en técnicas heurísticas/metaheurísticas, que permitirán dar solución a las incidencias que surgen diariamente en la empresa.

Palabras y frases clave: Optimización, planificación, algoritmos heurísticos

1. INTRODUCCIÓN

La ayuda a domicilio es un recurso que permite a las personas mayores y/o dependientes mantenerse viviendo en su casa a pesar de pasar por una situación de dependencia o necesidad de apoyo en distintas tareas de su vida diaria. La empresa Mayores presta servicios de ayuda a domicilio desde 1997, tanto en la ciudad de A Coruña como en los ayuntamientos limítrofes. Desde 2012 ofrece sus servicios en Ferrol y su comarca.

Así pues, los usuarios del servicio proporcionado por la empresa son aquellas personas que necesitan cuidados o ayuda en la realización de determinadas tareas y para mejorar su calidad de vida recurren a la empresa. El número de usuarios es muy variable, dependiendo de la zona de actuación. A modo de ejemplo, en la Figura 1 (a) aparecen localizados los 58 usuarios del ayuntamiento de Cerceda, poco numerosos y dispersos por tratarse de una zona más rural, mientras que la Figura 1 (b) presenta la localización de los 210 usuarios del ayuntamiento de Culleredo que, además de ser más numerosos, se encuentran en su mayor parte concentrados en la zona urbana.



(a) Cerceda (b) Culleredo

Figura 1: Localización de usuarios

Para prestar sus servicios, la empresa dispone de un grupo de auxiliares, que puede aumentar o disminuir en función de las incidencias que surgen diariamente. Las auxiliares son las personas encargadas de visitar los domicilios de los usuarios y realizar las tareas que les han sido asignadas (supervisión de la medicación prescrita, baño/ducha, barrer y fregar la habitación del usuario, etc.). Las tareas que deben realizar las auxiliares, y que pueden variar en función del usuario que reciba el servicio, se dividen en dos categorías:

- Con restricción horaria, ya que algunos servicios deben realizarse en un cierto abanico de horas (por ejemplo, levantar y acostar al usuario).
- Sin restricción horaria. Otras tareas pueden realizarse a cualquier hora del día (por ejemplo, limpiar el polvo o poner la lavadora).

De esta forma, la empresa debe llevar a cabo un plan de intervención para cada usuario, de modo que para cada usuario se determine la auxiliar que lo atenderá, el horario en el que la auxiliar estará en su domicilio y los servicios a realizar. Para ello habrá que tener en cuenta el horario de disponibilidad del usuario y las actividades que se deben realizar en cada tramo del día. El plan de intervención tiene una periodicidad, normalmente semanal, pero también hay servicios excepcionales, donde el usuario requiere de una auxiliar para la realización de ciertas tareas, en un horario ajeno al plan de intervención.

La empresa dispone de varias coordinadoras que se ocupan de gestionar el plan de intervención de los usuarios. Cada coordinadora tiene asignado un grupo de usuarios, en función de las zonas o barrios que les corresponde coordinar. Las coordinadoras se encargan de organizar manualmente los planes de trabajo de las auxiliares que tienen a su cargo. Para llevar a cabo esta organización, tienen que gestionar la bolsa de horas de cada auxiliar. Con la bolsa de horas se lleva la cuenta de las horas que cada auxiliar hace por encima o por debajo de su horario semanal.

Para facilitar el trabajo de las coordinadoras, se está creando una aplicación para digitalizar la planificación de las tareas. Esta aplicación se encarga de recoger los datos personales de las auxiliares y de los usuarios, así como de la implementación de una agenda que facilita la creación y modificación de la planificación de los servicios que deben realizar las auxiliares. En la Figura 2, se muestra un ejemplo de la apariencia de dicha aplicación.

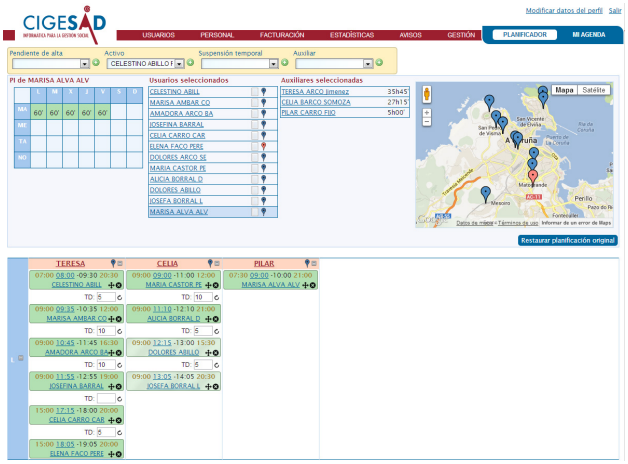


Figura 2: Aplicación para la gestión de tareas

Con nuestro trabajo, haciendo uso de diferentes técnicas de Investigación Operativa, se pretende dotar a esta aplicación de una herramienta que permita automatizar la modificación de las planificaciones en el caso de que se generen incidencias, es decir, permitirá recalcular de forma automática el plan de trabajo cuando se produzcan incidencias por parte de los usuarios o de las auxiliares. Los tipos de incidencias más comunes serían los siguientes:

- Que un usuario requiera un aumento de los servicios requeridos, por lo que se tendrían que asignar estos nuevos servicios a las auxiliares.

- Que aparezca un nuevo usuario, con lo que se tienen que asignar todos los servicios del usuario a las auxiliares.
- Que el usuario no pueda recibir a las auxiliares, porque está hospitalizado, de vacaciones, etc.
- Que una auxiliar no esté disponible a lo largo de un período, bien porque se ponga enferma o se vaya de vacaciones, con lo que se tendrían que asignar todos sus servicios a otras auxiliares.

2. SOLUCIÓN DEL PROBLEMA

Para resolver el problema descrito anteriormente, utilizaremos algoritmos basados en técnicas heurísticas presentes en la literatura. Estos algoritmos proporcionarán en poco tiempo una planificación semanal de los servicios que las auxiliares deben realizar, donde se incluirán todas las incidencias. Cada algoritmo heurístico tiene a su vez dos versiones, cuyos pasos se detallan a continuación:

Versión 1:

(1) Partimos de la solución inicial.

(2) Eliminamos los servicios no activos.

(3) Se ordena de forma aleatoria el conjunto formado por los nuevos servicios o servicios antiguos modificados.

(4) Ordenamos las auxiliares de acuerdo a su adecuación para realizar el servicio fijado en el paso (3) y se lo asignamos a la mejor auxiliar para este servicio.

(5) Una vez fijado el servicio y la auxiliar, colocamos el servicio donde menos solapamiento provoque.

(6) Se hace uso de la técnica heurística para eliminar el solapamiento. En caso de que no sea posible eliminar el solapamiento, se recurre a la siguiente mejor auxiliar de acuerdo al orden fijado en el paso (4) y volvemos al paso (5).

Si ya hemos utilizado todas las auxiliares disponibles y no se ha eliminado el solapamiento, creamos una auxiliar ficticia y le asignamos el servicio.

(7) Una vez eliminado el solapamiento, se hace uso nuevamente de la técnica heurística para minimizar los tiempos de descanso.

(8) Escogemos el siguiente servicio de acuerdo al orden fijado en el paso (3) y vamos al paso (4) hasta que todos los servicios estén asignados, en cuyo caso finalizaría el algoritmo.

Versión 2:

(1) Partimos de la solución inicial.

(2) Eliminamos los servicios no activos.

(3) Se ordena de forma aleatoria el conjunto formado por los nuevos servicios o servicios antiguos modificados.

(4) De acuerdo al orden fijado en el paso (3), vamos asignando cada uno de los servicios a la auxiliar que más se adecúa a dicho servicio.

Cada uno de los servicios se colocaría en la posición donde el solapamiento sea menor.

(5) Una vez asignados todos los servicios, utilizamos la técnica heurística para eliminar el solapamiento.

Para ello se van fijando los servicios de acuerdo al orden fijado en el paso (3) y, en caso de que haya solapamiento, se intenta eliminar.

Si no es posible, se recurre a la siguiente mejor auxiliar para el servicio, hasta que se elimine el solapamiento. Si no se puede conseguir, creamos una auxiliar ficticia y se lo asignamos.

Después de resolver los problemas de solapamiento para todos los servicios, pasamos al siguiente paso.

(6) En este paso aplicaremos de nuevo la técnica heurística para reducir los tiempos de descanso.

Como se puede observar, en ambas versiones se aplica la técnica heurística dos veces, una para resolver los solapamientos y otra para minimizar los tiempos de espera. La principal diferencia es que en la Versión 1 este procedimiento se realiza servicio a servicio y en la Versión 2 se hace para todos los servicios a la vez.

En cuanto a la solución inicial, en ambas versiones se considera la última planificación de la empresa.

Para la mejora de la solución, en ambos casos se trabajará con el vecindario de la solución fijada en cada paso del algoritmo. El vecindario se creará de acuerdo a los siguientes posibles movimientos o modificaciones de la solución:

Replanificar un servicio, que consiste en cambiar la hora de inicio y final de un servicio dado.

Desplazar servicios, que consiste en retrasar (o adelantar) la hora de inicio y fin de los servicios anteriores (o siguientes) a un servicio dado.

Intercambiar dentro, que consiste en intercambiar el horario de inicio entre dos servicios que realiza una auxiliar.

Para evaluar las soluciones en cada paso de los algoritmos, se tendrán en cuenta los siguientes objetivos:

- Eliminar el solapamiento.
- Minimizar el tiempo perdido entre servicios, es decir, el tiempo que una auxiliar ocupa en desplazarse desde la vivienda de un usuario a la siguiente y los tiempos de descanso que pueden tener las auxiliares entre servicios.
- Hay que equilibrar la bolsa de horas, es decir, asignar nuevos servicios a auxiliares que deben horas de trabajo y no asignar servicios a las auxiliares que han trabajado de más, aunque tengan huecos en su planificación. Este objetivo está relacionado con el orden de las auxiliares de acuerdo a su adecuación al servicio.
- No modificar en exceso las planificaciones dadas, puesto que es muy posible que en poco tiempo sea necesario volver a replanificar los horarios y una frecuente modificación de los horarios puede generar problemas para los usuarios. Por otro lado, los usuarios prefieren ser siempre atendidos por la misma auxiliar y para conseguirlo están permitidas las modificaciones de los horarios anteriores. Este objetivo se contempla con la elección de la solución inicial de acuerdo a la planificación anterior y con el orden de las auxiliares en función de su adecuación al servicio.

Con cualquiera de las dos versiones, se resuelven las incidencias y se consiguen soluciones factibles que proporcionan una planificación en la que los nuevos servicios han sido añadidos a la agenda de las auxiliares. En caso de que queden servicios asignados a la auxiliar ficticia, es necesario aumentar las horas de trabajo de alguna auxiliar o la contratación de más personal.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Economía y Competitividad a través del proyecto MTM2014-53395-C3-1-P, por el Centro para el Desarrollo Tecnológico Industrial a través del proyecto ITC-20151247, y por la Xunta de Galicia mediante las ayudas para Grupos de Referencia Competitiva ED431C-2016-015 y Centro Singular de Investigación de Galicia ED431G/01.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

BIG DATA EN PROBLEMAS DE CLASIFICACIÓN: APRENDENDO A COMPARTIMENTAR OS DATOS

Pablo Montero-Manso[1] e José A. Vilar[1]

[1] Grupo MODES, Departamento de Mátemáticas, Facultade de Informática, Universidade da Coruña.

RESUMO

Este traballo explora a posibilidade de aprender a realizar unha compartimentación dun conxunto de datos nun problema de clasificación cando o tamaño deste é inabordable computacionalmente. Como compartimentación enténdese separar o conxunto de datos en particións que serán tratadas de xeito completamente independente unhas de outras. Un novo dato a clasificar asígnase a unha única partición, a diferencia doutros métodos de clasificación distribuidos, como pode ser un esquema de votación que recolle o resultado da clasificación de cada unhas das particións dispoñible. A vantaxe de compartimentar os datos é inicialmente computacional, trivializando o procesado en paralelo, pero tamén pode ser inferencial (Dick et al., 2017).

A intratabilidade de chegar á solución óptima do problema de compartimentización lévanos a propor unha aproximación lineal, limitada tamén á división dos datos en dúas particións. A proposta polo tanto consiste en atopar un hiperplano separador combinando dous elementos: Por un lado maximízase o *balanceo*, é dicir, que as dúas particións reciban a mesma cantidade de datos. Ó mesmo tempo procurase que este hiperplano sexa de *mínima densidade* (Pavlidis et al., 2016), que percorra a zona de menor densidade posible. Esta primeira aproximación non precisa das etiquetas dos datos a clasificar polo que pode considerarse á súa vez como un método de cluster. Se se incluen as etiquetas, engádese a condición de respetar as proporcións das clases.

O bo funcionamento da proposta ilústrase con exemplos tradicionais en clasificación, mellorando en gran medida o tempo de cálculo, pero tamén a tasa de acerto en clasificación.

Palabras e frases chave: Big Data, Paralelo, Partición, Clasificación, Aprendizaxe Distribuido

REFERENCIAS

Dick, T., Li, M., Pillutla, V. K., White, C., Balcan, N. and Smola, A. (2017, April). Data Driven Resource Allocation for Distributed Learning. In Artificial Intelligence and Statistics (pp. 662-671).

Pavlidis, N. G., Hofmeyr, D. P., and Tasoulis, S. K. (2016). Minimum density hyperplanes. Journal of Machine Learning Research, 17(156), 1-33.

XII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 y 28 de octubre de 2017

THE PAINTING RULE IN MINIMUM COST SPANNING TREE PROBLEMS WITH MULTIPLE SOURCES

Adriana Navarro Ramos[1] and Gustavo Bergantiños[2]

[1] Ph.D student. Statistics and Operational Research. University of Vigo.
[2] University of Vigo.

ABSTRACT

We study minimum cost spanning tree problems with multiple sources and extend the painting rule to this context. We prove that this rule coincides with the generalized folk rule and provide a characterization of the rule based on meaningful properties.

Keywords: cost sharing, minimum cost spanning tree problems, multiple sources, folk rule, painting rule.

1. INTRODUCTION

Consider that a group of agents is interested in some service which is provided by several suppliers, also called sources. The connections have associated a cost and agents do not care whether they are connected directly or indirectly to such sources, but they want to be connected to all of them. This situation is a generalization of the classical minimum cost spanning tree problem where there is a unique source.

Given a cost spanning tree problem with multiple sources, the first issue to solve is to look for the least costly connection (a tree) that provides all the agents with the resource. Such tree can be obtained, in polynomial time, using the same algorithms as in the classical problem (for instance, Kruskal (1956) and Prim (1957)).

The second question we have to answer is how to allocate the cost of the obtained tree among the agents. This issue has been studied in the economic literature for similar problems. For the case that is studied in this research, Bergantiños *et al* (2016) extend different definitions of the folk rule, a well-known and widely studied solution for classical minimum cost spanning tree problems, to this new context. In Bergantiños and Lorenzo (2016) a set of several rules obtained through Kruskal's algorithm is defined and characterized.

Under a general framework of connection problems involving a single source, Bergantiños *et al* (2014) propose a cost distribution rule, called the *painting rule*. In addition to analyzing the properties of this rule and characterizing it, they show that for the painting rule coincides with the folk rule on the classical minimal cost spanning tree problem.

The objective of this work is to extend the painting rule, considering the same model as in Bergantiños *et al* (2016) and to prove that it coincides with the folk rule in the case of multiple sources.

2. THE PAINTING PROCEDURE

A minimal cost spanning tree problems with multiple sources problem is a triple (N, M, C) where N is a finite set of agents, M is a finite set of sources and C is a cost matrix that collects the cost of the direct link between any pair of agents and/or sources.

After the construction of a minimal tree t, we define the painting rule for minimal cost spanning tree problems with multiple sources as a two-phase procedure as follows:

> *Phase 1.* We build a tree t^* where all the sources are connected among themselves. This is achieved by removing the most expensive edge in the path that join two set of sources and use it to connect them.

Phase 2. Once the tree on Phase 1 is constructed, we define a rule for sharing its cost among the agents through a multi-stage procedure. In order to illustrate the procedure used to obtain the rule, assume that the nodes of the agents represent their houses and the edges are the roads connecting them to the sources. These roads need painting and there is only one machine to do this for each one. The machines cannot be moved to another road and all of them work at the same speed. At every stage, each agent is assigned to an edge while the path from her house to the sources has not been completely painted. The roads have painters assigned to them if the painting has not been completed. In each step, the agents assigned to an edge which is not completely painted share equally the time the painting machine is in operation. This can be read as their paying the same cost in that segment. At Stage 1, each agent is assigned to the first edge in the unique path from her house to the sources. At Stage s, each agent is assigned to the first unpaid edge in this unique path. If all edges in such a path have already been paid for in the previous stages, then this agent has finished her individual job. The last stage is reached when all agent have finished their job and the edges connecting the sources will be paid equally among the agents.

Consider a problem with 4 agents and 2 sources with the minimal tree depicted on Figure 1.

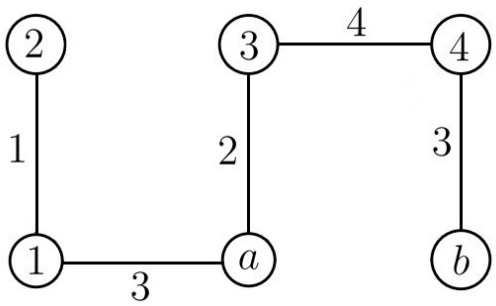


Figure 1: Original minimal tree.

Phase 1. Since the sources are not join directly, we create a new tree (Figure 2) where they are. We remove the most expensive edge in the path that from a to b and we use it to connect the sources; this is edge $(3, 4)$.

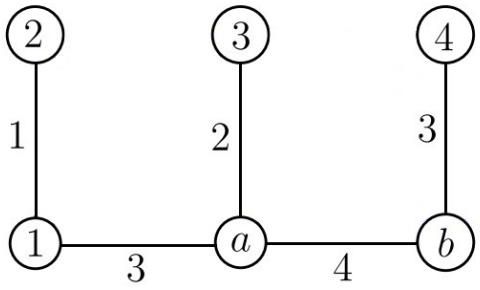


Figure 2: Tree obtained after Phase 1.

Phase 2. Figure 3 shows how the tree is painted.

Stage 1, in blue:

Agent	Assigned edge	Cost paid
1	$(1, a)$	1
2	$(2, 1)$	1
3	$(3, a)$	1
4	$(4, b)$	1

Table 1: Stage 1 (blue).

Then $(2, 1)$ is already paid.
Stage 2, in red:

Agent	Assigned edge	Cost paid
1	$(1, a)$	1/2
2	$(1, a)$	1/2
3	$(3, a)$	1
4	$(4, b)$	1

Table 2: Stage 2 (red).

Notice that edge $(3, a)$ is fully paid and that agent 3 has finished her job.
Stage 3, in green:

Agent	Assigned edge	Cost paid
1	$(1, a)$	1/2
2	$(1, a)$	1/2
3	$\varnothing$	0
4	$(4, b)$	1

Table 3: Stage 3 (green).

After this stage, all agents are connected to some source.
Stage 4, in black. The only edge to be painted is (a, b) and this is distributed equally among all agents. The next table 1 shows the amounts each agent pays.

Agent	Cost paid
1	1+1/2+1/2+1=3
2	1+1/2+1/2+1=3
3	1+1+0+1=3
4	1+1+1+1=4

Table 4: Distribution of costs among agents.

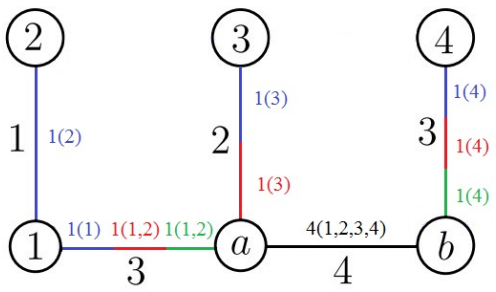


Figure 3: The painting procedure.

3. A CHARACTERIZATION

Cone-wise additivity (CA). If for two cost matrices C and C' there is a common ranking of the edges from cheapest to most expensive, then the cost shares on the problem $(N, M, C + C')$ are the sum of the cost shares on each problem. This property allow us to split the problem into a sum of smaller, simpler problems.

Cost monotonicity (CM). If a certain number of connection costs increase and the rest (if any) remain the same, no agent should end up better off.

Symmetry (SYM). If two agents are equivalent with respect to their connection costs, they should pay the same.

Isolated agents (IA). Assume there is an agent whose connection cost to any node is the same and greater than any other cost in the problem. This agent does not benefit from connecting to sources through other agents. Besides, suppose that there is a way of connecting all sources among themselves for free. Then, this agent should only pay her connection cost to any node.

Equal treatment of sources costs (ETSC). If the cost between two sources increases, then all agent should be affected in the same way.

The painting rule is the only rule that satisfies this five properties.

4. CONCLUSIONS

We have provided a new way of computing the folk rule in minimum cost spanning tree problems with multiple sources and with the characterization of the painting rule, we have obtained a new axiomatic characterization of the folk rule.

REFERENCIAS

Bergantiños G., Chun, Y., Lorenzo L. (2016). Minimum cost spanning tree problems with multiple sources: the folk rule. Mimeo, Universidade de Vigo.

Bergantiños, G., Gómez-Rúa, M., Llorca, N., Pulido, M., & Sánchez-Soriano, J. (2014). A new rule for source connection problems. European Journal of Operational Research, 234(3), 780-788.

Bergantiños G., Lorenzo L. (2016). Kruskal sharing rules in minimum cost spanning tree problems with multiple sources. Mimeo, Universidade de Vigo.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical society, 7(1), 48-50.

Prim, R. C. (1957). Shortest connection networks and some generalizations. Bell Labs Technical Journal, 36(6), 1389-1401.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Optimización de la producción en una fábrica de piensos.

Guido Ignacio Novoa-Flores[1], Luisa Carpente[1], Silvia Lorenzo-Freire[1]

[1] MODES Grupo de Investigación, Departamento de Matemáticas, Facultade de Informática, Universidade da Coruña, 15071, A Coruña, Spain.

RESUMEN

Consideramos el problema de organización de la producción en una fábrica de piensos gallega. En esta fábrica, se producen piensos a granel de distintas composiciones alimenticias para ajustarse a la necesidad de los clientes. Por tanto, se trata de organizar un horario de producción en un horizonte de tiempo para satisfacer las demandas y prioridades de los clientes. Debido a los diferentes tipos de pedidos que se procesan, es necesario limpiar y configurar las máquinas cuando se procesan pedidos de distinta composición, lo que produce paradas en la cadena de producción que trataremos evitar en la medida de lo posible. Proponemos un modelo de programación lineal para minimizar el número de veces que se cambia la maquinaria. Por la naturaleza de los escenarios presentes en la fábrica, la resolución del problema exacto se hace inabordable en un tiempo razonable. Finalmente, recurrimos a métodos heurísticos de resolución para diseñar un algoritmo que permitirá obtener rápidamente una planificación de la producción.

Palabras y frases clave: Problema de producción, Programación lineal, Algoritmos heurísticos.

1. INTRODUCCIÓN

En este trabajo se estudia el problema asociado a organizar la producción en una fábrica de piensos de una cooperativa agropecuaria gallega. La fábrica cuenta con aproximadamente 700 clientes fidelizados y entre 200 y 300 no fidelizados repartidos por todo Galicia, lo cual implica una gran cantidad de pedidos a satisfacer con una demanda muy variable (en la actualidad la fábrica produce anualmente unas 165000 toneladas). En cuanto al tipo de pedidos, la fábrica produce piensos genéricos para una amplia gama de animales de granja y, al mismo tiempo, permite a los clientes hacer peticiones individualizadas según las necesidades específicas de cada explotación. Este último tipo de pedidos representa un 85 % de la producción total de la fábrica, por tanto, es de vital interés para la fábrica optimizar esta parte de la producción para satisfacer la demanda de los clientes. No obstante, este tipo de pedidos presenta una gran variabilidad ya que se tienen que ajustar a las necesidades de cada explotación ganadera tanto en cantidad de pienso a producir como en composición alimenticia. Además, cada vez que se producen dos piensos con distinta composición de forma consecutiva es necesario limpiar las máquinas para evitar posibles contaminaciones y preparar la maquinaria para producir el siguiente pienso. Estos cambios en la maquinaria provocan paradas en la producción que habrá que evitar, en la medida de lo posible, para ahorrar tiempo. Todo lo mencionado se traduce en un elevado número de formas de organizar la producción por parte de los técnicos, que actualmente realizan la organización de forma manual. Por este motivo, es necesario modelar la situación en términos de un problema de optimización matemática con la finalidad de implementarlo en los servidores de la empresa para la automatización de la organización de la producción.

En la Sección 2, se expone el modelo de programación lineal propuesto para resolver este problema. Este modelo, incluye los parámetros necesarios para su resolución: tiempo de procesado de cada pedido, tiempo de limpieza de las máquinas después de procesar un pedido, tiempo de configuración de la maquinaria antes de iniciar un pedido, un código que se asocia al tipo de pedido, así como los días de holgura que tiene cada cliente para que su pedido sea procesado. Además, el modelo se ajusta a las restricciones propias de una cadena de producción. Finalmente, se expone la resolución de un problema ficticio así como el aumento exponencial de las dimensiones del problema a medida que aumenta el número de pedidos o el horizonte de días a programar. Este aumento provoca que la resolución de casos reales mediante el modelo exacto sea inabordable en tiempos razonables y sea necesario a recurrir a técnicas heurísticas que reduzcan considerablemente el tiempo de ejecución.

En la Sección 3, presentamos el algoritmo heurístico diseñado para resolver este problema. Este algoritmo, está basado en la técnica del enfriamiento simulado (*simulated annealing*) propuesto originalmente por Kirkpatrick et. al (1983). En primer lugar, construimos una solución inicial para la producción, a partir de la cual y de unos movimientos que definimos iremos recorriendo el espacio de soluciones posibles. En cada iteración, comparamos la solución actual y la resultante al realizar los intercambios mediante las funciones objetivo propuestas. Finalizamos la sección, exponiendo el funcionamiento del algoritmo con un ejemplo ficticio.

2. MODELO DE PROGRAMACIÓN LINEAL

A continuación, detallamos los parámetros necesarios para la formulación del modelo de programación lineal.

2.1 Parámetros

- Número de pedidos a procesar.
- Conjunto de días a ser programados.
- Número de unidades de tiempo disponibles para cada día.

Para cada pedido conocemos la siguiente información:

- Tiempo de procesado.
- Tiempo de limpieza después de procesar un pedido, en caso de que sea necesario.
- Tiempo de preparado de la maquinaria antes de procesar un pedido, en caso de que sea necesario.
- Ventana de tiempo asociada a un pedido, que indica el intervalo de tiempo dentro del cual se puede realizar el pedido.

2.2 Formulación del modelo

Como ya mencionamos anteriormente, si procesamos de forma consecutiva dos pedidos de distinto tipo, es necesario limpiar y preparar las máquinas para realizar el siguiente pedido, lo que conlleva una pérdida de tiempo. Por tanto, el objetivo será minimizar el número de cambios necesarios para producir todos los pedidos, respetando los horarios de apertura de la fábrica y las fechas en las que hay que entregar los pedidos.

Las restricciones que describen el problema son las propias asociadas a un problema de producción: todos los pedidos deben ser procesados respetando las ventanas de tiempo indicadas por los clientes, además de que los pedidos solo pueden ser procesados de uno en uno.

2.3 Ejemplo

A continuación, exponemos la resolución de un ejemplo ficticio usando el modelo de programación lineal antes descrito. En este caso, tenemos que organizar la producción de 9 pedidos, repartidos en tres tipos distintos, durante 3 días consecutivos. En el Cuadro 1 se recogen los tiempos de procesado (t_i), de limpieza después de procesar un pedido (c_i), de preparado de las máquinas antes de iniciar un pedido (b_i), las ventanas de tiempo de cada pedido ($[L_i, U_i]$), así como qué tipo de pedido es.

t_i	3	2	3	3	2	5	2	1	2
c_i	1	1	1	1	1	1	1	1	1
b_i	1	1	1	1	1	1	1	1	1
Tipo	A	B	B	C	B	A	A	A	C
L_i	1	1	1	1	1	1	1	1	23
U_i	11	11	33	33	33	33	33	33	33

Cuadro 1: Datos para el ejemplo.

En la Figura 1 exponemos la solución del problema mediante una planificación de los pedidos.

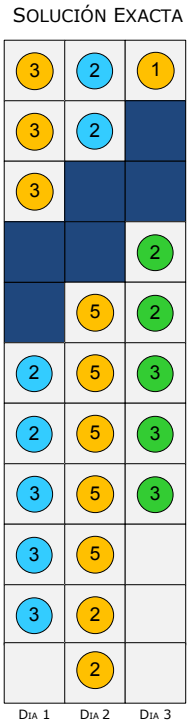


Figura 1: Solución obtenida mediante el modelo de programación lineal

En este caso, la función objetivo en la solución óptima tendría un valor de 6, es decir, que la producción estaría detenida durante 6 minutos debido a la limpieza y preparado de las máquinas.

El problema de programación lineal asociado tiene un total de 395721 restricciones y 1881 variables, para su resolución usamos el software Gurobi y el tiempo de resolución del problema fue de 30s. Nótese que a pesar de ser una situación ficticia y con pocos pedidos a organizar, el número de variables y restricciones es considerable. Actualmente, la fábrica procesa unos 25 pedidos por día (aunque con un comportamiento muy irregular) y su horario de apertura es desde las 07:00 am hasta las 21:00 pm. Por tanto, es razonable suponer que para una situación habitual de planificación de producción en la fábrica, el uso del modelo exacto de programación lineal sería ineficiente.

Finalizamos esta sección, exponiendo las dimensiones de los problemas de programación lineal generados con muestras de archivos reales (Cuadro 2).

Número de pedidos	Número de días a organizar	Restricciones	Variables
9	1	366962304	67329
14	1	777447340	96796
17	1	1915289385	164492
21	3	2919538169	580314

Cuadro 2: Dimensiones de muestras de problemas reales.

A la vista del Cuadro 2 se puede concluir que la resolución de situaciones reales de la fábrica sería inabordable en cualquier tiempo razonable usando el modelo exacto. Por esta razón, hemos diseñado el algoritmo heurístico que exponemos a continuación.

3. ALGORITMO HEURÍSTICO

El algoritmo heurístico usado en este trabajo está basado en la técnica de enfriamiento simulado propuesto por Kirkpatrick et. al (1983). En primer lugar, para implementar este método, necesitamos definir un espacio de soluciones y una función de coste. El algoritmo comienza con una solución inicial del espacio de soluciones dada y se exploran otras soluciones vecinas. En cada iteración, se comparan la nueva solución propuesta y la solución actual mediante la función de coste definida. Si la nueva solución propuesta es mejor que la actual, entonces actualizamos la solución. En caso contrario, hay una cierta de probabilidad de actualizar la solución actual por una peor. Esto sirve para evitar quedar atrapados en mínimos locales y recorrer más soluciones del espacio de soluciones.

3.1 Solución inicial

Comenzamos ordenando los pedidos por prioridad, esto es, por la cantidad de días disponibles para procesar cada pedido de acuerdo a las ventanas de tiempo. De esta forma, los pedidos que se tienen que procesar en un día tendrían prioridad de rango 1, los pedidos que se pueden procesar en dos días tendrían prioridad de rango 2, y así sucesivamente.

Una vez que tenemos los pedidos ordenados por prioridad, los vamos colocando en el día correspondiente. De esta forma, si tienen prioridad de rango 1 (un único día para ser procesado), los colocamos en el día indicado, si tienen prioridad de rango 2 (dos posibles días para ser procesados), los colocamos en uno de esos dos días al azar. Una vez colocados, los pedidos se ordenan por bloques del mismo tipo para reducir los tiempos de preparación y de limpieza de la máquina.

3.2 Cambios en la solución

Para mejorar la solución se proponen los $(\alpha_1, \alpha_2)-$intercambios. Estos intercambios consisten en lo siguiente:

1. Escogemos dos días al azar distintos.
2. Para cada día, elegimos los pedidos que puedan ser intercambiados al otro día respetando las ventanas de tiempo.
3. Fijamos α_1 y α_2 como un número al azar entre 0 y el número total de pedidos intercambiables.
4. Escogemos al azar α_1 y α_2 pedidos que intercambiamos entre los dos días seleccionados y, finalmente, se agrupan los pedidos por tipo.

3.3 Función objetivo

La función objetivo, se construye haciendo uso del orden lexicográfico, teniendo en cuenta las siguientes prioridades:

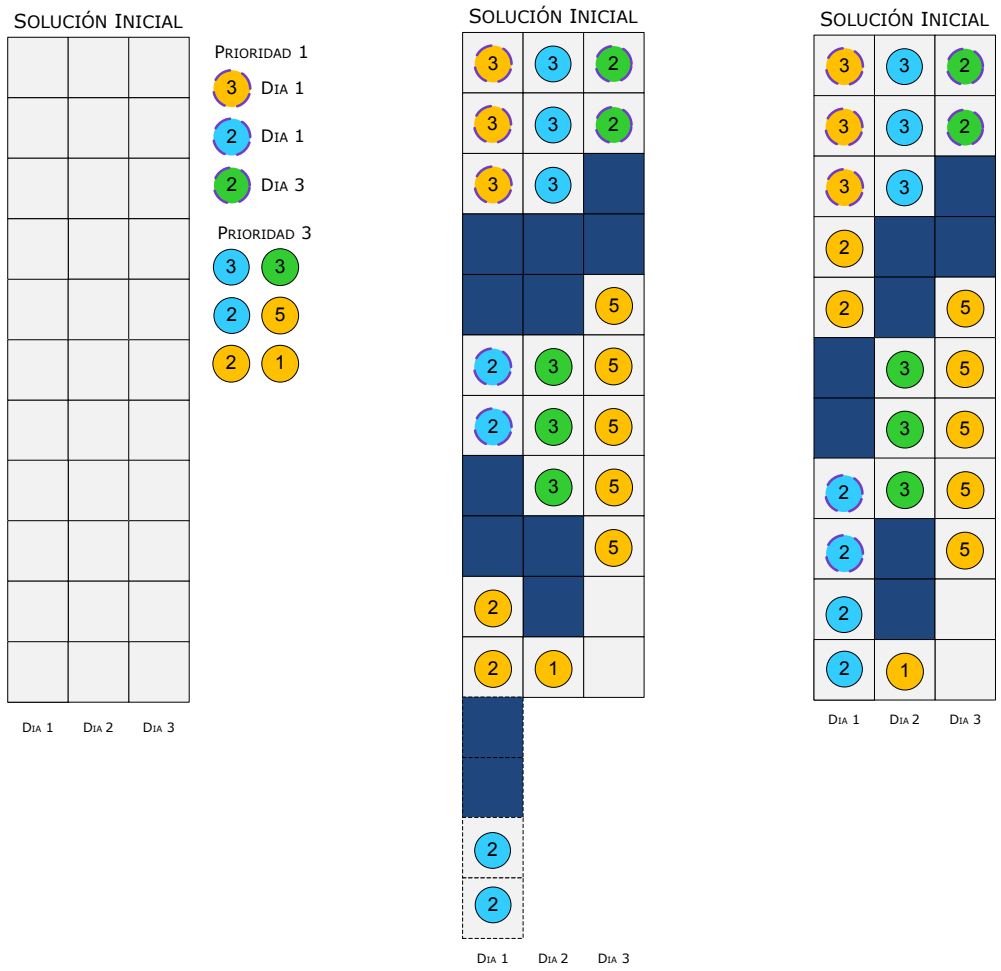
1. Factibilidad de la jornada laboral.
2. Minimizar los tiempos de preparado de las máquinas.
3. Equilibrio entre los tiempos libres al final de la jornada laboral

3.4 Ejemplo $(\alpha_1, \alpha_2)-$intercambios

A continuación, ilustramos el algoritmo propuesto mediante la resolución del ejemplo planteado en la sección anterior.

En la Figura 2a tenemos ordenados los pedidos por prioridad. De esta forma, los pedidos de prioridad 1 tienen que ser procesados el día correspondiente, sin posibilidad de ser desplazados a otro día, ya que en ese caso estaríamos ante una solución no factible. Por otro lado, los de prioridad 3 pueden ser procesados cualquiera de los tres días que estamos planificando. Por tanto, los colocamos aleatoriamente en uno de esos días, tal y como muestra la Figura 2b. Finalmente, ordenamos

los pedidos por el tipo al que pertenecen, obteniendo así la solución inicial (Figura 2c) con la que empezaría el algoritmo.

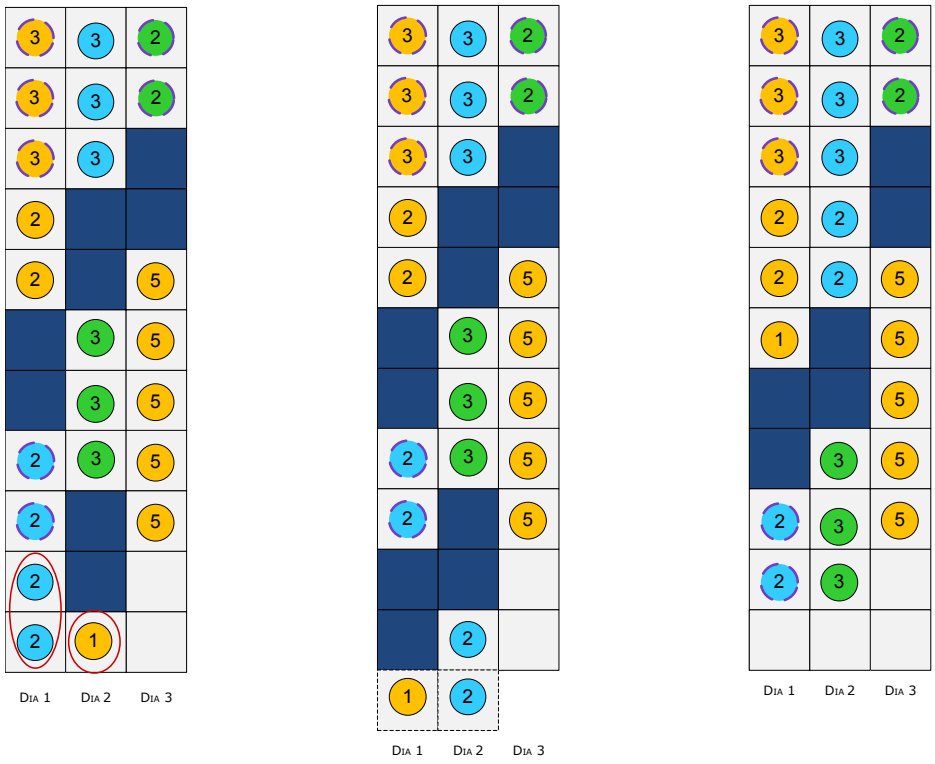


(a) Ordenar por prioridades (b) Colocar días aleatorios (c) Agrupar por tipo

Figura 2: Esquema de construcción de la solución inicial

Una vez que tenemos una solución inicial, comenzaríamos a realizar los $(\alpha_1, \alpha_2)-$intercambios para obtener nuevas soluciones y compararlas mediante las funciones objetivo definidas en el paso 3.3 del algoritmo.

Ahora, escogemos dos días al azar (supongamos los días 1 y 2). En el día 1, los pedidos intercambiables con el día 2 sin romper la factibilidad son el número 7 y el número 5, por lo que escogemos α_1 como un número al azar entre 0 y 2 (supongamos $\alpha_1 = 1$). Por su parte, del día 2 podemos intercambiar cualquiera de los tres pedidos respetando las ventanas de tiempo, por lo que tenemos que escoger α_2 como un número al azar entre 0 y 3 (supongamos $\alpha_2 = 1$). Finalmente, escogemos un pedido de cada día para ser intercambiado (supongamos que escogemos el pedido número 5 del día 1 y el número 8 del día 2 como muestra la Figura 3a). Acto seguido, realizamos el intercambio (Figura 3b) y, finalmente, reorganizamos la solución agrupando los pedidos por tipo (Figura 3c).



(a) Elección (α_1, α_2) (b) Intercambio de pedidos (c) Agrupar por tipo

Figura 3: Esquema de construcción de la solución inicial

Una vez hecho esto, procederíamos a comparar las soluciones antes de realizar el $(\alpha_1, \alpha_2)-$intercambio (Figura 2c) y la obtenida después de realizar el intercambio (Figura 3c). Esta comparación, la hacemos usando las funciones objetivo definidas en el paso 3.3 en el orden en el que aparecen.

En primer lugar, examinamos si ambas soluciones son factibles con la jornada de trabajo, esto es, que todos los pedidos que hemos asignado a ese día pueden realizarse dentro del horario de trabajo de la fábrica. En este caso, ambas soluciones son factibles con el horario de jornada laboral.

En tercer lugar, tenemos que ver si se reducen los tiempos de limpieza y preparado ocasionados al procesar de forma consecutiva dos pedidos de distinto tipo. En este caso, en el día 2 de la segunda solución (Figura 3c), vemos que solo hay que preparar las máquinas una vez, ya que ahora solo hay que procesar dos tipos de pedidos distintos, mientras que en la solución inicial (Figura 2c) se paraba la producción dos veces. Con esto, ya podríamos concluir que la nueva solución es mejor que la anterior y procederíamos a actualizar la solución actual.

A modo de ejemplo, si las soluciones obtenidas fuesen iguales con respecto a la comparación anterior, tendríamos que comprobar cual de ellas tiene mayor equilibrio en cuanto a los tiempos libres al final de la jornada. De esta forma, la solución de la Figura 2c tiene de holgura: 0, 0 y 2 unidades de tiempo, respectivamente para cada día, mientras que la solución de la Figura 3c tiene holguras respectivas de 1, 1 y 2 unidades de tiempo para cada día. Por tanto, hay mayor equilibrio en la segunda solución, lo cual la convierte en mejor solución según este criterio.

5. CONCLUSIONES Y TRABAJO FUTURO

En este trabajo, hemos propuesto un modelo matemático de programación lineal para la resolución del problema de producción en una fábrica de piensos. En el modelo se trataba de minimizar el número de veces que había preparar las máquinas cuando se procesaban, de forma consecutiva, dos pedidos de distinto tipo. Dado que estos pedidos se ajustan a la necesidad concreta de cada cliente, la fábrica se enfrenta constantemente a escenarios muy heterogéneos. Esto se traduce en un aumento exponencial de las dimensiones de los problemas a abordar a medida que aumentan el número de pedidos y la cantidad de días a programar, con lo cual, la resolución del problema mediante técnicas exactas resulta inabordable en la práctica mediante algoritmos exactos. Por esta razón, decidimos abordar el problema mediante algoritmos heurísticos usando la técnica del enfriamiento simulado obteniendo buenas soluciones en un tiempo razonable.

Por otro lado, ya que la fábrica dispone de la localización geográfica de los pedidos sería interesante utilizar esta información para diseñar conjuntamente las rutas de distribución con el problema de producción.

AGRADECIMIENTOS

Este trabajo ha sido financiada por el MINECO a través del proyecto MTM2014-53395-C3-1-P y por la Xunta de Galicia mediante el FEDER (Grupos de Referencia Competitiva ED431C-2016-015, Centro Singular de Investigación de Galicia ED431G/0, cuya financiación proecede de la Unión Europea a través del Fondo Europeo de Desarrollo Regional en un 80 %, y en el 20 % restante de la Xunta de Galicia) y la ayuda de apoyo a la etapa predoctoral en las universidades del sistema universitario gallego cofinanciadas parcialmente por el programa operativo FSE Galicia 2014-2020.

REFERENCIAS

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

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

PREDICTION INTERVALS FOR QUANTILE AUTOREGRESSION

Novo-Díaz, Silvia[1] e Sánchez-Sellero, César[2]

[1]Departamento de Matemáticas. Universidade da Coruña.
[2]Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela.

ABSTRACT

New methods are proposed to compute prediction intervals in quantile autoregression models, both under homoscedasticity and in general quantile autoregression models. The proposed methods are based on quantile estimation, bootstrap multipliers to mimic the variability in coefficients' estimation and bootstrap replicates of the future values. Consistency of the proposed methods is proven under homoscedasticity. Some simulations are provided to evaluate their performance for small to moderate sample sizes and to compare with already-existing methods in literature. An application to real data on unemployment in Spain is given.

Keywords: quantile autoregression, prediction intervals, bootstrap.

1. INTRODUCTION

One of the main goals in time series analysis is to provide a prediction and a prediction interval for a future observation at time $(n+k)$, based on the available realization of the series up to time n. We will focus on autoregressive linear models, where a new prediction interval will be proposed by means of quantile estimation and a bootstrap technique.

Classical methods to obtain prediction intervals for pth-order autoregressive [AR(p)] time series, as that proposed by Box and Jenkins (1976), are based on assuming that the error follows a normal distribution. As a consequence, these methods are not suited when the error distribution is not normal. For this reason, several authors have proposed alternative procedures based on the bootstrap. Stine (1987) suggests a bootstrap method that generates new values for the observed series and for future observations, using the estimated model and bootstrap errors drawn from the empirical distribution of re-scaled residuals. A drawback of Stine (1987)'s proposal is that it does not reproduce the randomness coming from the coefficients' estimation, which generally leads to under-coverage of prediction intervals. In order to include both the randomness of the future observation Y_{n+k} conditioned to the observed series and the randomness of coefficients' estimation, Thombs and Schucany (1990) propose to replicate the original series by means of the backward representation of the AR(p) model. Noting that the future observation only depends on the last p values of the observed series, the backward representation allows to fix these last p observations while reproducing new values for the other elements of the series. The empirical distribution of the re-scaled residuals, together with the coefficients estimated on the bootstrap series, are used to obtain bootstrap versions of the future observation. Cao et al. (1997) proposed a simplification to Thombs and Schucany (1990)'s method that omits the observed series generation. Their proposal saves computer time at the expense of a certain increase in coverage error. Pascual et al. (2001) suggest to generate new values of the observed series with the common forward representation of the AR(p) model. These new values are only used to obtain bootstrap replicates of the coefficients. Then, the last p values of the series are fixed and used to generate bootstrap values of the future observations. This way, both the randomness from the coefficients' estimation and the randomness from the future observation are reproduced, without the need of a backward representation. Pascual et al. (2001) also studied least absolute deviation estimators of the coefficients.

In this work a new method is proposed to obtain prediction intervals, that is based on least absolute deviation, or in general quantile estimation of the model coefficients, together with a boostrap method. Quantile regression is receiving a great attention in the last decades, due to its flexibility with respect to the error distribution (not necessarily normal) and possible heteroscedasticity. At the same time, quantile regression provides a more complete description of the response variable, because it allows to consider effects on the central and non-central locations in the response variable. In time series, quantile methods allows a more robust estimation of autoregression coefficients, particularly least absolute deviation is the option primarily considered in literature. Furthermore, general autoregressive models can be defined from a quantile perspective. Let us mention Koenker and Xiao (2006)'s model, where autoregressive coefficients are functions of the order $\tau \in (0,1)$ of the quantile of interest. This model implicitly includes heteroscedasticity in the autoregressive effect.

On Section 2 two new methods are proposed to obtain prediction intervals: one of them under homoscedastic autoregressive models and the other one for heteroscedastic quantile regression models. On Section 3 a simulation study is given to analyse the properties of the new method under homoscedastic models compared to the available literature. On Section 4 real data on the quarterly variation rate of unemployment in Spain are analysed following the techniques proposed in this paper.

2. THE PROPOSED METHODS

Let us consider an observed series coming from an AR(p) process of the form

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + a_t, \qquad t \in \{1, \ldots, n\} \tag{1}$$

where $\phi_0, \phi_1, \phi_2, \ldots, \phi_p$ are unknown coefficients, and the errors a_t are independent random variables with a common distribution, F_a. In classical time series analysis, the common distribution of the errors is assumed to be centred normal, with a certain variance σ^2. Here, normality will not be necessary for the proposed method. The fact that the errors follow a common distribution independent of the preceding values of the series, can be interpreted as homoscedasticity of the model. Note also that if the common distribution F_a has mean zero, then the linear combination in model (1) represents the mean. If F_a has τ-quantile equal to zero, with $\tau \in (0,1)$, then the linear combination in model (1) represents the τ-th quantile. Moreover, the same model can be transformed in an autoregressive model for any other quantile order, by only adding the corresponding quantile of F_a to the intercept.

Then, coefficients in model (1) can be estimated by least squares or by a quantile loss function of any chosen order τ, as follows:

$$\hat{\phi} = \arg \min_{\phi \in R^{p+1}} \sum_{t=p+1}^{n} \rho_\tau \left(Y_t - \phi^\top Z_{t,p} \right), \tag{2}$$

where $Z_{t,p} = (1, Y_{t-1}, \ldots, Y_{t-p})^\top$, $\phi = (\phi_0, \phi_1, \ldots, \phi_p)^\top$ and $\rho_\tau(u) = u(\tau - I(u < 0))$ for $u \in R$.

From the obtained fit, residuals can be computed as

$$\hat{a}_t = Y_t - \hat{\phi}_0 - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2} - \ldots - \hat{\phi}_p Y_{t-p}, \quad t = p+1, p+2, \ldots, n. \tag{3}$$

From a realization of the time series $Y_1, \ldots, Y_n$, we want a prediction interval for the future value Y_{n+k}. To this end, the following scheme is proposed:

Step 1. Fit the $AR(p)$ model given by (1) using a quantile loss function. Most commonly, least absolute deviation will be used. Obtain residuals from expression (3) and compute their empirical distribution.

Step 2. Draw a sample of independent positive random variables, w_i with $i = 1, \ldots, n$, and such that $E(w_i) = 1$ and $E(w_i^2) = 2$. The random variables are also independent of $\{Y_t\}_{t \in Z}$. For example, w_i can be independent and exponentially distributed with mean 1.

Step 3. Obtain an estimate of ϕ, by inserting the values w_i as multipliers to the loss function, that is:

$$\hat{\phi}^* = \arg \min_{\phi \in R^{p+1}} \sum_{t=p+1}^{n} w_t \rho_\tau (Y_t - \phi^\top Z_{t,p}), \tag{4}$$

Step 4. Draw bootstrap errors $\hat{a}^*_{n+j}$ for $j = 1, \ldots, k$ independently from the empirical distribution of the residuals obtained in Step 1.

Step 5. Define $Y^*_t = Y_t$ for $t = n-p+1, n-p+2, \ldots, n$ and compute future bootstrap observations following the expression:

$$Y^*_{n+j} = \hat{\phi}^*_0 + \hat{\phi}^*_1 Y^*_{n+j-1} + \hat{\phi}^*_2 Y^*_{n+j-2} + \ldots + \hat{\phi}^*_p Y^*_{n+j-p} + \hat{a}^*_{n+j} \quad j = 1, 2, \ldots, k.$$

Steps 2 and 3 try to mimic the randomness coming from the coefficients' estimation, while Steps 4 and 5 replicate the randomness related to the evolution of the time series up to time $(n + k)$. Bootstrap multipliers were already used in the context of quantile time series by Li et al. (2015).

To obtain the prediction intervals, Steps 2 to 4 are repeated B times, where B is a large enough number. One thousand is a common value, necessary for a level of 95%. For higher levels, a larger B would be required. The limits of the prediction intervals are obtained as the $\alpha/2$ and $(1 - \alpha/2)$ quantiles of the empirical distribution of the bootstrap values, say $(Y^{(1)}_{n+k}, \ldots, Y^{(B)}_{n+k})$.

Theorem 1. *Let $\{Y_t\}_{t \in Z}$ be an AR(p) process following the model given in (1). Under some regularity conditions, we have that, conditionally on the original realization of the time series, $(y_1, \ldots, y_n)$:*

(a) $\sqrt{n}(\hat{\phi}^* - \hat{\phi})$ converges to a normal distribution.

(b) Y^*_{n+k} converges in distribution to Y_{n+k}.

Proof. Part (a) is a direct consequence of Theorem 5 in Li et al. (2015). Part (b) can be proven following the same induction scheme given in the proof of Theorem 3.1 in Thombs and Schucany (1990). Here, Theorem 3.1 in Kreiss and Franke (1992) is used to prove that $\hat{a}^*_t$ converges in distribution to a_t.

Now, we will consider a more general autoregressive model, proposed by Koenker and Xiao (2006), given by

$$Y_t = \phi_0 (U_t) + \phi_1 (U_t) Y_{t-1} + \ldots + \phi_p (U_t) Y_{t-p}, \tag{5}$$

where $\phi_j : [0, 1] \to R$ with $j = 1, \ldots, p$ are functions defined in the unit interval $[0, 1]$ and $\{U_t\}_{t \in Z}$ is a sequence of independent random variables with uniform distribution on the interval $[0, 1]$. Note that under this model the vector of coefficients $\phi(\tau) = (\phi_0(\tau), \phi_1(\tau), \ldots, \phi_p(\tau))^\top$ depends on the quantile order τ. Thus, this model is suited to heteroscedastic time series. It can also be observed that the coefficients are random, and that they are dependent.

For those situations that are not well adjusted by model 1, and where the more general model 5 can be properly fitted, we propose a new method to obtain prediction intervals, given by the following scheme.

Step 1. Fit the $QAR(p)$ model given by (5), where functions $\phi(\tau)$ are estimated for a grid of values of τ, say, $0.05, 0.15, \ldots, 0.95$.

Step 2. Draw a sample of independent positive random variables, w_i with $i = 1, \ldots, n$, and such that $E(w_i) = 1$ and $E(w_i^2) = 2$. The random variables are also independent of $\{Y_t\}_{t \in Z}$. For example, w_i can be independent and exponentially distributed with mean 1.

Step 3. For each τ in the grid, obtain a bootstrap version of coefficients using the multipliers w_i (note that the same multipliers are used for all values of τ):

$$\hat{\phi}(\tau)^* = \arg \min_{\phi \in R^{p+1}} \sum_{t=p+1}^{n} w_t \rho_\tau \left(Y_t - \phi^\top Z^1_{t,p}\right), \tag{6}$$

Step 4. Draw independent random variables U_j for $j = 1, \ldots, k$, from the uniform distribution on the unit interval $[0, 1]$. Define $Y_t^* = Y_t$ for $t = n - p + 1, n - p + 2, \ldots, n$ and compute future bootstrap observations following the expression:

$$Y_{n+j}^* = \hat{\phi}_0^*(U_j) + \hat{\phi}_1^*(U_j)Y_n^* + \hat{\phi}_2^*(U_j)Y_{n-1}^* + \ldots + \hat{\phi}_p^*(U_j)Y_{n-p+1}^*, \quad j = 1, 2, \ldots, k.$$

The vector of bootstrap coefficients $\hat{\phi}(U_j)^*$ is approximated by $\hat{\phi}(\tau)^*$, where τ is the value in the grid considered in Step 1, that is closest to U_j.

Steps 2 to 4 are repeated B times and the prediction interval is obtained similarly to the previous method from the empirical distribution of the resulting bootstrap values.

3. SIMULATIONS

Now the new method proposed for the homoscedastic model given by (1) is compared with already-existing methods in literature. The simulated model is given by:

$$Y_t = 0.75Y_{t-1} - 0.50Y_{t-2} + a_t, \tag{7}$$

Two distributions are considered for the errors a_t: standard normal and exponential with mean 1. The considered sample sizes were $n = 25, 50$ and the prediction horizon was $k = 3$. The nominal coverage was 95%.

Coefficients' estimation is carried out by least squares or by least absolute deviation. We are going to evaluate the performance of several methods to obtain the prediction intervals: Box and Jenkins, Thombs and Schucany, Cao et al., Pascual et al. and the method proposed here.

$S = 500$ series are generated from model (7) with the true coefficients and error distribution. For each simulated series, every method produces a prediction interval at horizon 3. When boostrap is required, $B = 1000$ bootstrap replicates are used. Let us denote one such interval by (L_s, U_s) for $s = 1, \ldots, S$. In order to evaluate the performance of each method, $F = 1000$ future values are drawn from the true model and simulated series. Let us denote them by $Y_{n+k}^{s,f}$ for $s = 1, \ldots, S$ and $f = 1, \ldots, F$.

Then, the coverage conditioned to the s-th observed series can be approximated by

$$Cov_s = \frac{1}{F}\sum_{f=1}^{F} I(L_s < Y_{n+k}^{s,f} < U_s)$$

Clearly, a good method will be that producing coverages similar to the nominal level. We can also compute the length of intervals, given by

$$Length_s = U_s - L_s$$

Again, a short length will clearly be preferable. From the sample of coverages and lengths, we can compute their average and standard deviation.

An additional measure of the performance, considered by Thombs and Schucany (1990), is the Gamma-value. It is defined as the proportion of series with a conditional coverage larger than the nominal level, that is

$$Gamma = \frac{1}{S}\sum_{s=1}^{S} I(Cov_s \geq 1 - \alpha)$$

where $(1 - \alpha)$ denotes the nominal level.

In Table 1 results of these measures are given for each method of estimation, method to compute prediction intervals, error distribution and sample size. The proposed method, denoted by P, provides generally good results in terms of coverage compared to its competitors.

			Normal ($\mu = 0, \sigma = 1$)					Exponencial ($\lambda = 1$)				
n	Esti.	Method	Cover.	Std.err.	Length	Std.err.	Gamma	Cover.	Std.err.	Length	Std.err.	Gamma
25	LS	BJ	89.640	0.307	4.563	0.036	0.214	90.339	0.353	4.408	0.061	0.350
		TS	92.197	0.279	5.060	0.045	0.408	92.945	0.308	4.952	0.073	0.500
		CB	90.205	0.308	4.713	0.040	0.278	89.609	0.373	4.460	0.064	0.322
		PRR	92.535	0.247	5.055	0.042	0.396	93.119	0.302	4.911	0.072	0.512
	LAD	CB	91.051	0.302	4.994	0.046	0.326	90.408	0.371	4.689	0.066	0.368
		PRR	93.686	0.239	5.532	0.053	0.518	93.468	0.302	5.151	0.071	0.532
		P	94.796	0.225	6.128	0.079	0.612	94.688	0.270	5.592	0.087	0.620
50	LS	BJ	92.646	0.168	4.712	0.024	0.268	93.176	0.190	4.648	0.044	0.404
		TS	93.753	0.168	4.974	0.027	0.422	94.893	0.195	4.928	0.053	0.606
		CB	92.893	0.179	4.812	0.026	0.340	92.603	0.245	4.652	0.047	0.416
		PRR	93.883	0.163	4.983	0.027	0.452	94.926	0.192	4.901	0.051	0.608
	LAD	CB	93.249	0.181	4.916	0.029	0.350	92.686	0.263	4.763	0.048	0.404
		PRR	94.471	0.162	5.187	0.032	0.506	94.565	0.218	5.024	0.050	0.586
		P	94.994	0.161	5.347	0.034	0.614	95.294	0.177	5.084	0.052	0.630

Táboa 1: Results from S=500 series of $n = 25, 50$ observations, with F=1000 future values and B=1000 bootstrap replicates. Seed: 123456789. Nominal coverage: 95%. The following methods were considered: Box–Jenkins (BJ), Thombs and Schucany (TS), Cao et al. (CB), Pascual et al. (PRR) and the proposed method (P). Coefficients' estimation was made by least squares (LS) and least absolute deviation (LAD).

4. APPLICATION TO UNEMPLOYMENT DATA

An analysis of an economic real data example is obtained by means of the proposed methods. The data consists of the quarterly variation rate of unemployment in Spain from the second quarter of 1990 to the second quarter of 2016. These data were obtained by a survey *Encuesta de Población Activa*, conducted by the Spanish National Institute of Statistics (INE). Figure 1 below provides a representation of the series. The process seems to be stationary, although some atypical observations are detected. Some asymmetries are also observed, suggesting that a classical time series analysis could not be appropriate.

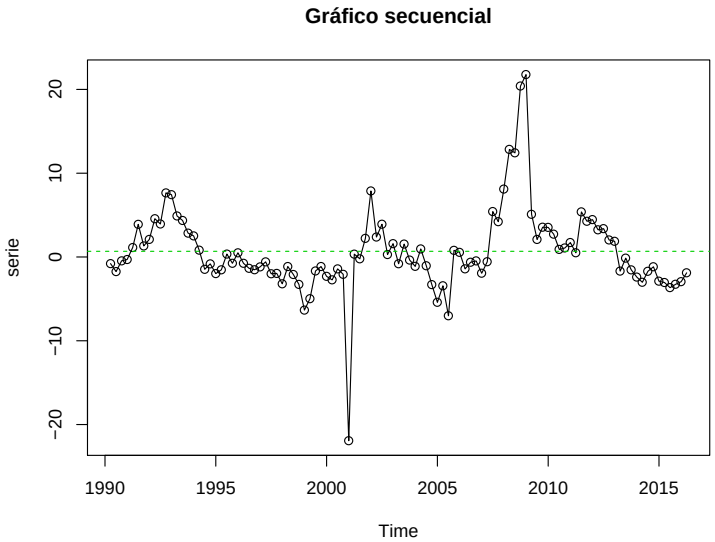


Figura 1: Quarterly variation rate of unemployment in Spain.

The representation of sample autocorrelations and sample partial autocorrelations, given in Figure 2, suggest an autoregressive model of order one. However, a scatterplot of each observation versus each preceding value in the series, given in Figure 3, shows a possible heteroscedastic behaviour.

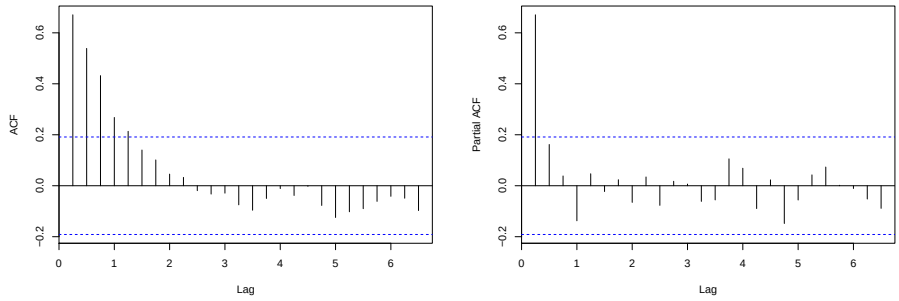


Figura 2: Left: sample autocorrelations; right: sample partial autocorrelations.

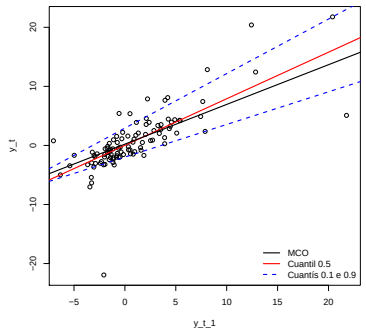


Figura 3: Scatterplot of Y_t versus Y_{t-1}.

Table 2 shows the prediction intervals and the true values for the Spanish unemployment data, obtained by each of the considered methods. It can be seen that the new method, adapted to the general quantile autoregression model, provides the shortest intervals, that still contain the true values.

	LS				LAD				
k	BJ	TS	CB	PRR	CB	PRR	P	QAR	Val. real
1	(-7.23,7.23)	(-6.17,6.87)	(-6.01,6.81)	(-5.97, 6.73)	(-5.82, 5.00)	(-5.86, 5.13)	(-5.93,5.02)	(-4.82,4.88)	-3.05
2	(-8.49,8.93)	(-9.91,9.55)	(-8.94,9.89)	(-10.21,9.82)	(-13.55,9.49)	(-12.87,9.90)	(-13.53,9.66)	(-5.30,6.73)	-3.80
3	(-8.44,8.97)	(-10.29,10.15)	(-10.35,10.67)	(-10.97,10.41)	(-14.17,10.77)	(-14.22,10.57)	(-15.07, 10.92)	(-5.85, 7.78)	-3.37
4	(-8.43,8.99)	(-10.37,10.64)	(-10.85,10.81)	(-10.64,10.99)	(-14.13,11.67)	(-14.94,10.82)	(-15.75,12.00)	(-6.09,8.67)	-5.08

Táboa 2: Prediction intervals for horizons 1, 2, 3 and 4 (from the third quarter of 2016 until the second quarter of 2017) obtained by the considered methods: Box–Jenkins (BJ), Thombs and Schucany (TS), Cao et al. (1997) (CB–LS) and (CB–LAD), Pascual et al. (PRR–LS) and (PRR–LAD), and the new method (P) and (QAR).

REFERENCES

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. San Francisco, Holden–Day.

Cao, R., Febrero-Bande, M., González-Manteiga, W., Prada-Sánchez, J.M. and García-Jurado, I. (1997). Saving computer time in constructing consistent bootstrap prediction intervals for autoregressive processes. Communications in Statistics–Simulation and Computation, 26, 961-978.

Pascual, L., Romo, J. and Ruiz, E. (2001). Effects of parameter estimation on prediction densities: a bootstrap approach. International Journal of Forecasting, 17, 83–103.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. Journal of the American Statistical Association, 101, 980-990.

Kreiss, J. P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving–average models. Journal of Time Series Analysis, 13, 297-316.

Li, G., Li, Y. and Tsai, C.-L. (2015). Quantile correlations and quantile autoregressive modeling. Journal of the American Statistical Association, 110, 246-261.

Stine, R. A. (1987). Estimating properties of autoregressive forecasts. Journal of the American Statistical Association, 82, 1072-1078.

Thombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. Journal of the American Statististical Association, 85, 410, 486–492.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

MODELIZACIÓN ESTADÍSTICA DEL PROCESO DE ENHEBRADO EN CONSTRUCCIÓN NAVAL

Déborah Otero[1], Ricardo Cao[2], Vicente Blasco[3], Javier Tarrío-Saavedra[2], Salvador Naya[2]

[1] Unidade Mixta de Investigación Universidade da Coruña – Navantia, Universidade da Coruña.
[2] Grupo MODES, Departamento de Matemáticas, ITMATI, CITIC, Universidade da Coruña.
[3] Navantia, Ferrol.

RESUMEN

En este trabajo se propone un modelo estadístico para estimar la probabilidad de enhebrado entre las bulárcamas que forman el armazón de los buques y los paneles compuestos de elementos longitudinales. El modelo desarrollado proporciona la probabilidad de enhebrado antes de que este se lleve a cabo, a partir de variables relacionadas con las dimensiones y estructura de panel y bulárcama. Para ilustrar y evaluar el procedimiento propuesto, se ha diseñado un estudio de simulación, asumiendo normalidad e independencia entre las variables estudiadas. La presente metodología se ha llevado a cabo en el marco de la Unidad Mixta de Investigación Navantia - Universidade da Coruña, que tiene por objeto el desarrollo de nuevas técnicas y tecnologías relacionadas con la Industria 4.0 para incrementar la competitividad del astillero de Ferrol y Fene.

Palabras y frases clave: Enhebrado, Construcción naval, Metrología, Probabilidad, Simulación.

1. INTRODUCIÓN

Una de las etapas más complejas y críticas en el proceso productivo de fabricación de unidades abiertas en la construcción naval consiste en el paso de los longitudinales a través de un elemento transversal. Este tipo de proceso se denomina enhebrado (Okumoto y Matsuzaki, 1997). A continuación, se definen los diferentes elementos que componen la estructura de un buque (Victoria, 1997), críticos en el proceso de enhebrado y, por tanto, necesarios para entender el modelo estadístico que se propone en este trabajo.

- Paño: Elemento prefabricado formado por la unión de chapas.
- Panel: Elemento prefabricado formado por el "paño" más la incorporación de perfiles longitudinales. Estos refuerzos longitudinales van colocados paralelos entre sí, a una distancia constante entre ellos.
- Unidad abierta: Conjunto prefabricado con base plana y abierto por la parte superior. Normalmente formado por el panel más la incorporación de otros elementos estructurales.
- Bulárcama: Elemento estructural, en sentido transversal de mayor sección que forman el armazón del buque y a las que van sólidamente unidas las planchas del forro exterior.

Aunque hay diferentes métodos de montaje de unidad abierta de acuerdo con cada astillero, uno de los métodos de montaje de panel más popular es el propuesto por Okumoto (2001) (Figura 1). En concreto, en los astilleros de Fene-Ferrol, se realiza el montaje utilizando este método. Primero se unen varias chapas, a continuación, después de un proceso de marcado, se incorporan varios longitudinales y por último se incorporan los elementos transversales.

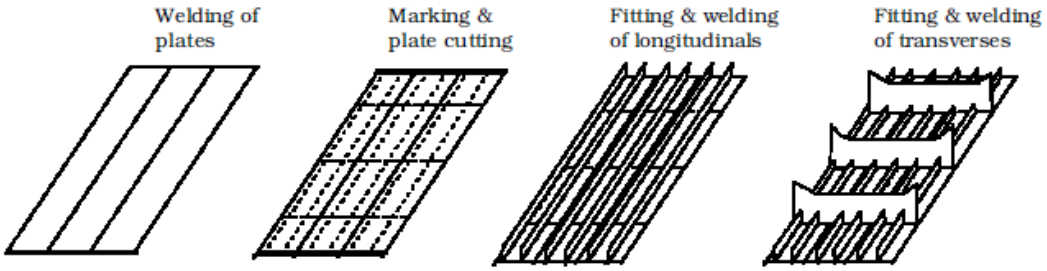


Figura 1: Procedimiento de montaje de unidad abierta. Fuente: Okumoto (2001).

A partir de la dificultad que conlleva la incorporación de un elemento transversal a un panel, surge la idea del siguiente estudio, cuyo objetivo es diseñar un modelo para predecir la probabilidad de enhebrado al inicio del proceso. Por enhebrar se entiende la incorporación de elementos transversales al panel (paño + longitudinales) (Figura 2). En particular, se estudia la probabilidad de incorporar una bulárcama a un panel. Dicho modelo, se prevé que sea de aplicación en los astilleros de Fene-Ferrol de Navantia.

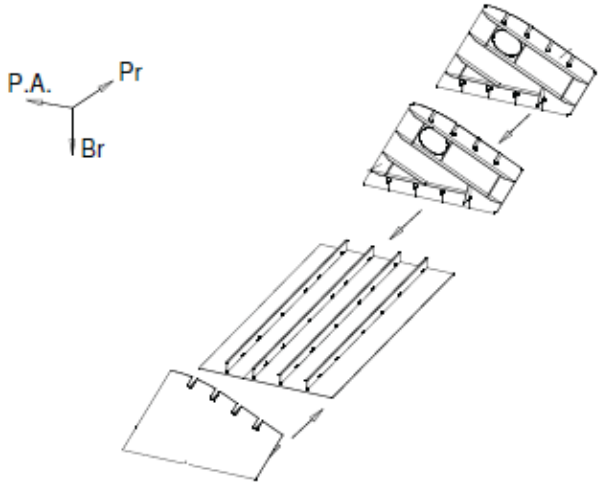


Figura 2. Conjunto de la estructura a enhebrar.

2. DISEÑO DEL MODELO

En primer lugar, se define un modelo simple que predice la probabilidad de enhebrado al inicio del proceso, para ello se supone que el paño tiene planicie perfecta, el hueco de las escotaduras (Figura 3), es decir, las aberturas de la bulárcama para paso de otro elemento, en este caso de los longitudinales, es perfectamente perpendicular y tiene altura suficiente para poder producirse el enhebrado y, por último, se supone que el panel tiene r longitudinales, por tanto, la bulárcama también tiene r escotaduras. Seguidamente se muestran las variables y los parámetros que se tienen para construir el modelo.

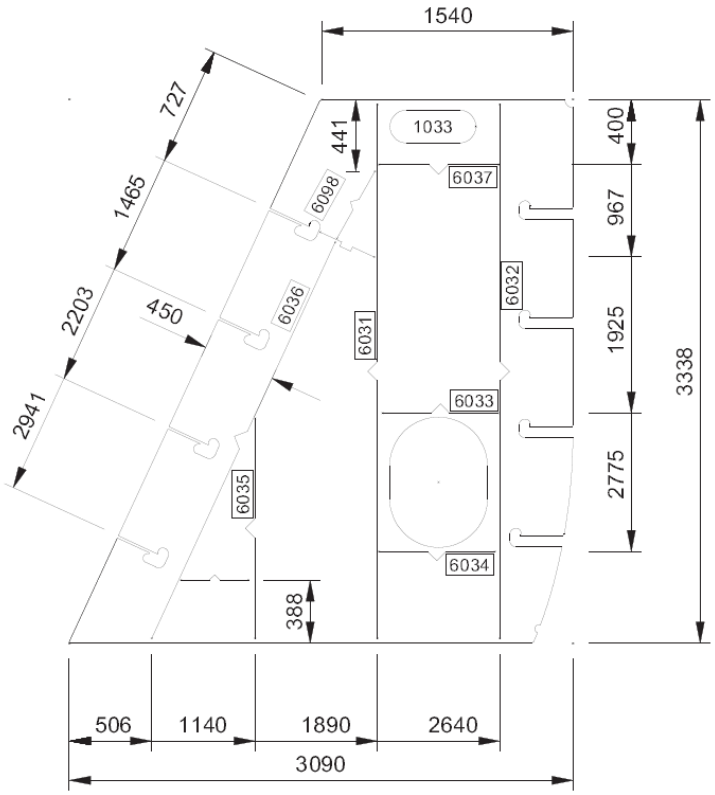


Figura 3. Escotaduras de enhebrado.

En el panel se define $\mathbf{X_i}$, con i = 1, …, *r*, como la distancia desde el canto hasta el centro del *i*-ésimo elemento longitudinal, $\mathbf{L_1}$ el ancho del panel y $\theta_\mathbf{i}$, con i = 1, …, *r*, el ángulo del *i*-ésimo elemento longitudinal con la chapa (≃90°). Por otro lado, en la bulárcama, se denota por $\mathbf{[Y_{i1}, Y_{i2}]}$, con i = 1, …, *r*, las distancias desde el canto hasta el inicio y hasta el final de la *i*-ésima escotadura y $\mathbf{L_2}$ el ancho de la pieza transversal.

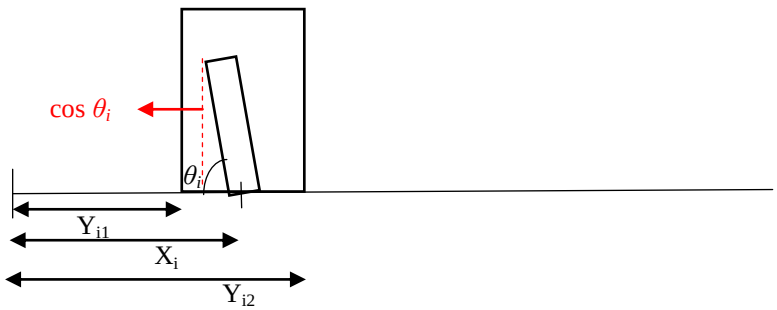


Figura 4. Variables que se definen para construir el modelo de enhebrado.

El enhebrado se producirá correctamente si los elementos longitudinales pasan a través de las escotaduras de la bulárcama sin producirse ninguna intersección entre ambos elementos. Sea *a* el espesor del longitudinal y *b* la altura de este, entonces, dependiendo del ángulo formado entre el longitudinal y la chapa (Figura 4), se pueden diferenciar tres casos:

1) $\theta_i = 90°$. El enhebrado se producirá correctamente cuando $\left[X_i - \frac{a}{2}, X_i + \frac{a}{2}\right] \subset [Y_{i1}, Y_{i2}]$. Es decir, se tiene que cumplir que $Y_{i1} \leq X_i - \frac{a}{2}$; $X_i + \frac{a}{2} \leq Y_{i2}$.

2) $\theta_i < 90°$. En este caso las condiciones que se deben cumplir son las siguientes:

$$Y_{i1} \leq X_i - \frac{a}{2} - b\cos\theta_i \ ; \ X_i + \frac{a}{2} \leq Y_{i2}$$

3) $\theta_i > 90°$. Se deberá cumplir que: $Y_{i1} \leq X_i - \frac{a}{2} \ ; \ X_i + \frac{a}{2} - b\cos\theta_i \leq Y_{i2}$.

Simplificando los tres casos anteriores, se concluye que el enhebrado se realizará correctamente cuando se verifican las siguientes condiciones:

Por un lado, debe cumplirse que

$$Y_{i1} \leq \min_{i=1,..,r}\left\{X_i - \frac{a}{2} - b\cos\theta_i \ , \ X_i - \frac{a}{2}\right\} = X_i - \frac{a}{2} + \min_{i=1,...,r}\{-b\cos\theta_i \ , 0\} =$$
$$= X_i - \frac{a}{2} - b\cos\theta_i \, I(\theta_i \leq 90°)$$

Por tanto,

$$\boxed{Y_{i1} \leq X_i - \frac{a}{2} - b\, I(\theta_i \leq 90°)\cos\theta_i}$$

Por otro lado,

$$\max_{i=1,...,r}\left\{X_i + \frac{a}{2} \ , \ X_i + \frac{a}{2} - b\cos\theta_i \ \right\} = X_i + \frac{a}{2} + \max_{i=1,...,r}\{0 \ , -b\cos\theta_i\} =$$
$$= X_i + \frac{a}{2} - b \ \cos\theta_i \, I(\theta_i > 90°) \leq Y_{i2}$$

Entonces

$$\boxed{X_i + \frac{a}{2} - b\, I(\theta_i > 90°)\cos\theta_i \leq Y_{i2}}$$

En consecuencia, la probabilidad de enhebrado se calcula según la siguiente expresión:

$$P(enhebrado) = P\left(Y_{i1} \leq X_i - \frac{a}{2} - b\, I(\theta_i \leq 90°)\cos\theta_i \, ,\right.$$
$$\left. X_i + \frac{a}{2} - b\, I(\theta_i > 90°)\cos\theta_i \leq Y_{i2} \, , |L_1 - L_2| < \varepsilon, \forall\, i = 1, \dots, r\right)$$

siendo ε el valor máximo permitido entre la diferencia del ancho del panel y el ancho de la bulárcama.

3. ESTUDIO DE SIMULACIÓN

Mediante el presente estudio de simulación se pretende evaluar el modelo diseñado en la sección anterior. A falta de datos reales, se define un primer modelo en el que las características de la simulación no están basadas en datos. Las suposiciones sobre las variables que forman el modelo son las siguientes.

En el panel, sea e la distancia desde la línea de trazado del primer longitudinal al canto, d la distancia entre los longitudinales, r el número de longitudinales y a el espesor de uno de ellos, entonces

X_1, la distancia desde el canto hasta el primer longitudinal se aproxima $X_1 \simeq e + \frac{a}{2}$. En el modelo a simular se supone que $X_1 \sim N(e + \frac{a}{2}, \sigma_1^2)$, es decir, que la distancia desde el canto hasta el primer longitudinal es una cantidad aleatoria con media $\boldsymbol{e + \frac{a}{2}}$ (su valor nominal) y una pequeña desviación típica, σ_1.

Por otro lado, sea $\xi_i = X_{i+1} - X_i$ i = 2, …, r-1 entonces en el modelo a simular se supone que la distancia entre longitudinales $\xi_i \sim N(d, \sigma_2^2)$.

Por tanto, $X_2 = X_1 + \xi_1$, $X_3 = X_2 + \xi_2$, … Es decir, $X_{i+1} = X_i + \xi_i$ i = 2 ,…, r-1.

Por último, dado que el ancho del panel, L_1, se puede aproximar por $L_1 \simeq L = 2e + (r\text{-}1)d + a$ entonces se supone que $L_1 \sim N(L, \sigma_3^2)$.

Por consiguiente, se tiene que simular

$X_1 \sim N(e + \frac{a}{2}, \sigma_1^2)$; $\xi_i \sim N(d, \sigma_2^2)$ i=2,...,r-1 y $L_1 \sim N(L, \sigma_3^2)$ todas ellas independientes.

En el caso de la bulárcama las suposiciones tomadas son las que se muestran a continuación:

Sea $\Delta_i = Y_{i,2} - Y_{i,1}$ i = 1, …, r la distancia desde el inicio hasta el final de la i-ésima escotadura, entonces $\Delta_i \sim N(a + 2\delta, \sigma_4^2)$, siendo δ la holgura de la escotadura (gap).

Por otro lado, la distancia entre longitudinales se puede aproximar mediante $d \simeq Y_{i+1,1} - Y_{i,2} + a + 2\delta$. Por tanto si $\eta_i = Y_{i+1,1} - Y_{i,2}$ i = 1, …, r-1 distancia desde el final de la i-ésima escotadura hasta el principio de la siguiente se tiene que $\eta_i \simeq d - 2\delta - a$. De esta forma, en el modelo a simular se supone que $\eta_i \sim N(d - 2\delta - a, \sigma_5^2)$.

$Y_{1,1}$, la distancia desde el canto hasta el principio de la primera escotadura se aproxima $Y_{1,1} \simeq e - \delta$. En el modelo a simular se supone que dicha distancia es una cantidad aleatoria con media $e - \delta$ y una pequeña desviación típica, σ_6, $Y_{1,1} \sim N(e - \delta, \sigma_6^2)$.

Por último, el ancho de la bulárcama se aproxima mediante $L_2 \simeq L$ entonces $L_2 \sim N(L, \sigma_7^2)$.

Por consiguiente, resumiendo lo anterior

$Y_{1,1}$ viene dada por un valor simulado de una distribución normal ($Y_{1,1} \sim N(e - \delta, \sigma_6^2)$)
$Y_{1,2} = Y_{1,1} + \Delta_1$
$Y_{2,1} = Y_{1,2} + \eta_1$
$Y_{2,2} = Y_{2,1} + \Delta_2$
$Y_{3,1} = Y_{2,2} + \eta_2$
…

Es decir,

$Y_{i+1,1} = Y_{i,2} + \eta_i$
$Y_{i+1,2} = Y_{i+1,1} + \Delta_{i+1}$ con i = 1, …, r-1

En conclusión, se tienen que simular

$Y_{1,1} \sim N(e - \delta, \sigma_6^2)$; $\Delta_i \sim N(a + 2\delta, \sigma_4^2)$ i=1,..., r; $\eta_i \sim N(d - 2\delta - a, \sigma_5^2)$ i=1,..., r-1 y

$L_2 \sim N(L, \sigma_7^2)$ todas ellas independientes.

Para finalizar, quedaría por simular el ángulo del i-ésimo elemento longitudinal con la chapa.
$\theta_i \simeq \frac{\pi}{2}$ i = 1, …, r entonces $\theta_i \sim N\left(\frac{\pi}{2}, \sigma_8^2\right)$.

Tras un cambio de impresiones con los expertos de Navantia los valores elegidos para las desviaciones típicas (σ_1, …, σ_8) en el modelo inicial han sido los siguientes: $\sigma_1 = 0.5$, $\sigma_2 = 0.5$, $\sigma_3 = 0$, $\sigma_4 = 0.1$, $\sigma_5 = 0.1$, $\sigma_6 = 0.1$, $\sigma_7 = 0.1$ y $\sigma_8 = 0.5$.

A continuación, se muestran algunos valores obtenidos de la probabilidad de enhebrado aproximada por simulación mediante este modelo para algunos valores concretos de las desviaciones típicas. Para

calcular esta probabilidad los valores de los parámetros fijos que intervienen en el modelo son los siguientes:
a = 8 (espesor del longitudinal), b = 200 (altura del longitudinal), L=5000 (ancho total del panel y de la bulárcama), ε = 2 (valor máximo permitido entre la diferencia del ancho del panel y el ancho de la bulárcama), $\delta =$ 1.5 (holgura de la escotadura (gap)), e = 650 (distancia desde la línea de trazado del primer longitudinal al canto), d = 650 (distancia entre longitudinales) y r = 4 (número de longitudinales).

Valores de las desviaciones típicas	Probabilidad de enhebrado
Desviaciones típicas elegidas por los expertos de Navantia: $\sigma_1 = 0.5$, $\sigma_2 = 0.5$, $\sigma_3 = 0$, $\sigma_4 = 0.1$, $\sigma_5 = 0.1$, $\sigma_6 = 0.1$, $\sigma_7 = 0.1$ y $\sigma_8 = 0.5$	99.702 %
Se modifica la desviación σ_1 (desviación de la distancia desde la línea de trazado del primer longitudinal al canto) y σ_3 (desviación del ancho del panel) respecto a las desviaciones elegidas por los expertos de Navantia: $\sigma_1 = 1$, $\sigma_2 = 0.5$, $\sigma_3 = 0.5$, $\sigma_4 = 0.1$, $\sigma_5 = 0.1$, $\sigma_6 = 0.1$, $\sigma_7 = 0.1$ y $\sigma_8 = 0.5$	77.732 %
Todas las desviaciones típicas iguales a 0.5: $\sigma_1 = 0.5$, $\sigma_2 = 0.5$, $\sigma_3 = 0.5$, $\sigma_4 = 0.5$, $\sigma_5 = 0.5$, $\sigma_6 = 0.5$, $\sigma_7 = 0.5$ y $\sigma_8 = 0.5$	90.743 %

Tabla 1. Probabilidad de enhebrado para diferentes valores de las desviaciones típicas.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la GAIN (Xunta de Galicia), Navantia (SEPI) y MINECO. La actividad de Salvador Naya, Ricardo Cao y Javier Tarrío ha sido también financiada por el MINECO a través del proyecto MTM2014-52876-R y por la Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 y Centro Singular de Investigación de Galicia ED431G/01), ayudas cofinanciadas por el FEDER.

REFERENCIAS

Okumoto, Y., Matsuzaki, S. I. (1997) Approach to accurate production of hull structures. Journal of ship production, 13 (3), 207-214.

Okumoto, Y. (2001) Dimensional error of assembled blocks. Journal of ship production, 17 (1), 8-15.

Victoria Meizoso, J. (1997) Principios de Ingeniería Naval. Tórtulo Edicións.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

DETECCIÓN DE ERRORES DE MEDIDA EN CONSTRUCCIÓN NAVAL

Déborah Otero[1], Ricardo Cao[2], Álvaro Brage[3], Javier Tarrío-Saavedra[2], Salvador Naya[2]

[1] Unidade Mixta de Investigación Universidade da Coruña – Navantia, Universidade da Coruña.
[2] Grupo MODES, Departamento de Matemáticas, ITMATI, CITIC, Universidade da Coruña.
[3] Navantia, Ferrol.

RESUMEN

El objetivo de este trabajo es proporcionar diferentes métodos para determinar los umbrales o valores críticos a partir de los cuales se detecten errores de medición en las chapas de acero utilizadas para la construcción de buques. Para ello se han utilizado tres métodos geométricos basados en el teorema del coseno, la fórmula de Herón y el cálculo del determinante de Cayley-Menger, respectivamente, para definir los criterios bajo los cuales no existiría error de medida. Se define una variable de discrepancia entre valores medidos y nominales, DIF, y se estima su distribución según el método geométrico utilizado, mediante el estudio de simulación de las dimensiones medidas de la chapa a las que se añade un error uniforme debido a la precisión del operario al medir. Se definen tres escenarios diferentes correspondientes a tres chapas con distintas dimensiones nominales. En cada uno de ellos, a partir de los valores nominales de la chapa se hallan los umbrales que permiten identificar un error en la medición de la pieza. Además, se incluye un estudio de la potencia de cada uno de los tres métodos para detectar errores de medida, así como de los métodos combinados resultantes de ellos, que permite elegir la metodología más eficiente en cada caso concreto.

Palabras y frases clave: Análisis dimensional, Construcción naval, Control estadístico de la calidad, Detección de anomalías, Metrología, Simulación.

1. INTRODUCCIÓN

El control dimensional de piezas en astillero es fundamental durante el proceso de producción de un buque. El control continuo de las dimensiones de las diversas piezas, elementos, bloques de los que se constituye el buque, previene la aparición de los errores de construcción, y reduce los costes de rectificado en elementos constructivos, contribuyendo a la construcción de buques de mayor calidad a un menor coste. Aunque existen diversos trabajos en los que se describen los procesos de control de calidad, entre los que se encuentra el control dimensional, a llevar a cabo en un astillero (Johnson et al., 2004; Paul et al., 2014), el control estadístico de la calidad (Montgomery, 2007) no ha sido aplicado de forma tan intensiva como en otras ramas de la industria como son la industria manufacturera, química o de automoción. De hecho, tal y como recoge la Metodología Seis Sigma, el control dimensional de piezas se llevaría a cabo en las etapas Definir, Medir, Analizar, Controlar y Mejorar, siendo aplicables todas las herramientas del control estadístico de la calidad relacionadas. De particular importancia en este ámbito son los estudios R y R, por los que se evalúa la idoneidad de un sistema de medición estudiando la variabilidad debida a la reproducibilidad y la repetitividad y los gráficos de control por variables, por medio de los cuales se pueden detectar anomalías en el sistema (Montgomery, 2007).

Este trabajo se centra dentro de las acciones de aplicación de la Industria 4.0 al sector naval. Concretamente en el contexto de la Unidad Mixta de Investigación (UMI), que la Universidad de A Coruña (UDC) tiene con el Astillero Navantia, que bajo el nombre de “Astillero 4.0. El astillero del futuro” pretende llevar las ideas de la industria 4.0 a un futuro astillero. El proyecto de la UMI surge a raíz de un trabajo previo de benchmarking que ha sido realizado por el Astillero Navantia y la buena relación ya existente con la UDC, donde varios grupos de investigación venían colaborando puntualmente con unidades de esta empresa. De este estudio del estado del arte, se detecta la necesidad de acometer una línea clara de optimización de los procesos necesarios para la construcción de los buques. El camino a

seguir se enclava en este proyecto de la UMI, en el que colaboran distintos grupos de investigación de la Universidad de A Coruña bajo la denominación de Astillero 4.0. El objetivo fundamental de esta Unidad Mixta es el desarrollo de nuevas técnicas y tecnologías que permitan incrementar la competitividad del astillero, para afrontar el desafío técnico industrial del programa de fragatas F110, mediante la mejora de los procesos productivos existentes y el desarrollo de otros nuevos, de tal forma que se incremente la productividad y se reduzcan tiempos y costes. Entre las acciones de esta unidad mixta se encuentra la del Control Estadístico de la Calidad.

A partir de la necesidad que tiene la industria naval en detectar cuándo una medida tomada por un operario es errónea, surge la idea del siguiente estudio, cuyo objetivo es determinar un umbral a partir del cual se considera que existe un error en la medición de una chapa de acero que será utilizada como elemento básico en la elaboración de los bloques que constituyen el buque. Este problema está por tanto relacionado con la evaluación del sistema de medida y con la detección de anomalías. Cada chapa, caracterizada por seis atributos dimensionales, se inspecciona antes de ser utilizada en sucesivos procesos. El umbral definido permite saber al operario encargado de realizar la medición que las dimensiones medidas de la pieza son correctas o, por lo contrario, se está produciendo un error en la medición. En este último caso, se debe realizar de nuevo la medición de la pieza y corregir el error. Además será especialmente relevante la detección de dicho error de medida especialmente cuando, en el momento de realización de las mediciones, no se disponga de las dimensiones nominales de la pieza que se está midiendo.

2. METODOLOGÍA

En la presente sección se muestra una breve descripción de los métodos utilizados para estimar el umbral a partir del cual se considera que existe un error en la medición que el operario realiza sobre chapas cortadas para su utilización en el astillero.



Figura 1: Esquema de las chapas de acero utilizadas en construcción naval. Es necesario establecer una nomenclatura para cada uno de sus vértices.

Para ilustrar los tres procedimientos propuestos, es necesario denotar previamente los vértices de cada chapa de acero como se indica en la Figura 1, conforme se puedan definir las seis medidas que caracterizan cada chapa de acero (AB, BC, CD, DA, AC y BD). Seguidamente se proponen tres métodos para medir las discrepancias entre las seis mediciones realizadas por un operario y el hecho de que exista una chapa plana de dichas dimensiones.

2.1. Método 1

Consiste en comprobar que los cuatro ángulos interiores de la chapa (no necesariamente rectangular), calculados a partir de los triángulos dados por las diagonales (AC y BD) y sus dos lados adyacentes suman aproximadamente 2π (360°). Para ello se utiliza el teorema del coseno, cuyo enunciado indica que, dado un triángulo cualquiera de lados a, b, c y ángulos $\widehat{A}$, $\widehat{B}$, $\widehat{C}$, opuestos a estos lados, entonces se cumple:

$$c^2 = a^2 + b^2 - 2ab\cos\widehat{C}$$

A partir de las mediciones de AB, DA y BD, basándose en el teorema del coseno, se puede obtener la estimación del ángulo $\widehat{A}$ (Figura 2) de la siguiente forma:

$$\widehat{A} = \arccos\frac{AB^2 + DA^2 - BD^2}{2 \cdot AB \cdot DA}$$

Análogamente se obtiene la estimación de los restantes ángulos

$\widehat{B} = \arccos\frac{AB^2+BC^2-AC^2}{2\cdot AB\cdot BC}$, $\widehat{C} = \arccos\frac{BC^2+CD^2-BD^2}{2\cdot BC\cdot CD}$ y $\widehat{D} = \arccos\frac{CD^2+DA^2-AC^2}{2\cdot CD\cdot DA}$,

a partir de los cuales se calcula la diferencia entre la suma de los ángulos interiores (estimados) y 2π:

$$\text{Dif} = \widehat{A} + \widehat{B} + \widehat{C} + \widehat{D} - 2\pi$$

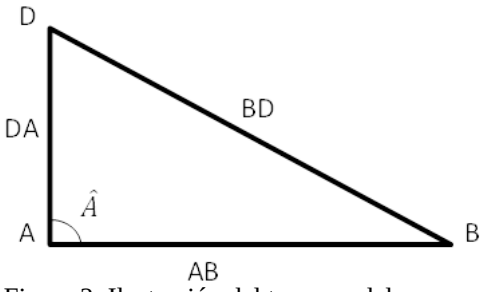


Figura 2: Ilustración del teorema del coseno.

2.2. Método 2

Consiste en comprobar que las dos triangulaciones posibles, obtenidas a partir del cuadrilátero que forma la chapa de acero cortada, están formadas, cada una, por un par de triángulos cuya suma de áreas es la misma, independientemente de la triangulación estudiada. Este segundo método se lleva a cabo utilizando la fórmula de Herón (Fleming y Varberg, 1991), que permite calcular el área de un triángulo conociendo las longitudes sus tres lados a, b y c:

$\text{Área} = \sqrt{s(s-a)(s-b)(s-c)}$, donde s es el semiperímetro del triángulo: $s = \frac{a+b+c}{2}$.

Se comprueba, utilizando la nomenclatura mostrada en la Figura 3, que la suma de las áreas de los triángulos ABC y ACD es aproximadamente la misma que la suma de las áreas de los triángulos ABD y BCD.

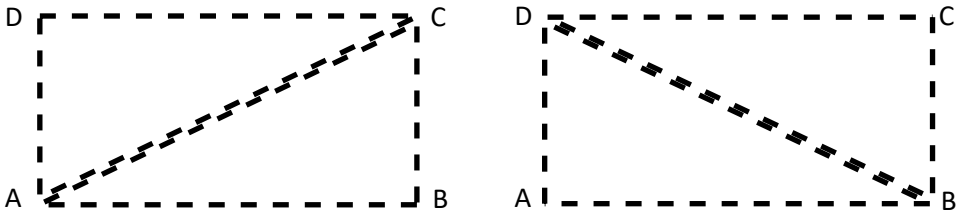


Figura 3: Triangulaciones posibles a partir de la plancha de acero cortada.

A partir de la fórmula de Herón se tiene:

$\text{Área (ABC)} = \sqrt{s_{ABC}(s_{ABC}-AB)(s_{ABC}-BC)(s_{ABC}-AC)}$, siendo $s_{ABC} = \frac{AB+BC+AC}{2}$
$\text{Área (ACD)} = \sqrt{s_{ACD}(s_{ACD}-CD)(s_{ACD}-DA)(s_{ACD}-AC)}$, siendo $s_{ACD} = \frac{CD+DA+AC}{2}$
$\text{Área (ABD)} = \sqrt{s_{ABD}(s_{ABD}-AB)(s_{ABD}-DA)(s_{ABD}-BD)}$, siendo $s_{ABD} = \frac{AB+DA+BD}{2}$
$\text{Área (BCD)} = \sqrt{s_{BCD}(s_{BCD}-BC)(s_{BCD}-CD)(s_{BCD}-BD)}$, siendo $s_{BCD} = \frac{BC+CD+BD}{2}$

A continuación, se calcula la diferencia de las sumas $\text{Dif}_{inicial} = \text{Área 1} - \text{Área 2}$, siendo
$\text{Área 1} = \text{Área(ABC)} + \text{Área (ACD)}$ y $\text{Área 2} = \text{Área(ABD)} + \text{Área (BCD)}$.

Dado que el área de un objeto es una medida 2-dimensional, se considera que lo más adecuado es calcular $\text{Dif} = \frac{Dif_{inicial}}{\text{Área promedio}}$ donde $\text{Área promedio} = \frac{\text{Área 1}+\text{Área 2}}{2}$, para que dicha medida sea adimensional.

2.3. Método 3

Se basa en comprobar que el volumen del tetraedro que tiene vértices en ciertos cuatro puntos, y cuyas aristas se corresponden con las seis mediciones obtenidas, es aproximadamente cero. Una forma de hallar el volumen del tetraedro (3-símplex) es mediante los determinantes de Cayley-Menger. Si se denota por CM al determinante de Cayley-Menger (Fleming y Varberg, 1991), entonces el n-volumen del n-símplex es

$$\sqrt{\frac{(-1)^{n+1}}{2^n(n!)^2}|CM|}$$

En el caso de dimensión 3, el que se está tratando, dicho volumen se calcula mediante la siguiente expresión:

$$\text{Vol} = \sqrt{\frac{(-1)^4}{2^3(3!)^2}|CM3|} = \sqrt{\frac{1}{8\cdot 36}|CM3|} = \sqrt{\frac{|CM3|}{288}},$$

siendo CM3 el determinante de Cayley-Menger para 3 dimensiones, definido por

$$\det\begin{bmatrix} 0 & AB^2 & AC^2 & DA^2 & 1 \\ AB^2 & 0 & BC^2 & BD^2 & 1 \\ AC^2 & BC^2 & 0 & CD^2 & 1 \\ DA^2 & BD^2 & CD^2 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

Para adimensionalizar el valor del volumen se considera calcular

$$\text{Dif} = \frac{\text{Vol}}{\text{Área promedio}^{3/2}}$$

3. ESTUDIO DE SIMULACIÓN Y RESULTADOS

Mediante el presente estudio de simulación se pretende estimar la distribución de *Dif* cuando no existe error en la medición de los operarios, es decir, cuando las medidas son correctas, además de medir la potencia de los tres métodos propuestos.

3.1. Estimación de la distribución de *Dif* y valores umbrales para el error de medida

En esta sección se presenta el esquema de simulación para estimar los valores o límites críticos a partir de los cuales se identificaría mediciones erróneas. Para cualquiera de los tres métodos definidos se asume que no habrá un error en la medición cuando *Dif* tome un valor pequeño. Para ello, dados los valores nominales de las seis dimensiones de una chapa, se suma a cada una de esas mediciones seis cantidades distintas (e_1, e_2, e_3, e_4, e_5, e_6) simuladas según una U(-δ, δ), que está asociada a la incertidumbre con la que los operarios realizan la medición (precisión de medida). El umbral considerado para detectar un error en la medición será el percentil que se considere oportuno seleccionar (95%, 97.5%, ó 99%, por ejemplo) de la distribución empírica de los valores simulados de *Dif*.

Se ha considerado $\delta = 0.5$ *mm* (precisión de las medidas realizados por los operarios) y se han calculado los percentiles con distintos valores nominales de las dimensiones. En la Tabla 1 se muestran las dimensiones nominales de las tres chapas consideradas, correspondientes a tres tipos de chapa de acero reales usadas para la construcción de buques en las instalaciones de Navantia (Ferrol).

Chapa 1	Chapa 2	Chapa 3
AB = CD = 11600	AB = CD = 6050	AB = CD = 3034
BC = DA = 2350	BC = DA = 2350	BC = DA = 2951
AC = BD = 11835.6	AC = BD = 6490.4	AC = BD = 4232.4

Tabla 1. Valores nominales de las dimensiones de tres chapas estudiadas.

Se ha calculado el valor de diferentes percentiles (0.5%, 2.5%, 5%, 90%, 95%, 97.5%, 99%, 99.5%) de la distribución empírica correspondiente a los valores simulados de *Dif*. Para los dos primeros

métodos, la discrepancia con la medida real es bilateral, es decir, los *Dif* simulados pueden tomar tanto valores pequeños (negativos) como grandes (positivos), sin embargo, para el método 3 es unilateral (los valores de *Dif* siempre son números positivos). Todos los resultados obtenidos se resumen en la Tabla 2, en la que se muestran los valores de los percentiles de *Dif*, estimados a partir de la simulación de 1 000 000 y 10 000 000 vectores 6-dimensionales de medidas (correspondientes a cada chapa de acero), obtenidos por cada uno de los métodos descritos y para cada tipo de chapa.

		Método 1						**Método 2**			**Método 3**		
		Resultados en radianes			Resultados en grados sexagesimales								
Percentil	***n*: número de chapas simuladas**	*Chapa 1*	*Chapa 2*	*Chapa 3*	*Chapa 1*	*Chapa 2*	*Chapa 3*	*Chapa 1*	*Chapa 2*	*Chapa 3*	*Chapa 1*	*Chapa 2*	*Chapa 3*
0.5%	1000000	-0.00131	-0.00126	-0.00143	-0.07498	-0.07197	-0.08199	-5.94e-08	-6.50e-08	-7.104e-08			
	10000000	-0.00131	-0.00126	-0.00143	-0.07487	-0.07194	-0.08195	-5.93e-08	-6.48e-08	-7.098e-08			
2.5%	1000000	-0.00105	-0.00098	-0.00114	-0.06038	-0.05643	-0.06531	-4.13e-08	-4.50e-08	-4.84e-08			
	10000000	-0.00105	-0.00098	-0.00114	-0.06023	-0.05625	-0.06520	-4.13e-08	-4.499e-08	-4.85e-08			
5%	1000000	-0.00091	-0.00083	-0.00098	-0.05202	-0.04760	-0.05607	-3.22e-08	-3.51e-08	-3.77e-08			
	10000000	-0.00091	-0.00083	-0.00098	-0.05194	-0.04750	-0.05601	-3.22e-08	- 3.50e-08	-3.77e-08			
90%	1000000										0.00483	0.00492	0.00503
	10000000										0.00482	0.00492	0.00502
95%	1000000	0.00075	0.00091	0.00082	0.04299	0.0 5229	0.04718	3.21e-08	3.51e-08	3.76e-08	0.00523	0.00534	0.00546
	10000000	0.00075	0.00091	0.00082	0.04291	0.05220	0.04713	3.22e-08	3.50e-08	3.77e-08	0.00523	0.00534	0.00545
97.5%	1000000	0.000895	0.001065	0.00098	0.05129	0.06102	0.05637	4.12e-08	4.496e-08	4.84e-08			
	10000000	0.000894	0.001064	0.00098	0.05121	0.06097	0.05631	4.13e-08	4.496e-08	4.85e-08			
99%	1000000										0.00588	0.00601	0.00617
	10000000										0.00588	0.00601	0.00616
99.5%	1000000	0.00115	0.00134	0.00128	0.0 6593	0.07671	0.07311	5.91e-08	6.47e-08	7.08e-08			
	10000000	0.00115	0.00134	0.00128	0.06587	0.07666	0.07311	5.93e-08	6.48e-08	7.096e-08			

Tabla 2: Percentiles de la distribución correspondiente a la discrepancia con la medida correcta, *DIF*.

No se observan muchos cambios en los valores de los percentiles al aumentar el número de piezas simuladas de un millón a diez millones. Pero, desde el punto de vista de la empresa, es importante saber si existen diferencias significativas entre los umbrales o límites proporcionados por tres métodos al cambiar las medidas nominales de la chapa. Con tal fin, se han estimado las funciones de densidad de la variable Dif obtenida por los tres métodos propuesto y para cada uno de los tres tipos de chapa que se diferencian en sus dimensiones (Figura 4). Para ello se ha utilizado el estimador kernel gausiano y ventana obtenida por métodos plug in (Sheather y Jones, 1991; Silverman, 1984). La Figura 4 muestra que, para distintos valores nominales de las seis medidas de la chapa, las funciones de densidad de los valores de los Dif simulados (usando n=10 000 000) son muy parecidas cuando se calcula la discrepancia entre las medidas tomadas con el método 2, sin embargo, con los otros dos métodos se aprecian ligeras diferencias. Esta afirmación se puede corroborar con la prueba U de Mann-Whitney (suma de rangos de Wilcoxon), sin asumir ningún supuesto acerca de la distribución de probabilidad de la población de la que se obtienen las muestras. La prueba U de Mann-Whitney se utiliza para contrastar si dos muestras, extraídas independientemente, proceden de la misma población. Es decir, determina si las dos muestras independientes son homogéneas y tienen la misma distribución. En este caso de estudio particular, mediante el test U, y para cada método, se contrasta si la función de distribución de los valores de Dif simulados es igual dependiendo de las dimensiones nominales de la chapa de acero. En la Tabla 3 se

muestra el p-valor obtenido correspondiente al contraste para cada uno de los métodos con dos de las medidas nominales. Como se puede observar, los valores de los Dif simulados obtenidos a partir del método 2 tienen la misma distribución independientemente del valor nominal de la chapa, el p-valor es mayor que cualquier nivel de significación estándar. Sin embargo, utilizando los métodos 1 y 3, la función de distribución es diferente al cambiar las dimensiones nominales de la chapa, obteniendo p-valores menores que cualquier nivel de significación estándar.

En consecuencia, para determinar el umbral a partir del cual se considera un error en la medición, es recomendable utilizar el método 2, pues de los métodos descritos es con el único que se obtienen resultados similares independientemente del valor nominal de la chapa. Dependiendo de lo conservadores que pretendamos ser con la hipótesis nula, se elegirá el percentil que se considere más adecuado. Por ejemplo, si se quiere que el 90% de las chapas tengan una medición correcta, se tomará como umbrales los valores de los percentiles 5 y 95 para los métodos 1 y 2, o el valor del percentil 90 para el método 3.

Método 1		Método 2		Método 3	
Chapa1 - Chapa2	Chapa1 - Chapa3	Chapa1 - Chapa2	Chapa1 - Chapa3	Chapa1 - Chapa2	Chapa1 - Chapa3
2.2e-16	1.068e-05	0.9756	0.8897	2.2e-16	2.2e-16

Tabla 3: *p*-valores correspondientes a la aplicación de la prueba U para contrastar si existen diferencias entre las distribuciones de DIF dependiendo de las dimensiones de las chapas de acero.

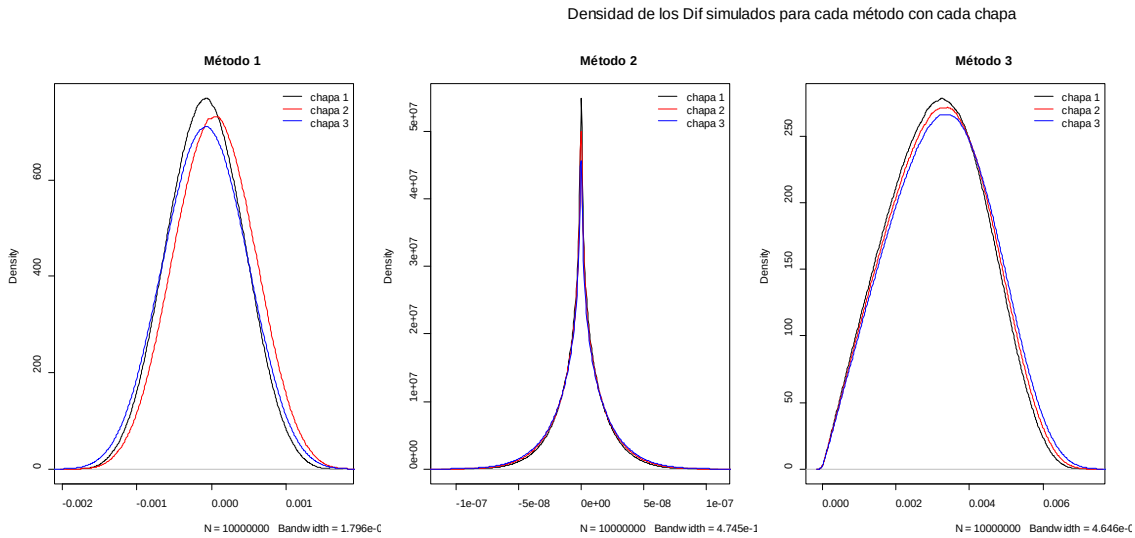


Figura 4: Función de densidad de la discrepancia con la medida correcta, *DIF*, para cada método y tipo de chapa de acero al que se aplica.

3.2. Estudio de la potencia de los tres métodos propuestos para detectar errores de medida

Con el fin de identificar cuál de los tres métodos propuestos anteriormente es mejor para determinar el umbral que se está buscando, en términos de sensibilidad de detección de cambios, se ha estudiado la potencia asociada a dichos métodos. Para ello se simula un error de medida en una de las dimensiones y, posteriormente, se calcula, por simulación, la probabilidad de que cada uno de los métodos detecte dicho error, es decir, que el valor de *Dif* correspondiente esté más allá de los umbrales calculados por el método 1, 2 ó 3. Para ello se han definido tres escenarios correspondientes a las tres chapas consideradas, cuyas longitudes nominales se muestran en la Tabla 1. En cada uno de ellos se simula un error de medida en una de las dimensiones. Se han alterado tres dimensiones de la chapa, AB (largo), BC (ancho) y AC (diagonal), y se han considerado diferentes magnitudes del error, en el intervalo (-2 mm, 2 mm), siguiendo las indicaciones del personal de Navantia. Para calcular la probabilidad de que cada método detecte el error de medida inducido se han simulado 100 000 chapas.

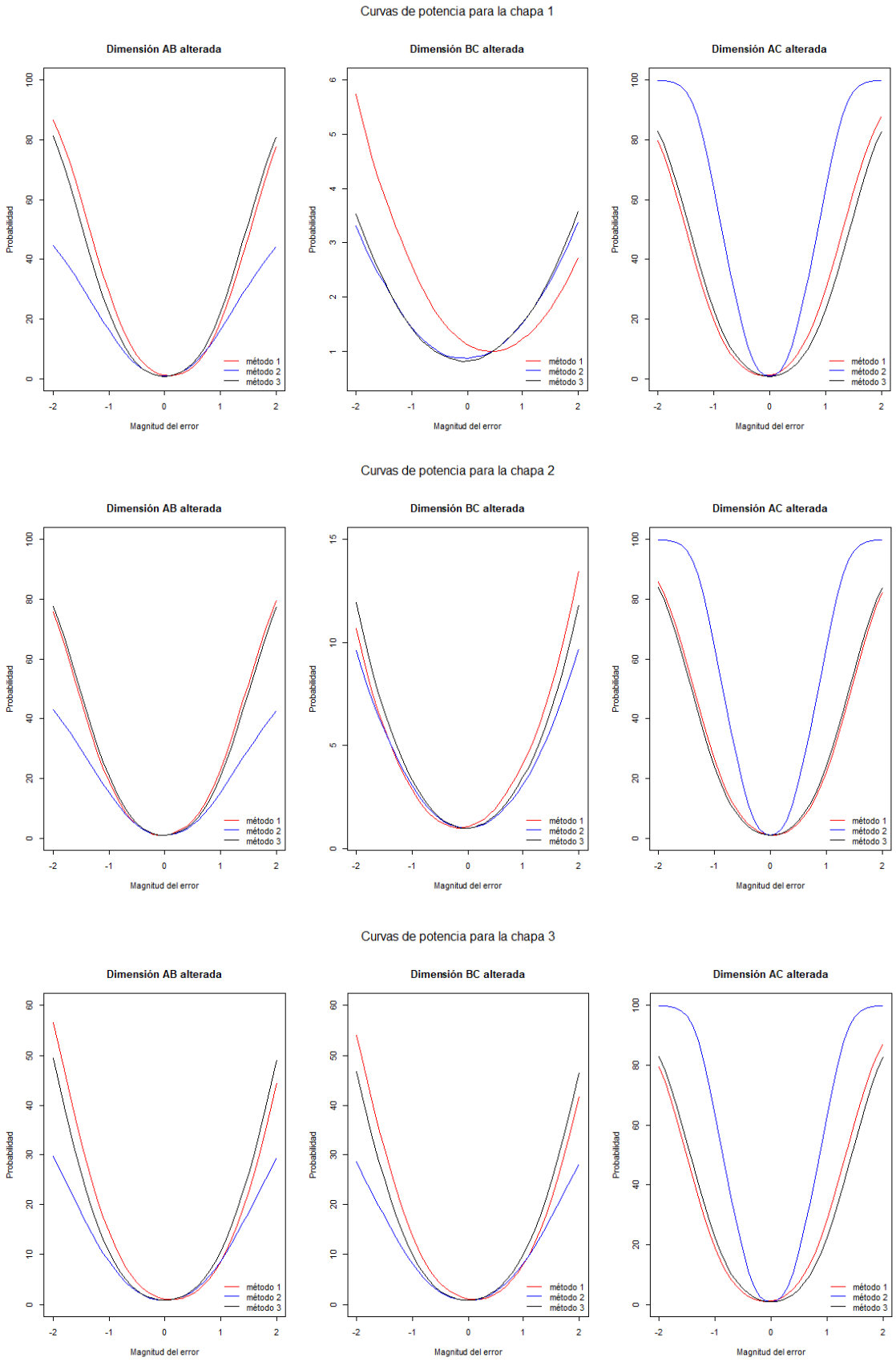


Figura 5: Curvas de potencia para los tres métodos, alterando las dimensiones AB, BC y AC de las tres chapas de acero estudiadas.

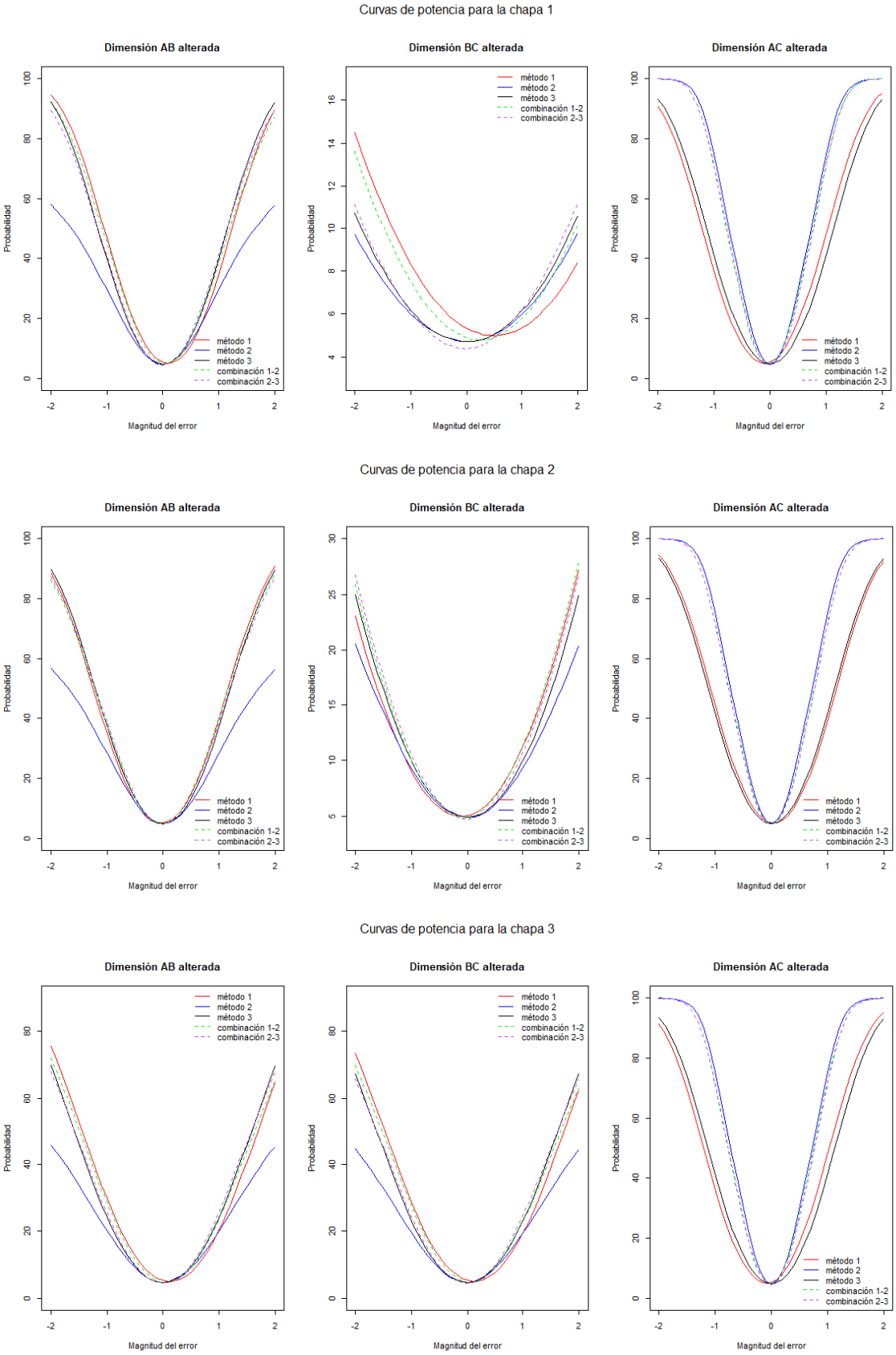


Figura 6: Curvas de potencia para los tres métodos y para la combinación de los mismos dos a dos, alterando las dimensiones AB, BC y AC de las tres chapas de acero estudiadas.

A partir de las curvas de potencia que se muestran en las distintas gráficas de la Figura 5, las cuales se encuentran en escalas diferentes, se puede concluir que el método mejor a utilizar para determinar un umbral a partir del cual se considera un error en la medición depende de la dimensión en la que ha se ha producido dicho error. Los resultados obtenidos son similares para distintas longitudes nominales de la chapa. Se observa que los métodos 1 ó 3 son la mejor opción si el error de medida se produce en la dimensión AB o BC, sin embargo, si el error está en la dimensión AC el método 2 sería más potente.

Ante estos resultados, no se puede identificar cuál de los tres métodos propuestos es mejor usar pues no se sabe en qué dimensión se va a cometer el error en el momento de medir la chapa. Por tanto, se propone utilizar un método combinado, es decir, usar el método 1 y 2 simultáneamente o el método 2 y 3 simultáneamente para determinar el umbral de medición. La combinación entre el método 1 y el método 3 se considera innecesaria ya que las potencias de ambos métodos son similares.

Por tanto, avanzando en el análisis para determinar cuál es el mejor método a utilizar para obtener un umbral que indique un error en la medición de la chapa, se muestran los gráficos donde se comparan las curvas de potencia de los tres métodos definidos con los métodos combinados propuestos anteriormente (Figura 6). Se ha realizado la combinación del método 1 y 2 simultáneamente y del método 2 y 3 simultáneamente. Estos métodos combinados identifican como error de medida cuando uno de los dos métodos lo identifica como tal. Es decir, se rechaza que la chapa esté bien medida cuando al menos uno de los dos métodos se salga de los umbrales permitidos. Para ello se fijó un nivel de significación, α, igual a 0.05 (5%) para los métodos no combinados y se consideró $\frac{\alpha}{2} = 2.5\%$ para los métodos combinados. Esta última consideración está motivada por la desigualdad de Bonferroni, según la cual

$$P(RC|_{MC}) = P(R1 \text{ U } R2|_{MC}) \leq P(R1|_{MC}) + P(R2|_{MC}) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

siendo los sucesos

MC: las medidas son correctas

RC: el método combinado rechaza que las medidas sean correctas

R_1: el método 1 rechaza que las medidas sean correctas

R_2: el método 2 rechaza que las medidas sean correctas

Se observa que los dos métodos combinados propuestos tienen un comportamiento parecido en los tres escenarios analizados. En cada uno de ellos, para las tres dimensiones de las chapas (AB, BC y AC) alteradas con un error de medida, los métodos combinados funcionan tan bien como el mejor de los dos que se combinan. La potencia del método combinado 1-2 es prácticamente igual de buena que la potencia del método 1, cuando se altera la dimensión AB o BC, y casi igual a la potencia del método 2, cuando se altera AC. Tanto el método 1 como el método 2 eran los mejores métodos en cada caso. Resultados similares se observan con el método combinado 2-3, la potencia de este método combinado es similar a la potencia del método 3, cuando se altera la dimensión AB o BC, y a la potencia del método 2, cuando se altera AC. Para finalizar, basándose en los resultados anteriores, aunque ambos métodos combinados se comportan de forma similar, se propone usar el método combinado 1-2 para determinar el umbral de medición.

AGRADECIMIENTOS

Este trabajo ha sido financiado por la GAIN (Xunta de Galicia), Navantia (SEPI) y MINECO. La actividad de Salvador Naya, Ricardo Cao y Javier Tarrío ha sido también financiada por el MINECO a través del proyecto MTM2014-52876-R y por la Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 y Centro Singular de Investigación de Galicia ED431G/01), ayudas cofinanciadas por el FEDER.

REFERENCIAS

Fleming, W., Varberg, D. (1991) Álgebra y trigonometría con geometría analítica. Pearson Educación.

Paul, E., Babu, D. E., Paul, B. (2014) Design of an Accuracy Control System in Ship Building Industry, 5, 504-509.

Johnson, G. W., Laskey, S. E., Robson, S., Shortis, M. R. (2004) Dimensional & accuracy control automation in shipbuilding fabrication: an integration of advanced image interpretation, analysis, and visualization techniques. Proceedings of Analysis and Visualization Techniques, 20th Congress ISPRS, Istanbul, Turkey.

Montgomery, D. C. (2007) Introduction to statistical quality control. John Wiley & Sons.

Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society series B, 53, 683–690.

Silverman, B. W. (1986) Density Estimation. London: Chapman and Hall.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

A novel distributed multimethod to solve challenging global optimization problems in Computational Systems Biology

David R Penas[1], Patricia González[2], Ramón Doallo[2] and Julio R Banga[3]

[1]Institute of Mathematics, Universidade de Santiago de Compostela, Santiago de Compostela
[2]Computer Architecture Group, Universidade da Coruña, A Coruña, Spain
[3]BioProcess Engineering Group, IIM-CSIC, Vigo, Spain

Abstract

Mathematical modelling is playing an increasingly important role in medical and biological research, as indicated by the many research efforts in the areas of Bioinformatics, Computational Biology and Systems Biology. Many problems in these areas can be formulated and solved using a mathematical optimization framework. A prominent example is the development of large-scale dynamic models of biological systems, where the problem of model calibration presents many challenges due to its multimodality and ill-conditioning.

Approximate global optimization methods, such as metaheuristics, can be used to solve these problems, although at the expense of a very large computational cost. Recent studies (Penas et al, 2015, 2017; Teijeiro et al, 2017), have focused on exploiting parallelism in order to obtain promissing results faster, using the features of current parallel computational infrastructures (many cores, multi-processor architectures, cloud computing or distributed programming). However, it is known that there is no general global optimization method that can perform equally well in on all sorts of problems. In general, the performance of a given global optimization method can be problem dependent.

Our recent work is focused on combining different methods and exploiting parallelism at the same time. In other words, we are developing a self-adaptive distributed multimethod where (i) multiple optimization methods are executed concurrently, (ii) they cooperate exchanging promising solutions, and (iii) they self-adapt their configuration during run time based on the observed performance of the individual methods. The goal is to obtain a very robust optimization tool, and improve the quality of the reported solution through a fair combination of diversification and intensification in the search. Preliminary tests using a set of challeging benchmarks from the Computational Systems Biology field show very encouraging results.

Keywords: Global Optimization, Parallel Cooperative Methods, Multimethod, Distributed Algorithms.

References

Penas, D.R., Banga, J.R., Gonzlez, P., and Doallo, R. (2015) Enhanced parallel Differential Evolution algorithm for problems in computational systems biology. Applied Soft Computing, 33, 86–99.

Penas, D.R., Henriques, D., Gonzlez, P., Doallo, R., Saez-Rodriguez, J., and Banga, J.R. (2017) A parallel metaheuristic for large mixed-integer nonlinear dynamic optimization problems, with applications in computational biology. PLOS ONE, 12(8).

Teijeiro, D., Pardo, X., Penas, D.R., Gonzlez,P. , Banga, J.R., and Doallo, R. (2017) A cloud-based enhanced differential evolution algorithm for parameter estimation problems in computational systems biology. Cluster Computing, 1-14.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Nuevas Necesidades de Previsión en los Actuales Marcos Presupuestarios

Marcos Rodríguez Rey[1]

[1]Instituto Galego de Estatística

RESUMEN

Los nuevos marcos normativos en los que se mueven las administraciones públicas al elaborar los prespuestos las obliga a implementar modelos que les permitan determinar la posición cíclica de la economía y su efecto sobre los presupuestos. Otra obligación derivada de dichos marcos es la obligación de obtener unas previsiones de ingresos tributarios. Este trabajo describe la aproximación que sobre estas tareas se ha tomado en Europa, España y Galicia.

Palabras y frases clave: predicción de series de tiempo, modelos de corrección del error, output gap.

1. INTRODUCCIÓN

La actual normativa europea sobre marcos presupuestarios establece la necesidad de introducir la planificación plurianual en la política económica de todas las administraciones públicas basada en los principios de estabilidad presupuestaria y sostenibilidad financiera.

Su cumplimiento en España se desarrolla con la Ley orgánica 2/2012 del 27 de abril, de Estabilidad Presupuestaria y Sostenibilidad Financiera. En Galicia por la Ley 2/2011, del 16 de junio de 2011, de Disciplina Presupuestaria y Sostenibilidad Financiera.

El principio de estabilidad presupuestaria se consigue cuando una administración pública no incurre en déficit estructural (el déficit no afectado por los componentes cíclicos de la coyuntura económica). Además, el gasto no podrá aumentar por encima de un valor de referencia, que en el caso de España se estableció para todas las administraciones públicas en la tasa de crecimiento potencial del Producto Interior Bruto del conjunto de España.

En este contexto, surge la necesidad de estimar el crecimiento potencial y el output gap de la economía gallega con un doble objetivo. Por una parte, comparar el crecimiento potencial de la economía gallega con la española, y por otra, estimar la descomposición del déficit en cíclico y estructural.

Otra de las necesidades que surge en este escenario es tratar de prever como evolucionarán los principales ingresos de la adminstración autónomica: Los ingresos tributarios. Además, existen otras posibles utilidades como la medición de tensiones inflacinistas, análisis de los factores de crecimiento, distinguiendo si son permanentes o transitorios.

Según la Orden del 4 de julio de 2016, en la que se dictan instrucciones para la elaboración de los escenarios presupuestarios de la Comunidad Autónoma de Galicia, el Instituto Galego de Estatística (IGE) es el organismo encargado de determinar estas variables para la economía gallega. Además, participará en la estimación de la de la recaudación de las principales figuras tributarias.

En el apartado 2 se trata de la predicción de ingresos tributarias mediante los modelos de corrección del error, según la metodología empleada por la Autoridad Independiente de Responsabilidad Fiscal.

El análisis de la posición cíclica y del output gap de la economía gallega se trata en el apartado 3 siguiendo la metodología que emplea la Comisión Europea en su análisis conjunto de la unión y de sus países miembros.

Por último, es importante mencionar que la estimación de estos modelos no sería posible sin el sólido Sistema de Cuentas(cuentas económicas anuales, cuentas económicas trimestrales,...) que el IGE realiza; y que permite disponer de un marco de variables de referencia contrastado desde diferentes ópticas con unos estándares de calidad muy altos en el ámbito regional.

2. PREDICCIÓN DE INGRESOS POR IMPUESTOS MEDIANTE MODELOS DE CORRECCIÓN DEL ERROR

Es frecuente realizar las predicciones de los ingresos tributarios para corto y media plazo, de 1 a 3 años, mediante los modelos de corrección del error. Empleando variables relacionadas para las que ya existen predicciones en el cuadro macroeconómico.

Estos modelos aprovechan una posible relación de cointegración entre las variables. Es decir, que aunque las variables sean no estacionarias, como es frecuente en economía, existe una relación lineal entre ellas que sí lo es. Mostrando una relación estable entre ellas a largo plazo.

Según plantearon Engle y Granger (1987), los modelos de corrección del error, en el caso de una variable explicada, consisten en dos ecuaciones de regresión para relacionar K variables explicativas y una explicada. Todas ellas no estacionarias e integradas de orden 1. Sin embargo, como ya se dijo, sí están cointegradas. Este supuesto debe de ser comprobado antes de realizar la estimación, por ejemplo con el test aumentado de Dickey y Fuller (ADF).

$$y_t = \sum_{i=1}^{K} \alpha_i x_{it} + \epsilon_{1t} \tag{1}$$

$$\Delta y_t = \sum_{i=1}^{K} \alpha_i \Delta x_{it} + \beta e_{t-1} + \epsilon_{2t} \tag{2}$$

La primera ecuación recoge la relación de largo plazo entre las variables. Pero, la novedad del procedimiento en dos pasos de Engel-Granger consiste en permitir la existencia de desajustes en el corto plazo contemplados en la segunda ecuación. En ésta aparecen las variables explicada y explicativas en diferencias pero, además se añaden los errores estimados de la primera ecuación retardados un período. De manera que el parámetro β captura la velocidad del ajuste en el corto plazo. Este parámetro tiene que ser negativo sino el sistema se alejaría constantemente de su equilibrio de largo plazo.

El análsis a posteriori de los residuos de las regresiones puede indicar que es necesario transformar el modelo, ya sea porque presentan una estructura autorregresiva o porque se alejan de las realizaciones esperables de un proceso ruido blanco. Entre las transformaciones más frecuentes se encuentran incluir en la ecuación retardos de las variables o realizar una transformación logarítmica de las mismas. Para comprobar si los residuos no presentan raíces unitarias con el test ADF, y por tanto las variables están cointegradas, hay que recurrir a los valores tabulados por MacKinnon para el caso de las regresiones.

El estudio exhaustivo de Duy y Thoma (1998) encuentra una buena capacidad predictiva en los modelos de corrección de error respecto a otros modelos dinámicos univariantes si se cumplen las condiciones expuestas. Por otra parte Engle y Granger (1987) encontraron que este modelo puede ser estimado consistentemente por mínimos cuadrados ordinarios (MCO).

A continuación empleamos el enfoque metodológico de la Autoridad Independiente de Responsabilidad Fiscal (AIReF) recogido en Cuerpo y Losada (2015), Aplicado al caso del Impuesto de las Personas Físicas (IRPF), el tributo con más capacidad recaudatoria. Pero, es fácilmente generalizable a otros tributos. La figura 1 representa la evolución de los ingresos devengados por los principales impuestos en el territorio fiscal común.

En el esquema propuesto la previsión de los ingresos devengados se subdivide en tres factores explicativos. En la ecuación 3 los ingresos devengados (ID) se descomponen en:

- VM, una variable relevante del cuadro macroeconómico.
- $\frac{BI}{VM}$, la relación existente entre la base imponible del impuesto y la variable macroeconómica.

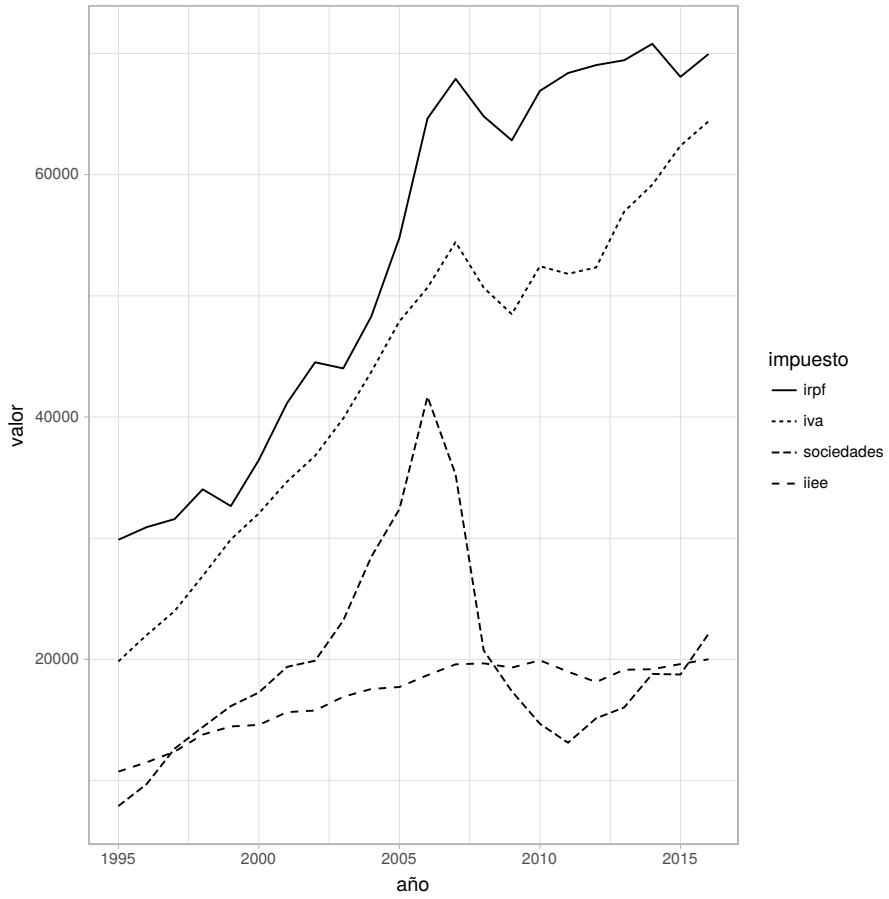


Figura 1: Ingresos devengados por los principales impuestos, millones de euros.

- $\frac{ID}{BI}$, indica el tipo efectivo medio del impuesto.

$$ID = VM \frac{BI}{VM} \frac{ID}{BI} \tag{3}$$

Si estamos interesados en los ingresos líquidos (IR), podemos introducir un factor adicional que recogería la mecánica recaudatoria, es decir, la diferencia entre el devengo de los impuestos y el momento en que se ingresan (ecuación 4).

$$IR = ID \frac{IR}{ID} \tag{4}$$

El proceso de previsión de los ingresos tributarios se realiza de una manera secuencial:

1. Se estima la relación entre las bases imponibles y las variables relevantes del cuadro macroeconómico. Es frecuente el empleo de los modelos de corrección del error. Empleando la relación estimada y las predicciones disponibles de las variables macroeconómicas, podemos predecir las bases imponibles.

2. Una vez predichas las bases imponibles se aplican los tipos efectivos para obtener una proyección de los impuestos devengados en una año concreto. En años de reforma fiscal, en los que se prevea una modificación de los tipos impositivos,... sería necesario realizar una simulación del impacto de dichos cambios en los tipos efectivos.

Existen impuestos que por su naturaleza no pueden ser tratados de forma homogénea. Sino que hay que aplicar el planteamiento precedente por separado a los diferentes ámbitos del impuesto.

Por ejemplo, el IRPF grava a las rentas del trabajo, a rentas del capital e incluso a rentas de la empresa (autónomos). Pero, incluso dentro de las rentas del trabajo la variable macroeconómica que mejor explica el comportamientos de las bases imponibles de los asalariados (remuneración de asalariados) no es la misma que ayuda a explicar las bases imponibles del gasto en pensiones.

La tabla 1 recoge las bases imponibles del IRPF separadas por los diferentes ámbitos del impuesto para el último año con datos definitivos según publica la Agencia Tributaria.

		millones de euros	%
Rentas del trabajo	Salarios	311.279	53,8
	Pensiones	123.380	21,3
	Otras Rentas	72.306	12,5
Rentas del capital	Capital mobiliario	18.011	3,1
	Inmuebles arrendados	16.872	2,9
	Ganancias patrimoniales	10.405	1,8
Rentas de la empresa		26.626	4,6

Tabla 1: Bases imponibles del IRPF en el territorio fiscal común. 2014

A modo de ejemplo: Los modelos de corrección del error que vinculan los ingresos devengados por los salarios con el número de asalariados a tiempo completo equivalentes y la remuneración media de la contabilidad nacional como variables explicativas, obtienen una elevada capacidad predictiva.

3. OUTPUT GAP

El output gap o brecha de producción se define como la relación porcentual entre la diferencia del PIB observado, o previsto, en terminos reales y del PIB potencial en relación al PIB potencial.

$$output\,gap = \frac{PIB\,observado - PIB\,potencial}{PIB\,potencial} 100 \tag{5}$$

La estimación del output gap es una tarea difícil porque el PIB potencial no es una variable observable. La metodología de la Comisión Europea (CE), en vez de realizar supuestos sobre las propiedades de la tendencia y el ciclo de la producción como una serie de tiempo, realiza supuestos basados en la teoría económica empleando el enfoque de la función de producción. Havik et al. (2014) recogen una exhaustiva actualización de la metodología empleada. Este enfoque tiene la ventaja de vincular las relaciones entre las variables con la teoría económica. Sin embargo, también obliga a realizar supuestos sobre la especificación de la función de producción, los rendimientos a escala o el progreso técnico, entre otros.

La especificación adoptada para la función de producción es Cobb Douglas con los factores trabajo (L) y capital (K). Además se incluye para cada uno de ellos dos factores correctores: nivel de capacidad utilizada (U) y nivel de eficiencia (E). Denominamos Y a la producción / PIB.

$$Y = (U_L L E_L)^{\alpha} (U_K K E_K)^{1-\alpha} = L^{\alpha} K^{1-\alpha}\, PTF \tag{6}$$

donde la productividad total de los factores (PTF) viene dada por el residuo de Solow:

$$PTF = (E_L^{\alpha} E_K^{1-\alpha})(U_L^{\alpha} U_K^{1-\alpha}) \tag{7}$$

que sintetiza tanto el grado de utilización de los factores productivos como su nivel tecnológico.

Esta especificación implica asumir el supuesto de una elasticidad de sustitución de los factores constante y unitaria. Si unimos a ello al supuesto de un grado elevado de competencia en los mercados, la elasticidad-producto de los factores productivos, α y $1-\alpha$ para el trabajo y el capital, representan su participación en las rentas. La estabilidad de estas participaciones es un hecho observado por la macroecnomía en las últimas décadas (hechos estilizados). La CE estima la participación del salario en las rentas para el conjunto de los estados miembros de forma común

y lo aplica en las estimaciones individuales de los distintos países. En el IGE empleamos la misma estimación empleada por la propia CE, que le asigna a α un valor de 0,63. La elasticidad-output del capital, $1-\alpha$, queda en 0,37.

La PTF no es directamente observable pero, para el caso de Galicia, se obtiene de las series de Cuentas Económicas difundidas por el IGE mediante la siguiente expresión:

$$PTF = \frac{PIB}{L^{\alpha}K^{1-\alpha}} \tag{8}$$

Para el empleo (L) se utiliza la variable de total de horas trabajadas, mientras que para obtener la serie del stock de capital se emplea el método del inventario permanente.

$$K_t = K_{t-1}(1-\delta) + I_t \tag{9}$$

Para el cálculo del output gap necesitamos estimar el PIB potencial, y éste lo obtenemos evaluando la función de producción con los factores potenciales en el período.

$$PIB_{pot} = PTF_{pot}L_{pot}^{\alpha}K_{pot}^{1-\alpha} \tag{10}$$

El uso potencial del factor trabajo es el más fácil de definir. Ya que el máximo potencial en la contribución del captial a la producción viene dado por la utilización plena del stock de capital en la economía. Además, como el stock de capital es un buen indicador de la capacidad del factor no tiene sentido suavizar su serie. Las series de inversión sí son muy volátiles, sin embargo representan una pequeña parte del stock de capital en un ejercicio determinado.

La estimación de la PTF potencial parte del hecho de que está compuesto de progreso técnico y el grado de utilización de los factores, como puede verse en (7). El primero es inobservable, pero del segundo sí que existen mediciones por medio de las encuestas de clima empresarial. Con estos supuestos se formula como un modelo bivariante de componentes inobservados, se obtiene como suma de dos componentes no observables. Uno representando el ciclo y otro la tendencia.

$$\begin{aligned} ptf_t &= p_t + c_t \\ u_t &= \mu_u + \beta c_t + e_{Ut} \\ e_{Ut} &= \delta_U e_{Ut-1} + a_{Ut} \qquad V(a_{Ut}) = V_U \end{aligned} \tag{11}$$

donde las letras minúsculas indican logaritmos de las equivalentes mayúsculas. El sistema anterior debe de ser completado con supuestos sobre las componente inobservadas para poder ser estimado:

$$\begin{aligned} \Delta p_t &= \mu_{t-1} \\ \mu_t &= \mu_p(1-\rho) + \rho\mu_{t-1} + a_{\mu t} \qquad V(a_{\mu t}) = V_{\mu} \qquad 0 <= \rho <= 1 \\ c_t &\sim AR(2) \end{aligned} \tag{12}$$

μ_t se formula como una tendencia amortiguada, donde μ_p corresponde al crecimiento promedio de la serie.

Expresado el modelo en Espacio de Estados y mediante el filtro de kalman puede ser estimado por máxima verosimilitud o por procedimientos bayesianos. Este último tiene la ventaja de que se puede incorporar información disponible a priori en el modelo, como la periodicidad del ciclo o la inercia de la tendencia.

El factor trabajo potencial se define como el total de horas trabajadas al año por los ocupados en la economía en el supuesto de plena utilización de los recursos disponibles y sin tensiones de inflación.

$$L_{pot} = POB_{pot}Tact_{pot}(1-NAWRU)H_{pot} \tag{13}$$

donde:

- POB_{pot}, representa a la población potencial de 16 a 64 años, eliminando la componente cíclica mediante el filtro de Hodrick-Prescott con parámetro $\lambda = 10$.

- $Tact_{pot}$, es la tasa de actividad potencia de 16 a 64 años, suprimiendo el ciclo de la serie por el mismo procedimiento de la variable anterior.
- H_{pot}, son las horas potenciales trabajadas al año por un trabajador, empleando el filtro de Hodrick-Prescott en las mismas condiciones que en las dos primeras variables.
- $NAWRU$, es la tasa de desempleo no aleceradora de los salarios. Se estima como una componente no observable en una curva Phillips. Dicho proceso se detalla a continuación.

La NAWRU se define implicitamente como el equilibrio de un sistema dinámico formado por las ecuaciones de oferta y demanda de trabajo. Precisamente la curva de Phillips resume ese proceso por el que los salarios se adaptan a las condiciones económicas. Con retardos en el ajuste como consecuencia de la existencia información limitada en la formación de expectativas o por rigideces institucionales. Ello implica que diferentes supuestos sobre la formación de expectativas implicarán diferentes especificaciones de la curva de Phillips. Por ejemplos, expectativas estáticas o adaptativas se traducen una curva tradicional Keynesiana de Phillips (TKP), mientras que expectativas racionales se traducen en una curva nueva Keynesiana de Phillips (NKP).

La TKP y la NKP responden a una misma concepción del mercado de trabajo, sólo se diferencia en la formación de las expectativas. La CE introdujo recientemente la NKP, para evitar las estimaciones de NAWRU excesivamente procíclicas que generaba TKP en determinadas cirscunstancias.

La estimación de la NAWRU se realiza de manera similar a la empleada para estimar la PTF potencial. Se formula en un modelo de componentes inobservados y se aprovecha la relación de la componente cíclica de la NAWRU con la inflación para obtener la tendencia como estimación de la componente potencial. En el caso de la PTF se aprovechaba la relación de su componente cíclica con el grado de utilización de los factores. Una diferencia en la formulación es que la componente tendencial en la PTF se expresaba como una tendencia amortiguada, mientras que en la NAWRU como un proceso integrado de orden 2.

$$\begin{aligned} u_t &= T_t + c_t \\ T_t &= T_{t-1} + d_t + \epsilon_t^T \\ d_t &= d_{t-1} + \epsilon_t^d \\ c_t &\sim AR(2) \\ \Delta clur &= \alpha \Delta clur_{t-1} + \beta_1(u_t - T_t) + \beta_2(u_{t-1} - T_{t-1}) + \Sigma_i \omega_i tot_{t-i} + \epsilon_t^{clur} \end{aligned} \tag{14}$$

donde $V(\epsilon_t^T) = V_T$, $V(\epsilon_t^d) = V_d$, $V(\epsilon_t^{clur}) = V_{clur}$, $\beta_1 < 0$ *y* $\beta_2 > 0$.

La tasa de paro observada (u) se obtiene como la suma una componente tendencial ($T \equiv NAWRU$) y una cíclica. Por último, se introduce la relación entre la componente cíclica de la tasa de paro y un indicador de inflación, el coste laboral unitario del factor trabajo (clur). Así formulado, es un curva de Phillips que adopta un enfoque mixto TKP / NKP, gobernado por el parámetro α.

Con $\alpha = 0$ tenemos un enfoque *forward looking* puro, coherente con el planteamiento NKP. Mientras que con $\alpha \simeq 1$, es una curva de Phillips Keynesiana tradicional, con expectativas estáticas o adaptativas (*backward looking*).

La variable *tot*, es una variable exógena que representa los términos de comercio o términos de intercambio. Hasta ahora, el modelo de mercado de trabajo planteado está formulado sobra las bases de una economía cerrada. Esta variable permite incluir el efecto sobre la formación de los salarios de la diferencia entre los precios domésticos y los precios de importación.

La estimación se realiza de manera semejante al caso de la PTF, expresando el modelo en Espacio de Estados y empleando el filtro de Kalman para estimar por máxima verosimilud o procedimientos bayesianos.

El IGE difunde las proyecciones de PIB real, potencial, output gap y las contribuciones de los factores productivos al crecimiento potencial de la economía de Galicia para un horizonte de medio plazo. Estas proyecciones han sido obtenidas empleando la metodología de la CE expuesta.

A la vista de la tabla 2, se observa que la previsión de que el PIB real alcance al PIB potencial se produce en el 2018. Período en el que el output gap pasa de negativo a positivo. Sin embargo, en

año	PIB real (%)	PIB potencial (%)	output gap
2016	3,1	0,6	-2,5
2017	3,1	0,8	-0,7
2018	2,7	0,9	0,9
2019	2,5	1,0	2,1
2020	2,4	1,2	3,3

Tabla 2: PIB real, potencial y output gap de Galicia

el período de predicción el PIB real está reduciendo su crecimiento, mientras que el PIB potencial lo está incrementando. Al final de dicho período el output gap será claramente positivo, indicando que la economía estará creciendo generando presiones inflacionistas. Para entonces, los ingresos tributarias estarán por encima de la situación de equilibrio y habrá que entender una parte de dichos ingresos como cíclicos.

La aportación del trabajo al crecimiento potencial es negativo en todo el período de predicción, ver tabla 3. Sin embargo, dicha contribución negativa se reduce paulatinamente. Por el contrario, la aportación del capital es positivo y creciente.

año	PTF	Capital	Trabajo
2016	0,8	0,3	-0,5
2017	0,8	0,3	-0,3
2018	0,8	0,3	-0,3
2019	0,8	0,4	-0,2
2020	0,8	0,5	-0,1

Tabla 3: Contribuciones al crecimiento potencial de Galicia

REFERENCIAS

Cuerpo, C. y Losada, R. (2015) Modelización y proyección de ingresos por el Impuesto sobre la Renta de las Personas Físicas. AIReF, documentos de trabajo 2015/02.

Duy, T. A. y Thoma, M.A. (1998) Modeling and forecasting cointegrated variables: Some practical experience. Journal of Economics and Business, 50, 291–307.

Engle, R. F. y Granger, C. W. J. (1987) Co-integration and error correction: Representation, estimation, and testing, Econometrica 55(2), 251–276.

Havik, K., Mc Morrow, K., Orlandi, F., Planas, C., Raciborski, R., Röger, W, Rossi, A., Thum-Thysen, A. y Vandermeulen, V.(2014) The Production Function Methodology for Calculating Potential Growth Rates & Output Gaps. European Commission, Economic Papers 535.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

UN MODELO DE PROGRAMACIÓN MATEMÁTICA PARA LA SELECCIÓN Y ASIGNACIÓN TEMPORAL DE RECURSOS AÉREOS DESTINADOS A LA CONTENCIÓN DE UN INCENDIO FORESTAL

Jorge Rodríguez-Veiga[1], María José Ginzo-Villamayor[2] y Balbina Casas-Méndez[3]

[1]ITMATI, Instituto Tecnológico de Matemática Industrial, Santiago de Compostela, España, jorge.rodriguez.veiga@usc.es

[2]Grupo de Investigación Modestya, Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela, España, mariajose.ginzo@usc.es

[3]Grupo de Investigación Modestya, Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela, España, balbina.casas.mendez@usc.es

RESUMEN

Determinar la planificación óptima que permita decidir el tipo y el número de recursos necesarios para extinguir un incendio forestal es una tarea que se ha abordado en la literatura utilizando modelos extraídos de la investigación operativa. En este trabajo se propone un modelo de optimización que contempla también la asignación de los recursos a diferentes periodos de tiempo. Esta asignación resulta de interés para los profesionales que combaten los incendios, que deben de cumplir con la normativa española que regula los periodos de trabajo y descanso de los pilotos, sin descuidar la restricción natural de no desatender los frentes más peligrosos del incendio. Se considera la solución exacta para contrastar el modelo y también se propone un algoritmo heurístico, diseñado específicamente para este problema, que nos permite obtener una solución de calidad rápidamente cuando nos enfrentamos a un problema de dimensiones reales. Por último creamos una interfaz gráfica que facilite la introducción de los datos y proporcione los resultados de manera práctica. La programación informática, tanto del modelo, como del heurístico y de la interfaz, la hemos realizado con el lenguaje `R`.

Palabras y frases clave: contención de incendios forestales, planificación de recursos aéreos, asignación de tiempos, programación entera, algoritmos heurísticos, interfaz gráfica.

1. INTRODUCCIÓN

El fenómeno de los incendios forestales se ha convertido en uno de los principales problemas ecológicos que sufren los bosques, debido a la alta frecuencia e intensidad que han adquirido en las últimas décadas y que conlleva graves consecuencias económicas (Butry *et al.*, 2001). Cada año se producen más de 45.000 incendios en el sur de Europa, que tienen como consecuencia la quema de cerca de 0,5 millones de hectáreas de bosques y otras tierras rurales (Camia *et al.*, 2009). Los periodos de sequía o el cambio climático no hacen más que acrecentar este fenómeno (Arno y Allison-Bunnell, 2002), y aunque el esfuerzo para prevenir incendios es importante (Martínez *et al.*, 2009), es esencial contar con herramientas que permitan una toma de decisiones eficiente cuando el fuego ya se ha producido y debe ser contenido (Mavsar *et al.*, 2013, Martell, 2015).

En Galicia (con una superficie de 29.574 km^2, el 69 % de la misma está formada por bosques), el organismo público regional responsable de combatir incendios forestales cuenta con 30 aviones en 2017, de los cuales 25 son helicópteros.

Es importante una buena planificación de los recursos de extinción que reduzca los costes y daños causados por el incendio. El análisis de los incendios desde la perspectiva de la optimización de los costes asociados se remonta a Headley (1916) y Sparhawk (1925). El modelo teórico que persigue una gestión de recursos basada en la minimización de costes, en sentido amplio, ha sido

denominado C+NVC (*Cost Plus Net Value Change*, Gorte y Gorte (1979)). Este modelo combina el objetivo de minimizar el coste provocado por el uso de recursos (terrestres y/o aéreos) con el de reducir los costes generados por la quema del terreno, la pérdida de materiales o las tareas de regeneración. Donovan y Rideout (2003a) proponen una reformulación del modelo inicial y Donovan y Rideout (2003b) proponen un modelo de programación lineal determinista que minimiza la función C+NVC y permite seleccionar los recursos a utilizar considerando el momento en que se puede contener el incendio con los recursos disponibles.

Lee (2006) propone una versión estocástica del modelo anterior, que tiene en cuenta que el perímetro del incendio en cada periodo de tiempo puede no ser un valor perfectamente determinado. Si bien este modelo puede ser más realista por acomodar la incertidumbre, su resolución puede incurrir en un mayor coste computacional. Hu y Ntaimo (2009) también incorporan estocasticidad en el modelo, vinculado en este caso a la posible aparición simultánea de varios incendios. Recientemente, Costafreda-Aumedes *et al.* (2015) utilizan un modelo de redes neuronales para analizar los costes y otros factores que influyen en el despliegue de diferentes recursos para la extinción.

Cabe señalar que el problema de la selección de recursos diseñado para la contención de incendios a coste mínimo, tiene similitudes con el problema de la mochila (Winston, 1994), un problema clásico de la investigación operativa en el que se busca seleccionar recursos dentro de un conjunto de posibles candidatos, sujeto a ciertas restricciones y con el objetivo de optimizar una cierta función de utilidad. Una peculiaridad notable es la componente temporal de nuestro problema, pues una vez elegido un recurso puede ser utilizado posteriormente durante varios periodos de tiempo. También la contención del incendio está ligada a un periodo de tiempo, de modo que cuanto más se tarde en contener el incendio, mayor será el área quemada.

El objetivo de este trabajo es adaptar la metodología C+NVC al caso de la contención de incendios en España. Hemos tenido en cuenta la Circular Operacional 16-B (1995) que establece como requisitos el tiempo máximo de vuelo diario para el piloto de un avión en un incendio forestal (8 horas) y el tiempo máximo de vuelo sin interrupciones (2 horas cada 40 minutos).

2. MARCO TEÓRICO Y ENFOQUE METODOLÓGICO

Proponemos un modelo de programación lineal binario cuya meta es seleccionar los recursos aéreos que se utilizarán, durante un día de trabajo, para la extinción de un incendio forestal. Al mismo tiempo, proporciona una planificación temporal de los recursos que se ajusta a los tiempos máximos de vuelo y los tiempos normativos de descanso. La función objetivo del modelo, siguiendo la metodología C+NVC, contempla minimizar el tiempo de extinción de incendios y, en consecuencia, minimizar los daños causados por el fuego y el tiempo de utilización de los recursos.

El modelo ha sido programado con el software libre `R` (The `R` Project for Statistical Computing, 2017) y su resolución ha sido contrastada mediante el solver Gurobi (Gurobi Optimizer, 2017) por medio de distintos ejemplos. Dado que la resolución exacta puede ser computacionalmente costosa cuando se consideran problemas que implican una cantidad elevada de recursos, se ha diseñado un algoritmo heurístico específico. La heurística también se ha programado en `R` y se ha contrastado su comportamiento analizando tanto la calidad de las soluciones que proporciona, como los tiempos de cálculo necesarios para su obtención.

La metodología heurística ya ha sido utilizada para la planificación de actividades en el contexto de la contención de incendios. Por ejemplo, Kim y Bettinger (2005) utilizan una búsqueda tabú que proporciona soluciones de alta calidad, pero no es computacionalmente eficiente. Young-Hwan *et al.* (2009) utilizan una heurística simple, que es rápida y produce soluciones de calidad. Ambos trabajos se enmarcan en la planificación óptima de actividades forestales. Pappis y Rachaniotis (2010a y 2010b) estudian el problema de planificar las jornadas de trabajo tratando este problema como un problema de horarios denominado con tiempos de procesamiento deteriorados, ya que un retraso en el inicio de las tareas de extinción puede causar un aumento en el tiempo de contención del incendio. Utilizando algunas ideas del algoritmo de Drozdowski (1996) para planificación en paralelo de tareas en un sistema multiprocesador, proponen un algoritmo heurístico para resolver el problema. Bilbao-Marón (2013) estudia la asignación de recursos a aeródromos para maximizar la utilidad de dichos recursos en relación con el riesgo de incendio y minimizar, a su vez, el coste de ejecución de la asignación. La formulación de este problema considera también, dentro de la definición de dicha función de utilidad, la distancia relativa entre el aeropuerto, el punto de

potencial riesgo de incendio y el recurso acuífero (lago, río o mar). Para su resolución eficiente se propone el uso de algoritmos de optimización basados en la denominada búsqueda de armonía.

Finalmente, en este trabajo se muestra una interfaz construida con la herramienta `R` que permite utilizar el algoritmo de una manera sencilla y generar informes de las actividades relacionadas con la planificación de los recursos.

3. FORMULACIÓN DEL PROBLEMA

Comenzamos introduciendo la notación necesaria, para posteriormente formular el problema.

Conjuntos

I**:** Conjunto de posibles aeronaves a seleccionar para combatir el incendio, $I := \{1 \dots n\}$.

T**:** Conjunto de periodos de tiempo en el que se tiene información de la evolución del incendio, $T := \{1 \dots m\}$.

Parámetros

Dentro de los parámetros distinguiremos entre aquellos relativos a las aeronaves, los relativos al incendio y los relativos a la normativa.

Parámetros relacionados con las aeronaves

A_i**:** Número de periodos de tiempo necesarios para llegar al incendio desde la situación actual de la aeronave $i \in I$.

CFP_i**:** Número de periodos de tiempo que ha trabajado sin descansos la aeronave $i \in I$ en el instante de realizar la planificación.

CRP_i**:** Número de periodos de tiempo de descanso que lleva realizados la aeronave $i \in I$ (en caso de estar en situación de descanso).

$CTFP_i$**:** Número de periodos de tiempo que ha trabajado la aeronave $i \in I$.

C_i**:** Coste de utilización de la aeronave $i \in I$ por cada periodo de tiempo que esté trabajando.

P_i**:** Coste fijo por selección de la aeronave $i \in I$.

BPR_i**:** Rendimiento de la aeronave $i \in I$ en condiciones de plena eficiencia.

Parámetros relacionados con el incendio

SP_t**:** Perímetro del incendio en el periodo de tiempo $t \in T$.

NVC_t**:** Incremento del coste del incendio en el periodo de tiempo $t \in T$ (consecuencia de las hectáreas de terreno afectadas por el fuego).

EF_t**:** Eficiencia de las aeronaves en el periodo de tiempo $t \in T$ (proporción de rendimiento). Este parámetro será dependiente de las condiciones meteorológicas.

$nMin_t$**:** Número mínimo de aeronaves trabajando en el incendio en el periodo de tiempo $t \in T$.

$nMax$**:** Número máximo de aeronaves trabajando en el incendio en un mismo periodo de tiempo.

Parámetros relacionados con la normativa de aviación

FP**:** Número máximo de periodos de tiempo volando sin descansos.

RP**:** Número de periodos de tiempo de descanso entre vuelos.

DFP**:** Número máximo de periodos de tiempo de trabajo en una jornada.

$FBRP$**:** Número de periodos de tiempo de viaje entre el punto de descanso y el incendio, y viceversa.

Además, por conveniencia, definiremos ciertos parámetros que son función de algunos de los ya introducidos.

Otros parámetros

PR_{it}**:** Rendimiento de la aeronave $i \in I$ en el periodo $t \in T$,

$$\forall i \in I, t \in T, \quad PR_{it} := BPR_i \cdot EF_t$$

PER_t**:** Incremento del perímetro del incendio en el periodo $t \in T$,

$$\forall t \in T, \quad PER_t := \begin{cases} SP_t & \text{si } t = 1 \\ SP_t - SP_{t-1} & \text{en otro caso} \end{cases}$$

l_i**:** Número de periodos de tiempo que le lleva llegar al incendio a la aeronave $i \in I$,

$$\forall i \in I, \quad l_i := A_i + \left\lfloor \frac{A_i + CFP_i}{FP} \right\rfloor \cdot (RP + 2 \cdot FBRP) - CRP_i$$

M'**:** Término de penalización por el incumplimiento del número mínimo de aeronaves.

M**:** Término suficientemente grande que contribuye a establecer, en cada instante, si el incendio está o no contenido. Basta tomar:

$$M := \max_{t \in T} SP_t + \sum_{i \in I, t \in T} PR_{it}$$

Variables

En este caso diferenciaremos entre las variables de decisión propias del problema y aquellas auxiliares que definiremos por conveniencia.

Variables del problema

s_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ es seleccionado para trabajar sobre el incendio en el periodo $t \in T$.

fl_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ vuela (sin trabajar sobre el incendio) en el periodo $t \in T$.

r_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ descansa en el periodo $t \in T$.

er_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ termina un descanso en el periodo $t \in T$.

e_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ termina su trabajo en el incendio en el periodo $t \in T$.

y_t**:** variable binaria que toma el valor 1 si el incendio no está contenido en el periodo $t \in \{0\} \cup T$. Se toma $y_0 = 1$.

μ_t**:** variable continua que contabiliza el número de recursos faltantes para llegar al mínimo en el periodo $t \in T$.

Variables auxiliares

u_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ se encuentra seleccionado en el periodo $t \in T$,

$$\forall i \in I, t \in T, \quad u_{it} := \sum_{j \in \{1,\ldots,t\}} s_{ij} - \sum_{j \in \{1,\ldots,t-1\}} e_{ij}$$

w_{it}**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ se encuentra trabajando en el periodo $t \in T$,

$$\forall i \in I, t \in T, \quad w_{it} := u_{it} - r_{it} - fl_{it}$$

z_i**:** variable binaria que toma el valor 1 cuando el recurso $i \in I$ es seleccionado para trabajar en el incendio en algun periodo de tiempo,

$$\forall i \in I, \quad z_i := \sum_{t \in T} e_{it}$$

Modelo

La formulación matemática siguiente modela la función objetivo y las limitaciones que se imponen para identificar la asignación óptima en el problema de la selección de recursos para la contención de un incendio forestal.

Función objetivo

$$\text{mín} \quad \sum_{i \in I, t \in T} C_i \cdot u_{it} + \sum_{i \in I} P_i \cdot z_i + \sum_{t \in T} NVC_t \cdot y_{t-1} + \sum_{t \in T} M' \cdot \mu_t \tag{1}$$

La función objetivo (1) minimiza la suma de los costes involucrados en la extinción del incendio forestal. El primer término indica el coste variable por utilizar los recursos seleccionados, el segundo, el coste fijo asociado a la utilización de cada uno de ellos y el tercero será el producido por las hectáreas de terreno quemado. El último término se incluye para penalizar el incumplimiento en el número mínimo de aeronaves.

Restricciones sobre la contención del incendio

$$\sum_{t \in T} PER_t \cdot y_{t-1} \leq \sum_{i \in I, t \in T} PR_{it} \cdot w_{it} \tag{2}$$

$$\forall t \in T, \quad M \cdot y_t \geq SP_t \cdot y_{t-1} - \sum_{i \in I, j \in \{1..t\}} PR_{ij} \cdot w_{ij} \tag{3}$$

La restricción (2) nos indica que en algún periodo el perímetro construido (esto es, sofocado o contenido) por los recursos tiene que cubrir todo el perímetro afectado por del incendio. (3) aporta coherencia con el hecho de si el incendio está contenido o no en cada periodo.

Restricciones sobre el comienzo y fin de la actividad

$$\forall i \in I, t \in T, \quad \sum_{j \in \{1,\ldots,t\}} s_{ij} - \sum_{k \in \{1,\ldots,t-l_i\}} s_{ik} - r_{it} \leq fl_{it} \tag{4}$$

$$\forall i \in I, \quad \begin{cases} s_{i1} + \displaystyle\sum_{t \in \{2,\ldots,m\}} (m+1) \cdot s_{it} \leq m \cdot z_i & \text{si } A_i = 0 \\ \displaystyle\sum_{t \in T} s_{it} \leq z_i & \text{en otro caso} \end{cases} \tag{5}$$

$$\forall i \in I, t \in T, \quad \sum_{\substack{j \in \{t-FBRP+1,\ldots,t\}: \\ t-FBRP+1 \geq 1 \text{ y } t-FBRP+1 \leq t}} fl_{ij} \geq FBRP \cdot e_{it} \tag{6}$$

Las restricciones (4) y (5) fuerzan a que las condiciones de intervención de las aeronaves en el incendio sean coherentes. Por un lado, (4) establece que si una aeronave es seleccionada para combatir el incendio, lógicamente deberá de volar hasta él. Por otra parte, (5) obliga a que si una aeronave ya está trabajando sobre el incendio ($A_i = 0$) en el momento de planificar una nueva selección, en caso de que se decida que esa aeronave continúe, ha de ser seleccionada desde el principio del periodo de tiempo considerado.

Por otro lado, la desigualdad (6) establece que tras la finalización del trabajo de una aeronave esta deberá de tener tiempo suficiente para volar de vuelta a una base.

Restricciones sobre los periodos de descanso

Para facilitar la formulación de este conjunto de restricciones, definimos unas nuevas variables auxiliares, cr_{it}, que constituyen un contador del número de periodos de tiempo volados sin descanso de cada aeronave en cada periodo. Además aquí se diferencian dos situaciones, pues en caso de que la aeronave no se elija en el instante inicial, y haya realizado un periodo de vuelo inferior a FP periodos, ha de ejecutar un descanso antes de ser seleccionada. De este modo, una aeronave que haya realizado un vuelo de menos de FP periodos, y no sea seleccionada en el instante inicial, ha de realizar un descanso completo de RP periodos antes de entrar a trabajar en el incendio. Sin embargo, si esta misma situación se produjese con una aeronave que ya ha realizado su periodo de descanso, no ha de esperar los RP periodos de tiempo antes de partir.

$\forall i \in I, t \in T$,

$$cr_{it} := \begin{cases} \sum\limits_{j\in\{1,\dots,t\}} (t+1-j+CFP_i-CRP_i)\cdot s_{ij} \\ -\sum\limits_{j\in\{1,\dots,t\}} (t-j)\cdot e_{ij} - \sum\limits_{j\in\{1,\dots,t\}} r_{ij} & \text{si } CFP_i = 0 \\ -\sum\limits_{j\in\{1,\dots,t\}} FP\cdot er_{ij} & \text{o } CFP_i \geq FP \\ \\ (t+CFP_i-CRP_i)\cdot s_{i1} + \sum\limits_{j\in\{2,\dots,t\}} (FP+t+1-j)\cdot s_{ij} \\ -\sum\limits_{j\in\{1,\dots,t\}} (t-j)\cdot e_{ij} - \sum\limits_{j\in\{1,\dots,t\}} r_{ij} & \text{en otro caso} \\ -\sum\limits_{j\in\{1,\dots,t\}} FP\cdot er_{ij} \end{cases}$$

Mediante el uso de estas nuevas variables, definimos las siguientes restricciones relativas a las condiciones de los periodos de descanso que ha de realizar una aeronave.

$$\forall i \in I, t \in T, \quad FP \geq cr_{it} \tag{7}$$

$$\forall i \in I, t \in T, \quad 0 \leq cr_{it} \tag{8}$$

$\forall i \in I, t \in T$,

$$\begin{cases} \sum\limits_{j\in\{t-RP+1,\dots,t\}} r_{ij} \geq RP\cdot er_{it} & \text{si } t-RP \geq 0 \\ \\ \sum\limits_{j\in\{1,\dots,t\}} r_{ij} \geq (RP-CRP_i)\cdot er_{it} & \text{en otro caso} \end{cases} \tag{9}$$

$\forall i \in I, t \in T$,

$$\begin{cases} r_{it} \leq \sum\limits_{j\in\{t,\dots,t+RP\}} er_{ij} & \text{si } t+RP \leq m \\ \\ r_{it} \leq 0 & \text{en otro caso} \end{cases} \tag{10}$$

$\forall i \in I, t \in T,$

$$\begin{cases} (-1 - FBRP - (t-1)) \cdot r_{it} + \sum\limits_{j \in \{1,\ldots,t+FBRP\}} (r_{ij} + fl_{ij}) \geq 0 & \text{si } t \leq FBRP \\ (-1 - FBRP - (m-t)) \cdot r_{it} + \sum\limits_{j \in \{t-FBRP,\ldots,m\}} (r_{ij} + fl_{ij}) \geq 0 & \text{si } t > FBRP \text{ y } t \geq m - FBRP \\ (-1 - 2 \cdot FBRP) \cdot r_{it} + \sum\limits_{j \in \{t-FBRP,\ldots,t+FBRP\}} (r_{ij} + fl_{ij}) \geq 0 & \text{en otro caso} \end{cases} \tag{11}$$

Las restricciones (7), (8), (9), (10) y (11) deben de considerarse conjuntamente. Todas ellas se establecen para forzar a las aeronaves a realizar los correspondientes descansos durante los tiempos establecidos.

Las dos primeras, (7) y (8), obligan a realizar descansos una vez que las aeronaves llegan al máximo tiempo de vuelo permitido sin descansos. La restricción (9), determina que después de los periodos de descanso obligatorios por normativa, la aeronave ya puede volver a trabajar sobre el incendio. La restricción (10) establece que cuando una aeronave realiza un descanso debe terminarlo. Por último, la restricción (11) acompasa los tiempos de vuelo entre el incendio y las bases donde se realizan los descansos con la realización de dichos descansos.

Restricciones sobre el tiempo de vuelo de las aeronaves

$$\forall i \in I, \quad \sum_{t \in T} u_{it} \leq DFP - CTFP_i \tag{12}$$

La restricción (12) establece que el tiempo de vuelo diario de una aeronave no puede ser superior al tiempo establecido por la normativa. Por tanto, también se ha de contemplar el tiempo que lleva trabajando la aeronave antes de haber sido seleccionada.

Restricciones sobre el número de aeronaves

$$\forall t \in T, \quad \sum_{i \in I} w_{it} \geq nMin_t \cdot y_{t-1} - \mu_t \tag{13}$$

$$\forall t \in T, \quad \sum_{i \in I} w_{it} \leq nMax \cdot y_{t-1} \tag{14}$$

La restricciones (13) y (14) determinan que no se podrá incumplir el número mínimo y máximo de aeronaves, respectivamente, mientras el incendio no esté contenido. Con respecto a la restricción (13), hemos de resaltar que puede ocurrir que no exista ninguna aeronave situada en el incendio en el instante inicial, y por tanto, el incendio estará desatendido durante un cierto tiempo. Este hecho inevitable ha de ser contemplado en el planteamiento del modelo. Por ese motivo, se incorpora al modelo la variable μ_t, que tomará el valor necesario para que se ajuste al número mínimo de aeronaves en el incendio en caso de que no se pueda alcanzar efectivamente.

Restricciones lógicas finales

$$\forall i \in I, \quad \sum_{t \in T} t \cdot e_{it} \geq \sum_{t \in T} t \cdot s_{it} \tag{15}$$

$$\forall i \in I, \quad \sum_{t \in T} w_{it} \geq z_i \tag{16}$$

$$\forall i \in I, t \in T, \quad r_{it} + fl_{it} \leq u_{it} \tag{17}$$

Por último, el siguiente conjunto de restricciones establece las últimas condiciones lógicas que deben de satisfacer las variables. De este modo, (15) establece que el periodo en el que es seleccionado un recurso para comenzar a combatir el incendio ha de ser anterior al periodo de finalización

de su trabajo; (16) especifica que toda aeronave seleccionada para trabajar ha de contribuir efectivamente a la contención del incendio, y (17) establece que los periodos de descanso y de vuelo también son considerados como periodos de uso de las aeronaves.

4. UN ALGORITMO HEURÍSTICO

Con el fin de reducir los tiempos de obtención de la solución exacta del modelo, se diseñó un algoritmo heurístico que, además de otorgar mayor control en la toma de decisiones, nos permite controlar el tiempo de obtención de decisiones factibles.

La principal idea del algoritmo reside en descomponer el problema de decisión de partida en dos subproblemas, también de variables enteras: un denominado problema 1, cuya finalidad es seleccionar los recursos que deben trabajar en el incendio para conseguir la extinción del mismo, y un problema 2, que permite planificar el momento en el que las aeronaves deben de entrar a combatir el incendio, de forma que se maximice su rendimiento y a la vez se cumpla la normativa de aviación relativa a periodos de descanso y tiempos máximos de vuelo.

Problema 1: selección de recursos	
Función objetivo:	Minimizar costes
Restricciones:	Contención del incendio
	Fin de la actividad
	Tiempo de vuelo de las aeronaves
	Número de aeronaves

Tabla 1: Modelo para la selección de los recursos aéreos y contención del incendio.

Problema 2: planificación horaria de recursos	
Función objetivo:	Maximizar rendimiento
Restricciones:	Tiempo de vuelo de las aeronaves
	Comienzo de la actividad
	Periodos de descanso
	Número de aeronaves
	No repetición de soluciones

Tabla 2: Modelo para la planificación horaria de los recursos aéreos.

Las Tablas 1 y 2 muestran los elementos más importantes que aparecen en los dos problemas que fundamentan nuestro algoritmo. Es importante mencionar la retroalimentación existente en ambos problemas, tal y como se muestra en la Figura 1 que representa el diagrama de flujo del algoritmo. El problema 1 necesita conocer el momento en el que las posibles aeronaves entrarán en el incendio (denotado por `S1[i,t]` en el diagrama) en caso de ser seleccionadas, y conocer también su futuro programa horario de trabajo (`Work[i,t]`) hasta que se le indique abandonar el

incendio. Esto es necesario para poder seleccionar un conjunto de aeronaves (`I_select`) capaces de contener el incendio en un periodo de tiempo (`m_select`), de tal forma que se minimicen los costes de la operación de contención. A su vez, el problema 2 hará uso de la selección de las aeronaves (`I_select`) y el periodo estimado de contención del incendio (`m_select`) para realizar la planificación horaria de los recursos aéreos (`S1[i,t]` y `Work[i,t]`), maximizando el rendimiento de las aeronaves seleccionadas en los periodos indicados.

Otro elemento importante en la ejecución del algoritmo es la factibilidad de las soluciones (`fact`), que se establece en función del cumplimiento o no de la restricción relativa al número máximo de aeronaves permitido en el incendio. De este modo, la factibilidad necesita controlar el número de aeronaves que excede la cota máxima de aeronaves en el incendio en cada periodo (considerando factible una solución cuando `fact=0`).

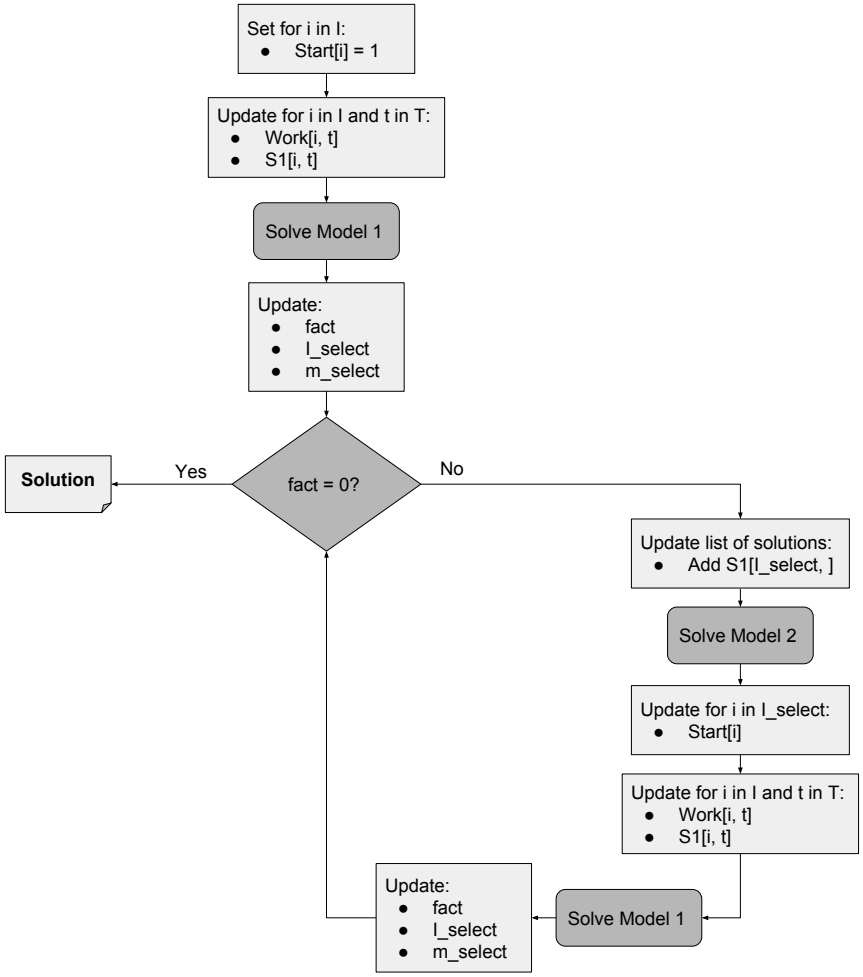


Figura 1: Diagrama de flujo del heurístico.

5. INTERFAZ

Con el fin de facilitar la ejecución de las dos propuestas (solución exacta del modelo y del heurístico) se ha creado una interfaz con el software `R` mediante el uso de la librería `shinydashboard`.

La interfaz permite una sencilla interacción a la hora de cargar la información de las aeronaves así como de la evolución del incendio y de la normativa española de aviación. Además ofrece la opción de ejecutar los modelos con distintos solvers (*gurobi*, *lpSolve* y *Rsymphony*, estos dos últimos de licencia libre). Una vez realizada la ejecución, se muestra una visualización de los resultados

tanto gráfica (mediante gráficos interactivos) como en forma de tablas y se permite la creación de informes para guardar la información de la instancia ejecutada en un archivo en formato html.

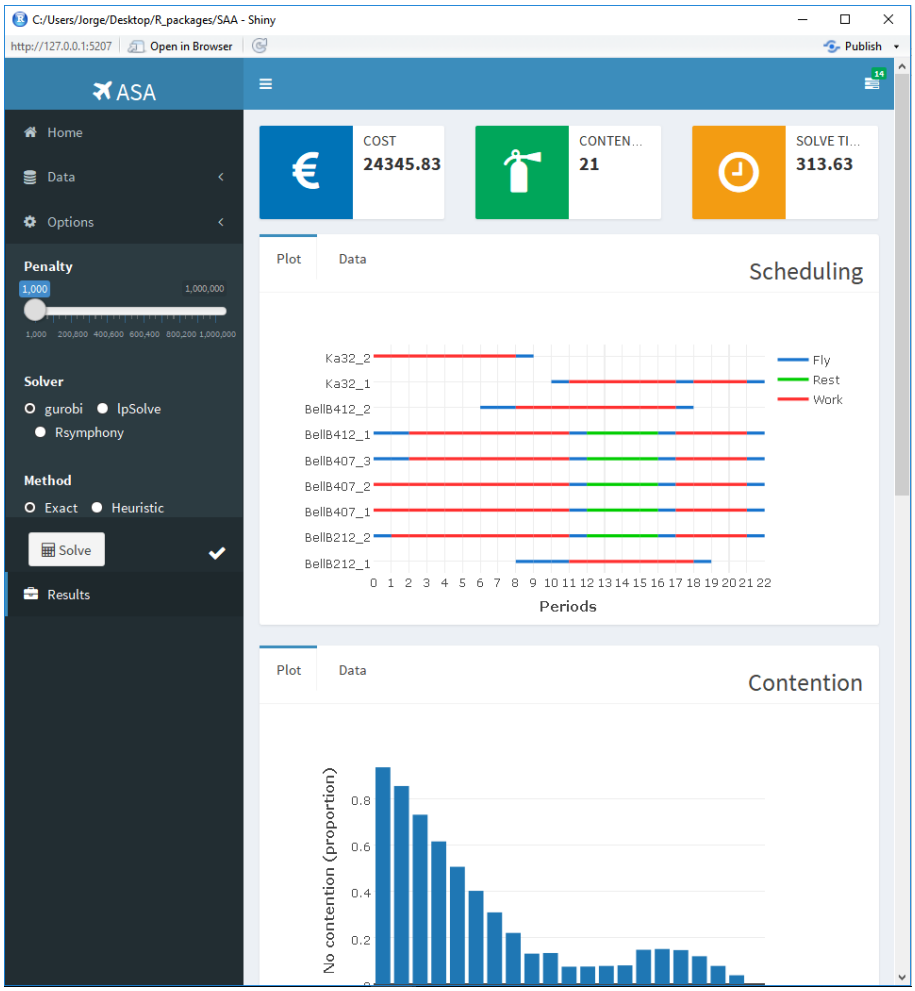


Figura 2: Datos de las aeronaves.

En la Figura 2 se muestra el resultado de resolver una instancia en la que se dispone de 9 aeronaves y se conoce la evolución del incendio durante 25 periodos de tiempo, siendo cada uno de ellos de 10 minutos, con lo que tenemos la evolución del incendio en los 250 minutos posteriores. Se tiene en cuenta además, la normativa regulada por la Circular Operacional 16-B (1995). Como vemos, se obtiene que el incendio se contendrá en el periodo 21 (a los 210 minutos) con un coste de operación de 24.345,83 €. La ventana *scheduling* nos muestra para cada aeronave seleccionada su planificación horaria que incluye los tiempos de trabajo, de vuelo y de descanso. Observamos, en particular, que en ningún periodo de tiempo el incendio queda desatendido. El tiempo que duró la ejecución fue de 314 segundos.

AGRADECIMIENTOS

Los autores de esta investigación agradecen las propuestas y comentarios realizados por A. Lorenzo y J. L. Sáiz-Díaz, de *Babcock International Group*, por M. Febrero-Bande, W. González-Manteiga y B. Pateiro-López, de la USC, y el apoyo financiero recibido desde el Ministerio de Economía y Competitividad de España a través de los proyectos MTM2014-53395-C3-2-P y MTM2016-76969-P, y desde ITMATI, a través del proyecto Enjambre.

REFERENCIAS

[1] Arno, S. F. and Allison-Bunnell, S. (2002) Flames in our forest: disaster or renewal? Island Press, Washington, DC.
[2] Butry, D. T., Mercer, D. E., Prestemon, J. P., Pye, J. M., and Holmes, T. P. (2001) What is the price of catastrophic wildfire? Journal of Forestry 99, 9-17.
[3] Bilbao-Marón, M. N. (2013) Advanced Meta-Heuristic Approaches and their Application to Operational Optimization in Forest Wildfire Management. Doctoral Thesis, Universidad de Alcalá, Spain.
[4] Camia, A., San-Miguel-Ayanz, J., Oehler, F., Santos De Oliveira, S., Durrant Houston, T., Kucera, J., Boca, R., Whitmore, C., Giovando, C., Amatulli, G., Libertà, G., Schmuck, G., Schulte, E., and Bucki, M. (2009) Forest Fires in Europe 2008. EUR 23971 EN. Luxembourg (Luxembourg): OPOCE; 2009. JRC53463.
[5] Circular Operacional 16-B (1995) Ministerio Español de Fomento. Available online at: http://www.aecaweb.com/informes/documentos/ INFORMES_Y_ESTUDIOS/anexo1aco16b.pdf. Last accessed March 6, 2017.
[6] Costafreda-Aumedes, S., Cardil, A., Molina, D. M., Daniel, S. N., Mavsar, R., and Vega-García, C. (2015) Analysis of factors influencing deployment of fire suppression resources in Spain using artificial neural networks. iForest-Biogeosciences and Forestry 9, 138-145.
[7] Donovan, G. and Rideout, D. (2003a) A reformulation of the Cost Plus Net Value Change (C + NVC) model of wildfire. Forest Science 49, 318-323.
[8] Donovan, G. and Rideout, D. (2003b) An integer programming model to optimize resource allocation for wildfire containment. Forest Science 49, 331-335.
[9] Drozdowski, M. (1996) Real-time scheduling of linear speedup parallel machines. Information Processing Letters 57, 35-40.
[10] Gorte, J. K. and Gorte, R. W. (1979) Application of Economic Techniques to Fire Management-A status Review and Evaluation. USDA Forest Service General Technical Report INT-53.
[11] Gurobi Optimizer (2017). Available online at: http://www.gurobi.com/. Last accessed March 6, 2017.
[12] Headley, R. (1916) Fire Suppression, District 5. United States Departament of Agriculture, Forest Service, Washington, 57 p.
[13] Hu, X. and Ntaimo, L. (2009) Integrated simulation and optimization for wildfire containment. ACM Transactions on Modeling and Computer Simulation 19, 1-29.
[14] Kali, A. (2016) Stochastic scheduling of single forest fire-fighting processor. Canadian Journal of Forest Research 46, 370-375.
[15] Kim, Y.-H. and Bettinger, P. (2005) Spatial optimization of fuel management activities. In: Bevers, M., Barrett, T.M. (Eds.), Systems Analysis in Forest Resources: Proceedings of the 2003 Symposium. General Technical Report PNW-656, USDA Forest Service, Portland, OR, pp. 205-214.
[16] Lee, W. J. (2006) A Stochastic Mixed Integer Programming Approach to Wildfire Management Systems. Doctoral thesis. Department of Industrial and Systems Engineering. Texas A&M University.
[17] Martell, D. L. (2015) A review of recent forest and wildland fire management decision support systems research. Current Forestry Reports 1, 128-137.
[18] Martínez, J., Vega-García, G., and Chuvieco, E. (2009) Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management 90, 1241-1252.
[19] Mavsar, R., González-Cabán, A., and Varela, E. (2013) The state of development of fire management decision support systems in America and Europe. Forest Policy and Economics 29, 45-55.
[20] Pappis, C. P. and Rachaniotis, N. P. (2010a) Scheduling in a multi-processor environment with deteriorating job processing times and decreasing values: the case of forest fires. Journal of Heuristics 16, 617-632.
[21] Pappis, C. P. and Rachaniotis, N. P. (2010b) Scheduling a single fire fighting resource with deteriorating fire suppression times and set-up times. Operations Research: An International Journal 10, 27-42.
[22] Rachaniotis, N. P. and Pappis, C. P. (2011) Minimizing the total weighted tardiness in wildfire suppression. Operations Research: An International Journal 11, 113-120.

[23] Sparhawk, W. N. (1925) The use of liability ratings in planning forest fire protection. Journal of Agricultural Research 30, 693-792.
[24] The R Project for Statistical Computing (2017). Available online at:
https://www.r-project.org/. Last accessed March 6, 2017.
[25] Winston, W. L. (1994) Operations Research Applications and Algorithms. Duxbury Press, Belmont, California.
[26] Young-Hwan, K., Bettinger, P., and Finney, M. (2009) Spatial optimization of the pattern of fuel management activities and subsequent effects on simulated wildfires. European Journal of Operational Research 197, 253-265.

XII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 y 28 de octubre de 2017

LOS VALORES PERDIDOS EN EL MUESTREO DE POBLACIONES FINITAS. TÉCNICAS DE IMPUTACIÓN

Gabriel Román Radío [1]

[1] Estudiante socio

RESUMEN

Los valores perdidos suponen un obstáculo - actualmente no bien resuelto - cuando es necesaria la respuesta completa - porque lo requiere la estimación a realizar o porque lo exigen las técnicas estadísticas a utilizar en el análisis. Las técnicas de imputación permiten asignar valores a los casos y variables con valores omitidos. No obstante, las técnicas de imputación no son neutrales en relación con los resultados obtenidos y la elección inadecuada de la misma podría ocasionar alteraciones en la distribución de los datos. Este trabajo expone la utilización de algunas estrategias vigentes para el análisis estadístico de matrices de datos incompletos y su aplicación en un supuesto.

Palabras y frases clave: valores perdidos, respuesta completa, estimación a realizar, técnicas de imputación y alteración en la distribución de los datos.

1. INTRODUCCIÓN

En el desarrollo teórico de la mayoría de técnicas y modelos estadísticos no es habitual considerar algunas cuestiones que se manifiestan en la aplicación práctica. Un dilema, al que con seguridad se ha enfrentado cualquier analista, es el de los datos faltantes, también denominados perdidos o incompletos. Este problema conduce a que, en la mayoría de los estudios muestrales y/o censales, el analista se encuentre, muy probablemente, con espacios en blanco en el área de respuesta, designados "no respuesta". La evaluación de ausencia de la no respuesta en un registro varía según varios factores como: el propio estudio, el grado de dificultad para rellenar el cuestionario o la medición de una unidad. Dependerá del analista valorar el registro como pérdida parcial o total si el número de no respuestas es considerable. Cuando no es posible ignorar la no respuesta, la técnica más empleada es completar los espacios en blanco o faltantes con valores plausibles: la imputación.

Son muchas las técnicas de imputación que han surgido, sobre todo desde la década de los sesenta del siglo pasado, que se sirven de enfoques univariantes y multivariantes a través de métodos de verosimilitud, regresión y descomposición de matrices en valores singulares. A pesar de este progreso, actualmente no se ha hallado una metodología capaz de reproducir los datos de modo totalmente satisfactorio, debido generalmente, a la compleja problemática que presentan las alteraciones de la distribución de los datos. El objetivo de este trabajo es exponer la utilización de algunas estrategias vigentes para el análisis estadístico de matrices de datos incompletos y constrastar sus resultados en un supuesto.

2. IMPUTACIÓN

La imputación consiste en obtener datos completos y consistentes para elaborar un análisis y aplicar inferencia estadística. Martos (1995) define la imputación como la etapa final del proceso de depuración de datos que consiste en reemplazar los valores faltantes por otros aceptables. En décadas anteriores, Cruz (1990), era habitual ignorar valores faltantes en alguna variable por los métodos de eliminación por lista (listwise deletion) o por pares (pairwise deletion). Las no respuestas se ignoraban en el análisis, provocando pérdida de información. Actualmente se reconoce la importancia de utilizar métodos de imputación que conserven características propias de la variable a ser imputada, lo que permite disponer de datos completos para el análisis, y aumentar la precisión de las estimaciones. Los valores imputados se clasifican en tres categorías:

1. Valores construídos con modelos estadísticos de predicción.
2. Valores observados como elementos de respuesta que son similares a aquellos que no responden.
3. Valores construídos mediante la opinión de expertos.

3. MODELOS DE GENERACIÓN DE DATOS FALTANTES

El patrón de la pérdida de los datos faltantes es un punto a considerar en la imputación; de ello depende el método seleccionado de imputación. La mayoría de los estadísticos de encuestas advierten de las diferencias que existen entre la falta de respuesta total y la no respuesta parcial. En Medina et al (2007), se establece que la no respuesta total se corrige eliminando las observaciones y ajustando los factores de expansión para que las unidades de la muestra se estimen sin sesgo. Para corregir la omisión parcial, lo común es aplicar el procedimiento Hot Deck o el método de promedios. Los distintos procedimientos de imputación establecen supuestos acerca del patrón de comportamiento de los datos omitidos y Rubin (1977) estableció la siguiente clasificación:

1. MCAR (Missing Completely At Random): la probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos. La ausencia de información no procede de ninguna variable presente en la matriz de datos, ya que la base de datos se interpreta como una matriz en donde los renglones son las unidades de observación y las columnas representan a las variables de interés (Rubin, 1977). Hay que señalar, que en la práctica, la hipótesis de este supuesto no siempre es satisfactoria porque la falta de respuesta se asocia a características personales.
2. MAR (Missing At Random): la probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable pero es dependiente de los valores de otras variables del conjunto de datos. En suma, la ausencia de datos está asociada a variables presentes en la matriz de datos. Por ejemplo, este supuesto, es usual en las encuestas de hogares que estudian el nivel de vida, donde la falta de respuesta en la variable ingreso en las familias con mayor renta está correlacionada con sus ingresos y por lo tanto no se cumple el supuesto. En cambio, cuando la omisión ocurre en la categoría ocupacional, la falta de información no necesariamente está correlacionada con el nivel de ingresos de las personas y por lo tanto sí se cumple el supuesto.
3. NMAR (Not Missing At Random): la probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable. Por ejemplo, en el estudio de las variables peso y edad; los sujetos con mayores valores de peso tienen un porcentaje de datos faltantes más elevado en esta variable para aquellos individuos con la misma edad. Este tipo de dato faltante también se denomina no ignorable.

4. MÉTODOS DE IMPUTACIÓN

Los métodos de imputación son una posible solución a la falta de respuesta parcial. La razón principal por la cual los investigadores realizan la imputación es para obtener un conjunto de datos completo y consistente al cual se le aplica inferencia estadística. En la fase de imputación se debe escoger cuidadosamente las variables objetivo, auxiliares, los criterios de imputación y el método preciso de imputación. Los distintos métodos de imputación de datos se diferencian en dos tipos generales: los que utilizan los valores existentes en algún registro de la variable a imputar o relacionados con ellos y los que utilizan un conjunto de campos - variables de control - de la encuesta; en todos los casos hay un registro donante. En este sentido, varios estudios como Goicoechea (2002) y Plateck (1986), clasifican las técnicas de imputación de la siguiente manera:

1. Técnicas fundamentadas en información externa: la información está basada en variables relacionadas con una encuesta perteneciente a otras bases de datos o reglas previas. Entre este tipo de técnicas podemos encontrar:
    - Métodos deductivos: los datos faltantes se deducen con cierto grado de certidumbre de otros registros completos del mismo caso, siguiendo algunas reglas específicas.

- Tablas Look - up: se hace uso de una tabla con información relacionada, como fuente de datos externa para imputar los datos faltantes.

2. Técnicas determinísticas: al repetir la imputación en varias unidades bajo las mismas condiciones se producirán las mismas respuestas. Entre este tipo de técnicas podemos encontrar:

   - Imputación de la media: se llena el vacío del dato faltante de cada variable con la media de los registros no faltantes en caso de variables cuantitativas, con la moda en caso de variables cualitativas. La desventaja es que puede ocasionar la modificación de las distribuciones de la variable reduciendo su varianza y no conservar la relación entre variables. La ventaja es la facilidad de aplicación.
   - Imputación de media de clases: las respuestas de cada variable son agrupadas en clases disjuntas con diferentes medias, y a cada registro faltante se le imputará con la media respectiva de su grupo. Las desventajas son las mismas que en el caso anterior, pero en menor proporción por estar agrupadas. La ventaja sigue siendo la facilidad de aplicación.
   - Imputación por regresión: se ajusta un modelo lineal que describa a Y, variable a imputar, para un conjunto X de variables auxiliares que se deben disponer. Resuelve el problema de la distorsión de la distribución de la variable a imputar, pero puede crear inconsistencias dentro de la base de datos, pues podría obtener valores imposibles.
   - Emparejamiento media: el valor de Y estimado es comparado con casos completos, y el caso más cercano correspondiente provee el valor imputado de Y.
   - Imputación por el vecino más cercano: se identifica la distancia entre la variable a imputar Y, y cada una de las unidades restantes o variables auxiliares mediante alguna medida de distancia, entonces se determina la unidad más cercana a Y, usando el valor de esta unidad cercana para imputar el dato faltante.
   - Algoritmo EM (Expectation Maximization): basado en la función de máxima verosimilitud, permite obtener estimaciones máximo verosímiles (MV) de los parámetros cuando hay datos incompletos con unas estructuras determinadas. Resuelve de forma iterativa el cálculo del estimador máximo verosímil mediante dos pasos en cada iteración (Little et al 1987). Este algoritmo resuelve un amplio rango de problemas, incluso los no usuales que surgen de la pérdida o información faltante, como la estimación de los componentes de la varianza.
   - Redes Neuronales: son sistemas de información procesados, que reconocen patrones de los datos sin algún valor perdido para aplicarlo a la información a imputar. Estas redes son más usadas para variables cualitativas que cuantitativas, siendo más adecuadas cuando la distribución no es lineal. No es aconsejable en caso de registros atípicos que distorsionan la red. Son costosos y requieren de capacitación del analista así como de software adecuado.
   - Modelos de series de tiempo: el problema se reduce a una situación en la cual hay una serie de tiempo, donde algunas observaciones están perdidas; se hace un uso óptimo de las interrelaciones entre observaciones en cada serie de tiempo, mediante el uso de un modelo adecuado.

3. Técnicas aleatorias o estocásticas: se repite el método de imputación bajo las mismas condiciones para una unidad y producen resultados diferentes. Dentro de este tipo de técnicas destacan:

   - Imputación aleatoria de un caso seleccionado: para cada caso con una celda faltante, se selecciona un donante aleatoriamente para ser asignado al dato faltante.
   - Imputación de un caso seleccionado entre clases: se realiza de igual forma que para el caso anterior, pero se lleva a cabo dentro de clases previamente creadas.
   - Imputación secuencial Hot Deck: cada caso es procesado secuencialmente. Si el primer registro tiene un dato faltante, éste es reemplazado por un valor inicial para imputar, pudiendo ser obtenido de información externa. Si el valor no está perdido, éste será el valor inicial y es usado para imputar el subsiguiente dato faltante. Una de las desventajas es que si el primer registro está perdido, se necesita un valor inicial, generalmente obtenido de manera aleatoria. Además si se imputan muchos registros se tiende a emplear el mismo registro donante, que conlleva la pérdida de precisión en las estimaciones.

- Imputación jerárquica Hot Deck: similiar al método secuencial anterior, se organiza dentro de clases que usan variables auxiliares en forma de una estructura jerárquica. Si el donante no es encontrado en un nivel de clasificación, las clases pueden ser colapsadas en grupos más amplios hasta que el donante sea encontrado.
- Imputación por regresión aleatoria: se realiza primero un procedimiento de regresión; luego un término residual es adicionado para imputar los valores de Y. Este término de error puede ser obtenido de diferentes maneras, una de ellas es a través de los residuos del modelo de regresión, generado con registros completos, eligiendo uno de éstos residuos aleatoriamente.
- Imputación por regresión logística: similar a la técnica anterior, pero para imputar variables binarias.

5. SUPUESTO

Para ejemplificar la utilización de la imputación, utilizaremos el conjunto de datos tsAirgap del paquete imputeTS de R. Dicho conjunto de datos se correponde con la serie cronológica de pasajeros mensuales de las líneas aéreas (con NAs), con el número total mensual de los pasajeros de las líneas aéreas internacionales, desde 1949 a 1960. En el paquete imputeTS también se incluye la serie temporal tsAirgapComplete que proporciona los valores verdaderos para los valores faltantes. El conjunto de datos tsAirgap tiene un formato que incluye 144 filas con 13 NAs. Se origina a través de un modelo Box & Jenkins y es un ejemplo comúnmente usado en la literatura de las series de tiempo, dado que sus características son una fuerte tendencia y una fuerte componente estacional, lo que la convierte en un gran ejemplo para la imputación en series de tiempo.

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1949	112	118	132	129	NA	135	148	148	NA	119	104	118
1950	115	126	141	135	125	149	170	170	NA	133	NA	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	NA	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	NA	NA	NA	374	413	405	355	306	271	306
1957	315	301	356	348	355	NA	465	467	404	347	NA	336
1958	340	318	NA	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	NA
1960	417	391	419	461	NA	535	622	606	508	461	390	432

Tabla 1: Representación del conjunto de datos tsAirgap donde los valores coloreados en rojo representan los valores faltantes.

Eliminamos el componente estacional de la serie tsAirgap, y realizamos la imputación utilizando interpolación en la serie desestacionalizada para después agregar el componente estacional otra vez obteniendo los resultados de la tabla 2.

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1949	112.0000	118.0000	132.0000	129.0000	121.3941	135.0000	148.0000	148.0000	131.6204	119.0000	104.0000	118.0000
1950	115.0000	126.0000	141.0000	135.0000	125.0000	149.0000	170.0000	170.0000	149.7682	133.0000	113.0095	140.0000
1951	145.0000	150.0000	178.0000	163.0000	172.0000	178.0000	199.0000	199.0000	184.0000	162.0000	146.0000	166.0000
1952	171.0000	180.0000	193.0000	181.0000	183.0000	218.0000	230.0000	242.0000	209.0000	191.0000	172.0000	194.0000
1953	196.0000	196.0000	236.0000	235.0000	229.0000	243.0000	264.0000	272.0000	237.0000	211.0000	180.0000	201.0000
1954	204.0000	188.0000	235.0000	227.0000	234.0000	274.9995	302.0000	293.0000	259.0000	229.0000	203.0000	229.0000
1955	242.0000	233.0000	267.0000	269.0000	270.0000	315.0000	364.0000	347.0000	312.0000	274.0000	237.0000	278.0000
1956	284.0000	277.0000	321.8776	321.4178	329.4329	374.0000	413.0000	405.0000	355.0000	306.0000	271.0000	306.0000
1957	315.0000	301.0000	356.0000	348.0000	355.0000	424.1798	465.0000	467.0000	404.0000	347.0000	305.0315	336.0000
1958	340.0000	318.0000	357.2362	348.0000	363.0000	435.0000	491.0000	505.0000	404.0000	359.0000	310.0000	337.0000
1959	360.0000	342.0000	406.0000	396.0000	420.0000	472.0000	548.0000	559.0000	463.0000	407.0000	362.0000	398.3917
1960	417.0000	391.0000	419.0000	461.0000	477.2415	535.0000	622.0000	606.0000	508.0000	461.0000	390.0000	432.0000

Tabla 2: Representación del conjunto de datos tsAirgap realizando imputación por interpolación donde los valores coloreados en azul representan los valores imputados.

Si comparamos los resultados de la imputación con los valores reales de la serie tsAirgapComplete (tabla 3), observamos que los valores imputados se acercan mucho a los valores reales.

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

Tabla 3: Representación del conjunto de datos tsAirgapComplete donde los valores coloreados en verde representan los valores reales.

6. CONCLUSIONES

En conclusión, los métodos de imputación nos permiten utilizar la inferencia estadística para obtener estimadores insesgados (o con pequeño sesgo) y de varianza mínima a través del ajuste de modelos que sustituyen la información faltante. Pero la imputación tiene ventajas y desventajas. Las ventajas de la imputación son que sus procedimientos permiten obtener un conjunto de datos completo posibilitando la redución del sesgo debido a la no respuesta. En este sentido, la imputación opera sobre los datos de forma que los resultados que se obtienen de los diferentes métodos suelen llevar a un análisis consistente. Las desventajas de la imputación son que la futura estimación obtenida del análisis no distinga entre las imputaciones y los datos reales. Los valores imputados pueden conducir a buenas estimaciones pero no son datos reales que aseguren una mejora en el sesgo respecto del sistema de datos incompletos. Si el método de imputación no es el adecuado, posiblemente aumente el sesgo y se sobreestime la varianza, al generar datos imputados inconsistentes que llevarían a la interpretación errónea de los resultados.

REFERENCIAS

Cruz, P. (1990) Del no sabe al no contesta: un lugar de encuentro para diversas respuestas. Reis, 52, 139–156.

Goicoechea, P. (2002) Imputación basada en árboles de clasificación. EUSTAT. Victoria.

Little, R. and Rubin, D. (1987) Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics. John Wiley & Sons. New York.

Martos, J. (1995) Tratamiento estadístico de la no respuesta en poblaciones finitas. Tesis, Universidad de Sevilla.

Medina, F. y Galván, M. (2007) Imputación de datos: teoría y práctica. CEPAL - Serie Estudios estadísticos y prospectivos nº 54.

Plateck, R. (1986) Metodología y tratamiento de la no-respuesta. EUSTAT. Vitoria.

Rubin, D. (1977) Assigment to Treatment Group on the Bases of a covariate. Journal of Educational Statistics, 2, 1–26.

Steffen, M. (2016) imputeTS: Time Series Missing Value Imputation. R package versión 1.8. http://www.cran.r-project.org/package=imputeTS.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

Analysis of incidence of chronic granulocytic leukemia and primary biliary cirrhosis using level set estimation

Paula Saavedra-Nieves[1] and Alberto Rodríguez-Casal[2]

[1]Universidade de Vigo
[2]Universidade de Santiago de Compostela

ABSTRACT

The question of whether the geographical incidence of disease shows any tendency towards clustering in geographical space due to environmental contamination factors can be analyzed in depth from the massive volume of available epidemiologic data currently. Figure 1 (first row) contains a data set that consists of the residential coordinates for 233 cases diagnosed of chronic granulocytic leukemia in the North West of England between 1982 up to 1998 together with 988 controls. Figure 1 (second row) also shows the geographical coordinates for 761 cases of primary biliary cirrhosis in the Northeast of England diagnosed between 1987 and 1994 and 3044 controls. See Henderson et al. (2002) and Diggle (2013) for a complete description of these two data sets, respectively.

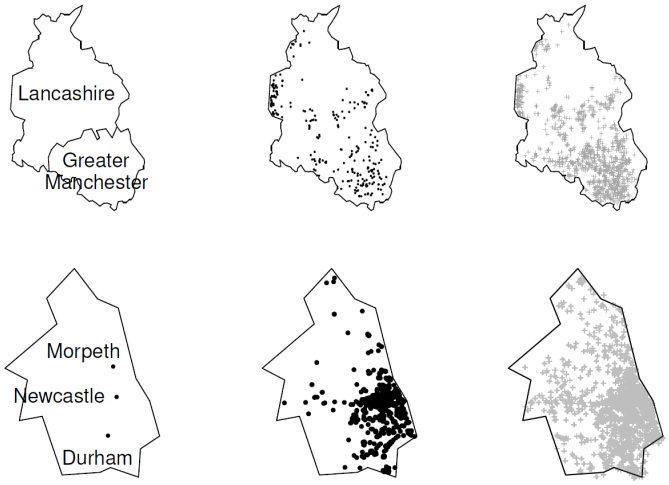


Figure 1: In the first row, sub-regions of Lancashire and Greater Manchester in the North West of England (left), distribution of 233 cases of diagnosed leukaemia (center) and 988 controls (right). In the second row, region of the Northeast of England (left), distribution of 761 cases of primary biliary cirrhosis between 1987 and 1994 (center) and 3044 controls (right).

Concretely, epidemiologists are interested in examining whether the distribution of some diseases like leukaemia or cirrhosis mirrors that of the population or whether there is evidence of clustering. The nonparametric estimation of density level sets can be a useful tool in this context. This theory estimates an unknown set $G(t) = \{x : f(x) \geq t\}$ from a random sample of points $X_1, \cdots, X_n$ of a random vector X, where f is the density of X and t, a positive level. For values of t close to zero, the level set $G(t)$ represents the effective support of the density f. However, if higher values of t are considered, it could represents the domain concentrated around the greatest mode. Nonparametric estimators of level sets are flexible since no assumptions are made on the data distribution. In exchange, smoothing parameters that are not so influential in set reconstructions must be estimated. See, for instance, Samworth and Wand (2010) or Walther (1997).

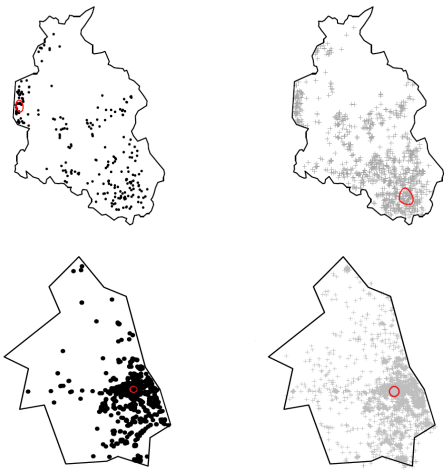


Figure 2: In the first row, nonparametric level set estimators for samples of cases (left) and controls (right) of diagnosed leukaemia. In the second row, nonparametric level set estimators for samples of cases (left) and controls (right) of diagnosed cirrhosis.

In this work, the main nonparametric methods for estimating density level sets will be reviewed. Then, level sets will be reconstructed from the samples of cases and controls presented in Figure 1. Figure 2 contains some estimators for leukaemia (first row) and cirrhosis (second row) obtained taking a large enough value for the level t. Clearly, cases and controls of leukaemia are not equally distributed. It is observed an excess of case intensity over that of population. Greater Manchester is one of the largest metropolitan region in North West England. However, Lancashire is a commercial and industrial county with high levels of environmental contamination that could be associated to the high incidence of this cancer. However, the greatest modes of cases and controls for cirrhosis appears in the same area of the map.

Key words: Density level set estimation. Smoothing parameters. Leukemia. Cirrhosis.

Acknowledgements. Authors acknowledge the financial support of the Spanish Ministry of Economy and Competitiveness through project MTM2016-76969-P (AEI/FEDER, UE) and the Government of Galicia through the ERDF (Grupos de Referencia Competitiva) ED431C 2016-040.

REFERENCES

Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. CRC Press.

Henderson, R., Shimakura, S. and Gorst, D. (2002). Modeling spatial variation in leukemia survival data. Journal of the American Statistical Association, 97, 965–972.

Samworth, R. J. and Wand, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. The Annals of Statistics, 38, 1767–1792.

Walther, G. (1997). Granulometric smoothing. The Annals of Statistics, 2273–2299.

XII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 y 28 de octubre de 2017

OPTIMAL DESIGNS IN ENZYME REACTIONS WITH INHIBITION BY HIGH SUBSTRATE CONCENTRATIONS

Santos-Martín, M.Teresa[1] and Rivas-López, M. Jesús[2]

[1],[2]Department of Statistics, University of Salamanca

ABSTRACT

In this paper, the methodology of the optimal designs of the experiments is used for estimating of a more precise form the different parameters in the most usual equations of modeling of the enzyme reactions when an inhibition by high substrate concentrations is presented.

Keywords and phrases: D-optimality, c-optimality, enzyme reaction, inhibition.

1. INTRODUCTION

The chief factors which determine the velocity of a particular reaction are enzyme concentration, substrate concentration, pH, temperature and the presence of activators or inhibitors.

Substrate concentration is one of the most important factors which determine the velocity of enzyme reactions. The enzyme first forms a complex with its substrate, and this subsequently breaks down giving the free enzyme and the products of the reaction. When the whole enzyme is bound in the complex enzyme-substrate, the reaction velocity reaches its maximum, saturation, and remains constant even increasing substrate concentration. The well–known Michaelis-Menten (MM) model

$$E[v] = \frac{Vs}{K_m + s} \quad s \in [0, b] \tag{1}$$

describes the saturation of enzyme systems. It corresponds to a rectangular hyperbola which asymptotes are $v = V$ and $s = -K_m$. Here v is the reaction velocity, V is the maximum velocity of the reaction, s is the concentration of substrate and K_m is the Michaelis constant or half–saturation constant, the value of s where $v = K_m/2$. The value of s_0 is assumed to be user–selected. There are several methods to estimate the parameters V and K_m, based on transformations of model (1) to a linear plot and application of ordinary linear regression techniques (see, e.g. Dixon and Webb (1979)). Raaijmaakers (1987) bet on application of maximum likelihood techniques for estimating the parameters in the MM model. The problem of designing experiments for the MM model has been studied by many authors (see, e.g. Duggleby (1979), Dette and Wong (1999), López–Fidalgo and Wong (2002) or Dette and Biedermann (2003)).

But, it is not uncommon to find that, while the MM equation is obeyed el lower substrate concentrations, the velocity falls off again at high concentrations. An usual cause of this phenomenon is the following. Many enzymes have several groups, each combining with a particular part of the substrate molecule. In the effective enzyme–substrate complex one substrate molecule is combined with all these groups, but it is possible to imagine an ineffective complex in which some of the binding groups are combined with one substrate molecule and the other groups with an additional substrate molecule. In high substrate concentrations, where the substrate molecules tend to crowd on the enzyme, the chances of formation of ineffective complexes, with two or more substrate molecules with an active site, increase. In this case the equation describing the reaction velocity is given by the inverse quadratic polynomial.

$$E[v] = \frac{Vs}{K_m + s + s^2/K_i} \tag{2}$$

where v is the reaction velocity, V is the maximum velocity of the reaction, s is the concentration of substrate and K_m is the Michaelis constant or half–saturation constant and K_i is a new constant due to inhibition. There are again several methods to estimate the parameters based on transformations of model (2) to a plot and application of regression techniques (see, e.g. Dixon and Webb (1979)). Some of these techniques are carried aout in two stages. First spreading out the data for the lower substrate concentrations, so that the greater part of the plot represents the region without inhibition in where MM model is obeyed, and estimating so V and K_m. And then using these estimations for estimating K_i.

In 1969, Westley suggested that if high–substrate inhibition is due to two substrate molecules binding in incorrect orientation to the enzyme, it is possible that a proportion of the bound substrate molecules may be disoriented even at low substrate concentration at which no substrate inhibition is apparent. This would be in fact a case of competitive inhibition in which substrate competes with itself. So, in this case

$$E[v] = \frac{Vs}{K_m + s(1 + K_m/K_i)} \tag{3}$$

The construction of D–optimal and c–optimal designs using Elving's method for these to last models is considered in this paper. D–optimal designs provide the best estimation of all the parameters jointly. And c–optimal designs deal with the best estimation of a linear combination of the parameters.

2. OPTIMAL DESIGNS IN ENZYME REACTIONS

As a first step, some main results for optimal designs for nonlinear regression models are introduced. These models are interesting because models (1) up to (3), in previous section, are of this type. Working with optimal designs for nonlinear models is more difficult than dealing with linear ones, because the designs are dependent on unknown parameters.

Let $Y = f(x, \theta) + \epsilon$, be a general regression model where $f(x, \theta)$ is a nonlinear function of θ. Y is the response for $x \in \mathcal{X}$ where $\mathcal{X}$ is the design space and $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^T$ is the vector of unknown parameters. The random error ϵ is normally distributed with mean zero and constant variance σ^2. If $f(x, \theta)$ is differentiable with respect to θ, its gradient vector is $\nabla f(x, \theta) = \partial f(x, \theta)/\partial\theta$. An approximate design ξ is a collection of n different support points $x_1, \ldots, x_n$ and weights $\xi(x_1), \xi(x_2), \ldots, \xi(x_n)$, defining a probability on $\mathcal{X}$. The Fisher information matrix $M(\xi, \theta)$ from a design ξ reflects the information given by ξ. This matrix is

$$M(\xi, \theta) = \sum_{i=1}^{n} \xi(x_i) \nabla f(x_i, \theta) \nabla f(x_i, \theta)^T.$$

For independent observations, $M^{-1}(\xi, \theta)$ is proportional to the asymptotic covariance matrix for the maximum likelihood estimate for θ.

The goal of the optimal design is the maximization of a function ϕ of the information matrix, called criterion. For nonlinear models, the information matrix depends on some unknown parameters, so the design must be obtained from a best guess θ^0. In this case, the optimal design is said locally optimal design.

The local D–optimality for a guess θ^0 maximizes the determinant of the information matrix to obtain the best estimations for the whole set of parameters. A design ξ^* is locally D–optimal design if $\nabla f(x, \theta^0)^T M^{-1}(\xi^*, \theta^0) \nabla f(x, \theta^0) \leq k$ for all $x \in \mathcal{X}$ and with equality on the support points (General Equivalence Theorem for D–optimality)

The local c-optimality criterion minimizes the variance of a linear combination of the unknown parameters $c^T\theta$. This criterion obtains the best estimation for an only parameter θ_i when c is the canonical vector of order i. In this case, the θ_i–optimal design minimizes the variance of the estimator for θ_i. It is reached by minimizing the ith element in the diagonal of $M^{-1}(\xi, \theta)$.

In the literature, there exist several results to characterize the D–optimal design for the MM model. A best guess K_m^0 is used for K_m, the parameter with a nonlinear behavior. Following proposition gives the D–optimal design for model (1).

Proposition The locally D–optimal design for the MM model (1) with $K_m = K_m^0$ in the design space $[0, b]$ is equally supported on two points, $K_m^0 b/(b + 2K_m^0)$ and b, the highest limit of the interval. That is,

$$\xi_D^* = \left\{ \begin{array}{cc} K_m^0 b/(b + 2K_m^0) & b \\ 1/2 & 1/2 \end{array} \right\}$$

Haines (1992) shows explicit formulae for the support points for the D–optimal design for the three parameters inverse quadratic polynomial model:

$$f(x, \theta) = \frac{x}{\beta_0 + \beta_1 x + \beta_2 x^2} \tag{4}$$

It is possible to rewrite model (2) as a particular case of (4), setting $\beta_0 = K_m/V$, $\beta_1 = 1/V$ and $\beta_2 = 1/(K_i V)$. Through Haines (1992) results the following proposition provides the expression of the support points for the locally D–optimal design, for the best guesses K_i^0 and K_m^0.

Proposition The locally D–optimal design for model (2) with $K_i = K_i^0$ and $K_m = K_m^0$ is equally supported on three points with values in a geometric progression. This locally D–optimal design is:

$$\xi^* = \left\{ \begin{array}{ccc} \rho^{-1}\sqrt{K_i^0 K_m^0} & \sqrt{K_i^0 K_m^0} & \rho\sqrt{K_i^0 K_m^0} \\ 1/3 & 1/3 & 1/3 \end{array} \right\}$$

with

$$\rho = \frac{\delta + \sqrt{\delta^2 - 4}}{2}, \quad \delta = \frac{\gamma + 1 + \sqrt{\gamma^2 + 6\gamma + 33}}{2}, \quad \gamma = \sqrt{\frac{K_i^0}{K_m^0}}.$$

In other respect, the MM model is defined, mathematically, by $y = \dfrac{bx}{a + x}$ so establishing the equalities $b = \dfrac{VK_i}{K_i + K_m}$, $a = \dfrac{K_i K_m}{K_i + K_m}$ it is possible to put model (3) in that form.

So following proposition shows its locally D–optimal design for the best guesses K_i^0 and K_m^0.

Proposition The locally D–optimal design for model (3) with $K_i = K_i^0$ and $K_m = K_m^0$ is the equally supported design:

$$\xi^* = \left\{ \begin{array}{cc} \frac{K_i^0 K_m^0 b}{b(K_i^0 + K_m^0) + 2K_i^0 K_m^0} & b \\ 1/2 & 1/2 \end{array} \right\}$$

In order to obtain c–optimal designs Elfving's method is a well known graphical procedure (Elfving, 1952). To avoid the difficulty on the construction of the Elfving set López–Fidalgo & Rodríguez–Díaz (2004) proposed a computational procedure for finding c–optimal designs. This procedure can be used for finding c–optimal designs for models (2) and (3). The procedure must work on $\nabla v(s, \theta)$ cause these models are non-linear models, where $\theta = \{V, K_m, K_i\}$.

3. EXAMPLE

In 2000, Lizcano, Tipton and Unzeta studied the activity of ox lung microsomal semicarbazide-sensitive amine oxidase (EC 1.4.3.6; SSAO) towards bensylamine increased 20-fold during incubation al 37C. Incubation os this preparation at 37C resulted in increases in fluorescence and activity similar to those with the unactivated enzyme. Benzilamine oxidation was inhibited, uncompetitively with respect to oxygen, by high substrate concentrations. Activation resulted in an increase in V for benzylamine oxidation, with no significant alterations in the K_m or the K_i for high-substrate inhibition by benzilamine.

SSAO from crude microsomal membranes was preincubated at different times at 37C and then assayed using benzilamine at different concentrations $(25 - 4000\mu M)$. Table 1 shows the changes in the kinetic parameters for benzilamine oxidation by SSAO after 3 hours of preincubation os the enzime at 37C. Values are the means ± standard deviations.

Preincubation (h)	$K_m(\mu M)$	$K_i(\mu M)$	$V\ (nmol \cdot min^{-1} \cdot mg^{-1})$
0	40 ± 5	1310 ± 75	2.09 ± 0.15
3	46 ± 5	1300 ± 37	25.56 ± 0.98

Tabla 1: Kinetic parameters without and after 3 hours of preincubation

If model (2) is supposed, and concentration of substrate can vary from 25 up to $4000\mu M$, the D– and θ_i–optimal designs will be:

Without preincubation
Initial values:
$V = 2.09, \quad K_m = 40, \quad K_i = 1310$

$$\xi_D = \left\{ \begin{array}{ccc} 28 & 229 & 1887 \\ 0.333 & 0.333 & 0.333 \end{array} \right\}$$

$$\xi_V = \left\{ \begin{array}{ccc} 25 & 232 & 2954 \\ 0.217 & 0.538 & 0.245 \end{array} \right\}$$

$$\xi_{K_m} = \left\{ \begin{array}{ccc} 25 & 232 & 2954 \\ 0.524 & 0.330 & 0.146 \end{array} \right\}$$

$$\xi_{K_i} = \left\{ \begin{array}{ccc} 25 & 232 & 2954 \\ 0.119 & 0.301 & 0.580 \end{array} \right\}$$

After 3h preincubation
Initial values:
$V = 25.56, \quad K_m = 46, \quad K_i = 1300$

$$\xi_D = \left\{ \begin{array}{ccc} 31 & 245 & 1921 \\ 0.333 & 0.333 & 0.333 \end{array} \right\}$$

$$\xi_V = \left\{ \begin{array}{ccc} 25 & 246 & 3006 \\ 0.228 & 0.524 & 0.248 \end{array} \right\}$$

$$\xi_{K_m} = \left\{ \begin{array}{ccc} 25 & 246 & 3006 \\ 0.536 & 0.318 & 0.146 \end{array} \right\}$$

$$\xi_{K_i} = \left\{ \begin{array}{ccc} 25 & 246 & 3006 \\ 0.127 & 0.299 & 0.574 \end{array} \right\}$$

But supposing model (3), and concentration of substrate can vary from 25 up to $4000\mu M$, the D– and θ_i–optimal designs will be:

Without preincubation
Initial values:
$V = 2.09, \quad K_m = 40, \quad K_i = 1310$

$$\xi_D = \left\{ \begin{array}{cc} 26 & 150 \\ 0.500 & 0.500 \end{array} \right\}$$

$$\xi_V = \left\{ \begin{array}{ccc} 33 & 103 & 120 \\ 0.076 & 0.494 & 0.430 \end{array} \right\}$$

$$\xi_{K_m} = \left\{ \begin{array}{ccc} 29 & 41 & 111 \\ 0.425 & 0.485 & 0.090 \end{array} \right\}$$

$$\xi_{K_i} = \left\{ \begin{array}{ccc} 75 & 135 & 141 \\ 0.036 & 0.500 & 0.464 \end{array} \right\}$$

After 3h preincubation
Initial values:
$V = 25.56, \quad K_m = 46, \quad K_i = 1300$

$$\xi_D = \left\{ \begin{array}{cc} 28 & 150 \\ 0.500 & 0.500 \end{array} \right\}$$

$$\xi_V = \left\{ \begin{array}{ccc} 29 & 88 & 150 \\ 0.230 & 0.477 & 0.293 \end{array} \right\}$$

$$\xi_{K_m} = \left\{ \begin{array}{c} 108 \\ 1.000 \end{array} \right\}$$

$$\xi_{K_i} = \left\{ \begin{array}{c} 108 \\ 1.000 \end{array} \right\}$$

REFERENCES

Dette, H. and Wong, W.K. (1999). E-Optimal Designs for the Michaelis-Menten Model. Statistics & Probability Letters, 44(4), 405-408.

Dette, H. and Biedermann, S. (2003). Robust and Efficient Designs for the Michaelis–Menten Model. Journal of the American Statistical Association, 98, 679-686.

Dixon, M & Webb, E.C. (1979). Enzymes (3 Ed.). Longman. New York.

Duggleby, R.G. (1979). Experimental Designs for Estimating the Kinetic Parameters for Enzyme-Catalysed Reactions. Journal of Theoretical Biology, 81, 671-684.

Elfving, G. (1952). Optimum Allocation in Linear Regression Theory. The Annals of Mathematical Statistics, 23(2), 255-262.

Haines, L.M. (1992). Optimal design for inverse quadratic polynomials. South African Statistical Journal, 26, 25-41.

Lizcano, J.M., Tipton, K.F. and Unzeta, M. (2000). Time-dependent activation of the semicarbazide-sensitive amine oxidasa (SSAO) from ox lung microsomes. Biochemical Journal, 351(3), 789-794.

López–Fidalgo, J. and Wong, W.K. (2002). Optimal Designs for the Michaelis–Menten Model. Journal of Theoretical Biology, 215, 1-11.

López–Fidalgo, J. and Rodríguez–Díaz, J.M. (2004). Elfving's Method for m–dimensional Models. Metrika 59, 235-244.

Raaijmaakers, J.G.W. (1987). Statistical Analisys of the Michaelis-Menten Equation. Biometrics, 43, 793-803.

Song, D & Wong, W.K. (1998). Optimal two-points designs for the Michaelis-Menten model with heteroscedastic errors. Communications in Statistics, 27, 1503-1516.

Westley, J. (1969). Enzymic Catalysis. Harper & Row. New York.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

CARGA DE POBOACIÓN ESTACIONAL DOS CONCELLOS DE GALICIA

Torrado Solar, Pablo[1] e López Vizcaíno, María Esther[1]

[1] Instituto Galego de Estatística

RESUMO

A carga de poboación dun concello pode ser moi diferente da súa poboación residente xa que as persoas se desprazan entre concellos por motivos de turismo, estudos, traballo... Nos últimos anos, os encargados de deseñar as políticas públicas teñen un interese cada vez maior por coñecer a poboación que realmente soportan os concellos. O obxectivo deste traballo é expoñer a metodoloxía que emprega o Instituto Galego de Estatística para estimar a carga de poboación dos concellos de Galicia.

Palabras e frases chave: carga de poboación que pernoita, carga de poboación diúrna, método das compoñentes, splines, regresión loxística, optimización numérica

1. INTRODUCIÓN

O Padrón municipal de habitantes é o rexistro administrativo onde constan os residentes de cada concello. O artigo 15 da Lei 4/1996, do 10 de xaneiro, pola que se modifica a Lei 7/1985, do 2 de abril, reguladora das bases do réxime local establece que toda persoa que viva en España está obrigada a empadroarse no concello no que resida habitualmente e que no caso de que resida en máis dun concello deberá empadroarse unicamente naquel no que resida durante máis tempo ao ano.

O criterio de adscrición da poboación ao seu concello de residencia fai que o Padrón municipal de habitantes teña serias limitación á hora de analizar a carga de poboación que soportan os concellos, pois as persoas non só teñen interacción co seu concello de residencia senón tamén con aqueles concellos aos que se desprazan por motivos de turismo, estancias en segundas vivendas, estudos, traballo, problemas de saúde...

A información sobre a carga de poboación é de vital importancia para a toma de decisións que afecten ás entidades locais (concellos) e ás súas agrupacións (mancomunidades, comarcas ou áreas metropolitanas) como poden ser a dotación de servizos sanitarios, construción de plantas de depuración de augas residuais, planificación do transporte público...

O Instituto Nacional de Estadística (INE) introduciu no Censo de poboación e vivendas do 2001 o concepto de poboación vinculada co fin de aproximar a carga de poboación dos concellos. O INE define a poboación vinculada a un concello como o conxunto de persoas que residen, traballan, estudan ou pasan máis de 14 noites ao ano nese concello. Agora ben, o concepto de poboación vinculada presenta limitacións:

- Non inclúe a poboación que se despraza por problemas de saúde (ingresos e consultas externas) nin a que pasa 14 ou menos noites ao ano nun concello distinto ao de residencia por motivos que non sexan o traballo nin os estudos (fins de semana, vacacións...).
- A carga de poboación presenta estacionalidade polo que resultaría interesante dispor de información trimestral, pero o Censo de poboación e vivendas só proporciona datos con data de referencia o 1 de novembro.
- A periodicidade do Censo de poboación e vivendas é decenal, así pois, existen períodos de tempo moi longos sen información actualizada.
- O Censo de poboación e vivendas do 2011 non se realizou como unha investigación exhaustiva senón como unha operación estatística baseada na combinación de rexistros administrativos e unha enquisa por mostraxe, polo tanto, non é posible obter información con tanto nivel de desagregación como no correspondente ao 2001.

No Decreto 188/2016, do 22 de decembro, polo que se aproba o Programa estatístico anual da Comunidade Autónoma de Galicia para 2017 introduciuse unha liña de actuación salientable para abordar o crecente interese por coñecer a carga de poboación que soportan os concellos de Galicia.

2. CONCEPTOS DE CARGA DE POBOACIÓN

O método das compoñentes úsase frecuentemente en demografía para calcular a poboación no ano *a+1* a partir da poboación no ano *a*, sumándolle os fluxos de entrada por nacemento e inmigración e restándolle os fluxos de saída por defunción e emigración.

Adaptando o método das compoñentes ao estudo da carga de poboación, defínense os conceptos de carga de poboación que pernoita e carga de poboación diúrna que se empregan neste estudo:

- A **carga de poboación que pernoita (CPP)** representa o número medio de persoas que pernoitan no concello e obtense a partir do número de residentes, sumándolle o número medio de persoas que pernoitan e non residen e restándolle o número medio de persoas que residen e non pernoitan:
  CPP = poboación residente + saldo medio de persoas aloxadas en establecementos turísticos + saldo medio de persoas aloxadas en segundas vivendas + saldo medio de persoas asociado ás pernoitas por estudos + saldo medio de persoas asociado ás pernoitas por traballo + saldo medio de persoas asociado ás pernoitas por problemas de saúde
- A **carga de poboación diúrna (CPD)** representa o número medio de persoas que están presentes no concello, independentemente de onde pernoiten, e obtense a partir do número de residentes, sumándolle o número medio de persoas que entran e restándolle o número medio de persoas que saen:
  CPD = poboación residente + saldo medio de persoas aloxadas en establecementos turísticos + saldo medio de persoas aloxadas en segundas vivendas + saldo medio de persoas asociado á mobilidade por estudos + saldo medio de persoas asociado á mobilidade por traballo + saldo medio de persoas asociado á mobilidade por problemas de saúde

Debe terse en conta que a estimación da carga de poboación dos concellos é unha estatística de síntese, polo tanto, a súa metodoloxía está en constante actualización e revisión co fin de incorporar novas fontes estatísticas que permitan mellorar a estimación das compoñentes da carga de poboación que pernoita e da carga de poboación diúrna ou incluso incorporar ás súas definicións novas compoñentes asociadas a motivos de desprazamento como as compras, trámites administrativos...

Dado o elevado grao de estacionalidade que presentan os desprazamentos das persoas, sobre todo os debidos ao turismo e ás estancias en segundas vivendas, realízanse estimacións trimestrais da carga de poboación que pernoita e da carga de poboación diúrna.

3. METODOLOXÍA

Nesta epígrafe exporase a metodoloxía que utiliza o IGE para estimar cada unha das compoñentes da CPP e da CPD, é dicir, as asociadas á poboación residente, turismo, estancias en segundas vivendas, estudos, traballo e problemas de saúde.

3.1. RESIDENTES

A operación estatística Cifras de poboación elaborada polo INE proporciona unha medida cuantitativa da poboación residente en España así como nas súas comunidades autónomas, provincias e illas. Trátase dunha estatística de síntese que emprega como información o Censo de poboación e vivendas do 2011, as Estatísticas do movemento natural da poboación, a Estatística de migracións e a Estatística de adquisición da nacionalidade española de residentes.

As cifras de poboación publicadas polo INE non chegan a detalle municipal, agora ben, na operación estatística Cifras poboacionais de referencia o IGE desagrega a nivel municipal as cifras de poboación que o INE establece para as provincias galegas.

Para o estudo da carga de poboación dos concellos de Galicia tómanse esas cifras poboacionais de referencia como poboación residente, en detrimento do Padrón municipal de habitantes.

A data de referencia das cifras poboacionais de referencia é o 1 de xaneiro, así pois, para o trimestre *t* do ano *a* tómase como poboación residente dun concello a media ponderada das cifras poboacionais de referencia dese concello para os anos *a* e *a+1* (P_a e P_{a+1}):

$$P_a^t = \frac{w \times P_a + (12 - w) \times P_{a+1}}{12}$$

onde *w=10,5* para *t=1*, *w=7,5* para *t=2*, *w=4,5* para *t=3* e *w=1,5* para *t=4*.

3.2. TURISMO

Na CPP trátase de aproximar para cada concello de Galicia a carga media diaria de poboación que pernoita en establecementos turísticos por trimestre. Empréganse os microdatos das seguintes enquisas elaboradas polo INE e que teñen periodicidade mensual:

- Enquisa de ocupación hoteleira
- Enquisa de ocupación en apartamentos turísticos
- Enquisa de ocupación en cámpings
- Enquisa de ocupación en aloxamentos de turismo rural
- Enquisa de turismo de residentes

Para a poboación que entra nos concellos

A variable de interese é o número de pernoitas correspondente á información agregada de todas as enquisas de ocupación a nivel de concello. Hai que ter en conta que estas enquisas non están preparadas para dar información a nivel de concello, só a nivel de provincia, área e punto turístico.

Para suavizar a variabilidade da información a nivel de concello, para cada mes, emprégase a media de dous anos consecutivos, é dicir, para o mes *m* do ano *a* considérase a media das pernoitas por concello dos datos do mes *m* dos anos *a-1* e *a*. A continuación, a partir dos datos mensuais, calcúlase para cada trimestre o total de pernoitas por concello.

As enquisas de ocupación non ofrecen información para todos os concellos de Galicia, só ofrecen información para aqueles nos que hai puntos de mostraxe. Nalgúns dos concellos onde non existen puntos de mostraxe hai establecementos turísticos, polo tanto, é posible que teñan pernoitas. Para estimar o número trimestral de pernoitas neses concellos séguese o seguinte procedemento:

1. Para cada trimestre axústase un modelo aditivo xeneralizado con distribución binomial negativa para os concellos onde hai puntos de mostraxe. No modelo de regresión, a variable resposta é o número trimestral de pernoitas e empréganse como variables explicativas:
   - Número de prazas abertas nos hoteis (*prazas_ht*)
   - Número de prazas abertas nos apartamentos turísticos (*prazas_ap*)
   - Número de prazas abertas nos cámpings (*prazas_cp*)
   - Número de prazas abertas nos aloxamentos de turismo rural (*prazas_tr*)
   - Indicador da área turística (*area*)
   - Indicador de concello costeiro - non costeiro (*costeiro*)

   A información do número de prazas abertas nos establecementos turísticos provén dos directorios de establecementos turísticos que elabora o INE.

   Así pois, o modelo de regresión empregado para cada trimestre é:

   $$pernoitas = c + s(prazas_ht) + s(prazas_ap) + s(prazas_cp) + s(prazas_tr) + area + costeiro$$

   onde as *s* representan funcións de suavizado (splines).
2. Unha vez axustado o modelo de regresión, emprégase para facer a predición do número trimestral de pernoitas naqueles concellos nos que non hai puntos de mostraxe pero si hai prazas abertas nos establecementos turísticos.
3. A suma das pernoitas obtidas non coincide coa estimación facilitada polas enquisas de ocupación a nivel galego. Así pois, para cada provincia, axústanse os datos de pernoitas multiplicando o dato obtido para cada concello polo factor de axuste:

$$\lambda_p = \frac{\sum_{concellos\ da\ provincia\ p} pernoitas\ (enquisas)}{\sum_{concellos\ da\ provincia\ p} pernoitas(regresión)} \times \frac{pernoitas\ en\ Galicia\ (enquisas)}{\sum_{concellos\ de\ Galicia} pernoitas\ (enquisas)}$$

Para transformar o número trimestral de pernoitas de cada un dos concellos en carga media diaria de poboación, hai que dividilo entre o número de días do trimestre.

Para a poboación que sae dos concellos

Para cada trimestre, o procedemento para estimar o número de pernoitas a nivel de concello consiste en aproximar o número total de pernoitas en establecementos turísticos que realizaron os residentes en Galicia e distribuílo entre os concellos.

Inicialmente, calcúlase o número total de pernoitas en establecementos turísticos que realizaron os residentes en Galicia como a suma das realizadas en establecementos turísticos do territorio nacional e das realizadas en establecementos turísticos do estranxeiro. O número de pernoitas en establecementos turísticos do territorio nacional obtense directamente das enquisas de ocupación e o número de pernoitas en establecementos turísticos do estranxeiro obtense da Enquisa de turismo de residentes aplicando un

coeficiente de enlace para garantir a coherencia entre a Enquisa de turismo de residentes e as enquisas de ocupación.

Para rematar, hai que distribuír o número total de pernoitas en establecementos turísticos que realizaron os residentes en Galicia entre os concellos. A distribución entre os concellos realízase en dous pasos; en primeiro lugar, distribúese entre catro grupos de concellos (concellos de menos de 10.000 residentes, concellos entre 10.000 e 20.000 residentes, concellos entre 20.000 e 50.000 residentes e concellos de 50.000 ou máis residentes) empregando a distribución proporcionada pola Enquisa de turismo de residentes e, en segundo lugar, dentro de cada grupo de concellos distribúese proporcionalmente ao número de residentes de cada concello.

O número trimestral de pernoitas de cada un dos concellos hai que transformalo en carga media diaria de poboación, para elo divídese entre o número de días do trimestre.

A metodoloxía descrita tamén se emprega para o cálculo da compoñente da CPD asociada ao turismo. Asúmese que esta poboación pasa o día completo no concello de destino.

3.3. SEGUNDAS VIVENDAS

Na CPP, para cada trimestre, o obxectivo é estimar para cada concello de Galicia a carga media diaria de poboación que pernoita en segundas vivendas por motivos non laborais nin de estudo. O procedemento para a estimación trimestral emprega como fontes de información o Censo de poboación e vivendas do 2011, a Enquisa de turismo de residentes e a Enquisa de ocupación en apartamentos turísticos:

1. A partir da Enquisa de turismo de residentes calcúlase o número total de pernoitas de entrada e saída en segundas vivendas por motivos distintos aos desprazamentos ao centro de estudos ou traballo.
   O número total de pernoitas de entrada obtense como o número de pernoitas asociadas ás etapas que realizaron os residentes en territorio nacional en segundas vivendas situadas en Galicia e o número total de pernoitas de saída obtense como o número de pernoitas asociadas ás etapas que realizaron os residentes en Galicia en segundas vivendas situadas en territorio nacional. Non se consideran as etapas correspondentes ás viaxes con motivo principal os desprazamentos ao centro de estudos ou traballo.
2. Na Enquisa de turismo de residentes considéranse pernoitas en segundas vivendas aquelas que se realizaron en vivendas completas en alugueiro (incluídos os apartamentos turísticos), habitacións en alugueiro en vivendas particulares, vivendas en propiedade, vivendas de familiares, amigos ou empresas cedidas gratuitamente, vivendas de uso compartido ou vivendas intercambiadas.
   Agora ben, os apartamentos turísticos son establecementos turísticos e non segundas vivendas, polo tanto, é necesario restar do número total de pernoitas de entrada e saída en segundas vivendas que proporciona a Enquisa de turismo de residentes o número total de pernoitas de entrada e saída en apartamentos turísticos que proporciona a Enquisa de ocupación en apartamentos turísticos.
3. O Censo de poboación e vivendas do 2011 permite coñecer para cada concello de Galicia o número anual de pernoitas de entrada e saída das persoas que ao longo do ano pasaron máis de 14 noites nun concello distinto ao de residencia por motivos non laborais nin de estudo (ao esixir que pasasen máis de 14 noites ao ano no mesmo concello, hipoteticamente elimínase a poboación que se aloxou en establecementos turísticos).
   Aplicando a distribución municipal de entrada e saída que proporciona o Censo de poboación e vivendas do 2011 ao número total de pernoitas de entrada e saída en segundas vivendas obtense o número de pernoitas de entrada e saída para cada concello de Galicia.
4. Para rematar, hai que transformar o número de pernoitas de entrada e saída obtido para cada concello de Galicia en carga media diaria de poboación. Así pois, divídese o número de pernoitas de entrada e saída entre o número de días do trimestre.

Esta metodoloxía tamén se emprega para estimar a compoñente da CPD asociada ás segundas vivendas. Asúmese que esta poboación pasa o día completo no concello de destino.

3.4. ESTUDOS

Na CPP trátase de aproximar para cada trimestre a carga media diaria de poboación que pernoita por estudos en cada concello de Galicia.

O rexistro de matriculados no Sistema Universitario Público de Galicia permite coñecer para cada estudante o concello de residencia e o concello de estudos. Así pois, pódese calcular para cada combinación de concello de residencia e concello de estudos o número de estudantes matriculados.

Para cada combinación de concello de residencia e concello de estudos determínase cantos deses estudantes matriculados pernoitan no concello de residencia e cantos no concello de estudos. Para isto, asúmese que os non residentes en Galicia pernoitan no concello de estudos e establécese para os residentes en Galicia o seguinte procedemento:

1. O Censo de poboación e vivendas do 2011 permite coñecer para cada residente en Galicia que estea realizando estudos universitarios en Galicia e que non traballe o seu concello de residencia, concello de estudos e concello onde pernoita os días lectivos. Así pois, para cada combinación de concello de residencia e concello de estudos dispónse do número de estudantes e da súa desagregación segundo estean desprazados ou non.
2. A partir desa información axústase o modelo de regresión loxística:

$$p(desprazado = 1) = \frac{\exp(\beta_0 + \beta_1 tempo + \beta_2 campus)}{1 + \exp(\beta_0 + \beta_1 tempo + \beta_2 campus)}$$

   onde *desprazado* toma o valor 1 cando o estudante pernoita no concello de estudos e 0 cando pernoita no concello de residencia, *tempo* é o tempo necesario para realizar por estrada o traxecto entre as capitais do concello de residencia e o concello de estudos e *campus* indica o campus universitario do estudante.
3. Unha vez axustado o modelo de regresión loxística estímase para cada combinación de concello de residencia e concello de estudos a probabilidade de que un estudante estea desprazado. Estas probabilidades estimadas permiten repartir para cada combinación de concello de residencia e concello de estudos o número de estudantes matriculados en estudantes desprazados e non desprazados.

Posteriormente, estímase para cada concello de Galicia o número de non residentes que pernoitan no concello por estudos (poboación de entrada) e o número de residentes que pernoitan noutro concello de Galicia por estudos (poboación de saída).

Para rematar, hai que transformar a poboación de entrada e a poboación de saída asociada a cada concello de Galicia en carga media diaria de poboación. Para isto, hai que multiplicar a poboación de entrada e a poboación de saída asociada a cada concello de Galicia polo cociente entre o número de días lectivos e totais do trimestre.

Na CPD é necesario estimar para cada trimestre a carga media diaria de poboación asociada á mobilidade por estudos en cada concello de Galicia. A información de base procede de tres fontes de información:

- A Estatística de ensinanza non universitaria de Galicia: permite saber para cada concello de Galicia o número de estudantes matriculados nos distintos niveis de ensinanzas non universitarias de réxime xeral (ensinanzas ordinarias: infantil, primaria, secundaria obrigatoria, bacharelato, ciclos formativos e programas de cualificación profesional inicial/formación profesional básica; ensinanzas de educación especial e ensinanzas de adultos: básicas iniciais, secundaria, bacharelato, ciclos formativos e formación profesional dual).
- O sistema de Xestión administrativa de educación: para os distintos niveis de ensinanzas non universitarias de réxime xeral, permite coñecer para cada concello de Galicia o número de estudantes matriculados nun centro educativo dese concello, o número de estudantes residentes no concello que están matriculados nun centro educativo de Galicia e o número de estudantes que residen no concello e están matriculados nun centro educativo dese concello.
- O rexistro de matriculados no Sistema Universitario Público de Galicia: permite ter para cada combinación de concello de residencia e concello de estudos o número de estudantes matriculados nunha universidade pública de Galicia.

A carga media diaria de poboación asociada á mobilidade por estudos é a suma da carga media diaria de poboación asociada á mobilidade por estudos non universitarios e da carga media diaria de poboación asociada á mobilidade por estudos universitarios.

Mobilidade por estudos non universitarios

Agrúpanse os distintos niveis de ensinanzas non universitarias de réxime xeral en tres grupos:

Grupos de estudos	
1	Ensinanzas ordinarias de infantil e primaria
2	Ensinanzas ordinarias excepto infantil e primaria: secundaria obrigatoria, bacharelato, ciclos formativos e programas de cualificación profesional inicial/formación profesional básica Ensinanzas de educación especial
3	Ensinanzas de adultos: básicas iniciais, secundaria, bacharelato, ciclos formativos e formación profesional dual

Para cada concello de Galicia, estímase a carga media diaria de poboación dun trimestre mediante o seguinte procedemento:

1. A Estatística de ensinanza non universitaria de Galicia permite obter para cada grupo de estudos o número de estudantes matriculados no concello, que denotaremos por $AM_g;\ g = 1,2,3$
2. O sistema de Xestión administrativa de educación permite calcular para cada grupo de estudos as seguintes proporcións:

$$P_g^1 = \frac{número\ de\ estudantes\ que\ residen\ e\ están\ matriculados\ no\ concello}{número\ de\ estudantes\ matriculados\ no\ concello};\ g = 1,2,3$$

$$P_g^2 = \frac{número\ de\ estudantes\ que\ residen\ e\ están\ matriculados\ no\ concello}{número\ de\ estudantes\ residentes\ no\ concello\ e\ matriculados\ en\ Galicia};\ g = 1,2,3$$

3. A poboación de entrada aproxímase por:

$$Poboación\ de\ entrada = \sum_{g=1}^{3} AM_g \times (1 - P_g^1)$$

4. A poboación de saída aproxímase por:

$$Poboación\ de\ saída = \sum_{g=1}^{3} \frac{AM_g \times P_g^1}{P_g^2} \times (1 - P_g^2)$$

5. Para rematar, asúmese que os estudantes non universitarios pasan a metade do día no concello de estudos, polo tanto, pódese obter a carga media diaria de poboación multiplicando a poboación de entrada e a poboación de saída por:

$$0{,}5 \times \frac{días\ lectivos\ do\ trimestre}{días\ totais\ do\ trimestre}$$

A necesidade de combinar a información da Estatística de ensinanza non universitaria de Galicia e do sistema de Xestión administrativa de educación débese a que a Estatística de ensinanza non universitaria de Galicia proporciona información do número de estudantes matriculados pero non recolle o concello de residencia dos estudantes. O sistema de Xestión administrativa de educación recolle tanto o concello onde están matriculados como o concello de residencia dos estudantes, agora ben, este rexistro administrativo presenta limitacións como que non inclúe a todos os centros educativos, non se somete a un proceso de depuración tan minucioso como a Estatística de ensinanza non universitaria de Galicia... Así pois, resulta adecuada a estratexia de combinar esas dúas fontes de información.

Mobilidade por estudos universitarios

O rexistro de matriculados no Sistema Universitario Público de Galicia permite calcular para cada concello de Galicia a poboación de entrada como o número de estudantes non residentes que cursan estudos nunha universidade pública dese concello e a poboación de saída como o número de estudantes residentes que cursan estudos noutro concello de Galicia nunha universidade pública.

Ademais, para cada concello de Galicia, pódese desagregar a poboación de entrada e a poboación de saída segundo os estudantes pernoiten no concello de residencia ou no concello de estudos. Para isto, emprégase o procedemento que se utilizou na estimación da carga media diaria de poboación que pernoita por estudos en cada concello de Galicia.

Esta desagregación resulta importante xa que se establece que os estudantes universitarios que pernoitan no concello de residencia pasan a metade do día no concello de estudos pero os que pernoitan no concello de estudos pasan neste o día enteiro.

Tendo en conta esa suposición e o número de días lectivos e totais do trimestre, pódese transformar a poboación de entrada e a poboación de saída asociada a cada concello de Galicia en carga media diaria de poboación multiplicando a parte que pernoita no concello de residencia por:

$$0{,}5 \times \frac{días\ lectivos\ do\ trimestre}{días\ totais\ do\ trimestre}$$

e a parte que pernoita no concello de estudos por:

$$\frac{días\ lectivos\ do\ trimestre}{días\ totais\ do\ trimestre}$$

3.5. TRABALLO

Para calcular a compoñente da CPP asociada á mobilidade por traballo o obxectivo é ter unha matriz de cargas medias diarias de poboación na que figuren os concellos de residencia e traballo dos traballadores. Para isto, empréganse os datos dos afiliados en alta laboral nas distintas mutualidades: Seguridade Social, Mutualidad General de Funcionarios Civiles del Estado (MUFACE), Instituto Social de las Fuerzas Armadas (ISFAS) e Mutualidad General Judicial (MUGEJU).

O procedemento para obter a carga media diaria de poboación asociada á mobilidade por traballo en cada concello de Galicia é:

1. Para cada combinación de concello de residencia e concello de traballo, calcúlase a carga media diaria de poboación asociada aos traballadores afiliados a cada unha das mutualidades.
2. Para cada concello, a carga media diaria de entrada é a suma das cargas medias diarias de poboación asociadas aos traballadores que traballan pero non residen no concello e a carga media diaria de saída é a suma das cargas medias diarias de poboación asociadas aos traballadores que residen pero non traballan no concello.
3. A carga media diaria de poboación de cada concello de Galicia correspóndese á diferenza entre a carga media diaria de entrada e a carga media diaria de saída.

Matriz de cargas medias diarias de poboación: afiliados á Seguridade Social

A matriz de cargas medias diarias de poboación asociada aos traballadores afiliados á Seguridade Social constrúese utilizando información dos ficheiros de afiliacións e contas de cotización á Seguridade Social facilitados trimestralmente pola Tesorería General de la Seguridad Social (TGSS), Padrón municipal de habitantes do INE, Directorio de empresas e unidades locais do IGE e Enquisa de poboación activa levada a cabo polo INE e IGE.

No ficheiro de afiliacións á Seguridade Social constan os datos identificativos dos afiliados, polo tanto, podemos asignar a cada afiliación como concello de residencia do afiliado o que aparece no Padrón municipal de habitantes. Para distribuír estas afiliacións por concello de traballo hai que distinguir entre afiliacións por conta propia e afiliacións por conta allea:

- Para as afiliacións por conta propia tómase como concello de traballo o concello do domicilio do réxime, excepto para as afiliacións para as que non ven cuberto, para as que se toma como concello de traballo o concello de residencia do afiliado.
- Para as afiliacións por conta allea coñécese o concello da conta de cotización á que está adscrita cada afiliación. Agora ben, non se pode tomar como concello de traballo de cada afiliación o concello da súa conta de cotización, xa que as empresas só están obrigadas a dispor dunha conta de cotización por cada colectivo de traballadores con peculiaridades de cotización e provincia na que realiza unha actividade económica.

  Cruzando o ficheiro de contas de cotización á Seguridade Social co Directorio de empresas e unidades locais, pódese indicar para cada conta de cotización os concellos onde a empresa ten establecementos asociados a esa conta de cotización e repartir por concellos o seu número de traballadores. Para as afiliacións asociadas a contas de cotización non multilocalizadas tómase como concello de traballo o concello onde a conta de cotización ten establecementos.

Neste punto, coñécese o concello de residencia de cada afiliación e o concello de traballo das afiliacións por conta propia e das afiliacións por conta allea adscritas a contas de cotización non multilocalizadas.

Aproximadamente o 25% das afiliacións por conta allea corresponden a contas de cotización multilocalizadas. Para distribuír esas afiliacións por concello de traballo constrúense 314 modelos de regresión loxística, un por cada concello de Galicia, co fin de estimar:

$$p_{ij}: probabilidade\ de\ que\ un\ afiliado\ por\ conta\ allea\ que\ reside\ no\ concello\ i\ traballe\ no\ concello\ j; i, j = 1 \dots 314$$

Para cada *j=1...314*, a partir das afiliacións por conta allea asociadas a contas de cotización non multilocalizadas, axústase un modelo de regresión loxística con variable dependente:

$$Y_j\text{: } 1 \; se \; o \; afiliado \; traballa \; no \; concello \; j \; e \; 0 \; se \; traballa \; nun \; concello \; distinto \; de \; j$$

e variables independentes:

$$D_j\text{: } distancia \; entre \; o \; concello \; de \; residencia \; do \; afiliado \; e \; o \; concello \; j$$

$$I^A = \frac{afiliacións \; con \; concello \; de \; traballo \; o \; concello \; de \; residencia \; do \; afiliado}{afiliacións \; con \; concello \; de \; residencia \; o \; concello \; de \; residencia \; do \; afiliado}$$

$$I_j^B = \frac{afiliacións \; con \; concello \; de \; traballo \; o \; concello \; de \; residencia \; do \; afiliado}{afiliacións \; con \; concello \; de \; traballo \; o \; concello \; j}$$

Se para *i,j=1...314* denotamos por N_i^r o número de afiliacións por conta allea adscritas a contas de cotización multilocalizadas con concello de residencia *i* (dato coñecido) e N_j^t o número de afiliacións por conta allea adscritas a contas de cotización multilocalizadas con concello de traballo *j* (dato coñecido), pódese estimar o número de afiliacións correspondentes a contas de cotización multilocalizadas con concello de residencia *i* e concello de traballo *j* mediante:

$$\widehat{N_{ij}} = N_i^r \times \widehat{p_{ij}}$$

agora ben, estas estimacións non cumpren:

$$\sum_{i=1}^{314} \widehat{N_{ij}} = N_j^t;\; j = 1 \dots 314$$

$$\sum_{j=1}^{314} \widehat{N_{ij}} = N_i^r;\; i = 1 \dots 314$$

Polo tanto, é necesario calcular uns $\widehat{N_{ij}}^c \geq 0; i,j = 1 \dots 314$ tales que:

$$Min \sum_{i=1}^{314} \sum_{j=1}^{314} \left(\widehat{N_{ij}} - \widehat{N_{ij}}^c\right)^2$$

$$\sum_{i=1}^{314} \widehat{N_{ij}}^c = N_j^t;\; j = 1 \dots 314$$

$$\sum_{j=1}^{314} \widehat{N_{ij}}^c = N_i^r;\; i = 1 \dots 314$$

Unha vez resolto este problema de optimización numérica utilizando o servidor NEOS e Gurobi como optimizador de referencia, considéranse como estimacións definitivas das $p_{ij}; i,j = 1 \dots 314$:

$$\widehat{p_{ij}}^c = \frac{\widehat{N_{ij}}^c}{N_i^r}; i,j = 1 \dots 314$$

Finalmente, para construír a matriz de cargas medias diarias de poboación asociada aos traballadores afiliados á Seguridade Social hai que ter en conta que a carga media diaria de poboación que supón cada afiliación calcúlase ponderando mediante a súa xornada (proporción de horas de traballo asociadas a esa afiliación respecto ás horas de traballo dun traballador a tempo completo na mesma empresa) a carga media diaria de poboación que supoñen as horas de traballo efectivo dun traballador a tempo completo coa situación profesional (conta propia, conta allea) e actividade económica do establecemento no que traballa correspondentes a esa afiliación, que se calcula a partir da Enquisa de poboación activa.

Matrices de cargas medias diarias de poboación: afiliados a MUFACE, ISFAS e MUGEJU

A partir dos ficheiros de afiliados enviados polas secretarias xerais de MUFACE, ISFAS e MUGEJU constrúense as correspondentes matrices de afiliados. Nos ficheiros de afiliados a MUFACE, ISFAS e MUGEJU consta o concello de traballo de cada afiliado; en canto ao concello de residencia, só se pode tomar o do Padrón municipal de habitantes para os afiliados a ISFAS xa que para os afiliados a MUFACE e MUGEJU non se dispón dos seus datos identificativos, así pois, hai que tomar o que consta nos ficheiros de afiliados como concello de residencia.

Posteriormente, transfórmanse as matrices de afiliados en matrices de cargas medias diarias de poboación multiplicándoas pola carga media diaria de poboación que supoñen as horas efectivas de traballo dun asalariado do sector público a tempo completo segundo a Enquisa de poboación activa.

3.6. SAÚDE

Para cada trimestre e concello, o Servizo Galego de Saúde (Sergas) facilita ao IGE o número total de pernoitas por ingreso hospitalario nese concello dos non residentes e o número total de pernoitas por ingreso hospitalario noutro concello de Galicia dos residentes. Para obter a compoñente da CPP divídese o número trimestral de pernoitas de entrada e o número trimestral de pernoitas de saída entre o número de días do trimestre.

A compoñente da CPD asociada á mobilidade por problemas de saúde é a suma da carga media diaria de poboación asociada aos ingresos hospitalarios e da carga media diaria de poboación asociada ás consultas externas en hospitais.

O procedemento para obter a carga media diaria de poboación asociada aos ingresos hospitalarios é o que se utilizou para calcular a compoñente da CPP asociada ás pernoitas por problemas de saúde. Asúmese que esta poboación pasa o día completo no concello de destino.

Para cada trimestre e concello, o Sergas facilita ao IGE o número total de consultas externas en hospitais dese concello dos non residentes e o número total de consultas externas en hospitais doutro concello de Galicia dos residentes. A carga media diaria de poboación asociada ás consultas externas en hospitais obtense multiplicando o número trimestral de entradas e o número trimestral de saídas por:

$$\frac{0{,}25}{días\ do\ trimestre}$$

4. CONCLUSIÓNS

O IGE aproveita as fontes estatísticas que proporcionan información sobre a poboación residente e as persoas que se desprazan por motivos de turismo, estancias en segundas vivendas, estudos, traballo e problemas de saúde para levar a cabo a estatística de síntese Carga de poboación estacional dos concellos de Galicia. A CPP e a CPD tratan de cuantificar a poboación que realmente soportan os concellos galegos, aspecto que xoga un papel cada vez máis importante no deseño das políticas públicas.

REFERENCIAS

Vinuesa, J. (2005) De la población de hecho a la población vinculada. Cuadernos geográficos de la Universidad de Granada, núm. 36, páx. 79-90

Módenes, J. A. (2008) Movilidad espacial, habitantes y lugares: retos conceptuales y metodológicos para la geodemografía. Estudios geográficos del CSIC, núm. 264, páx157-178

Sánchez, J. A., Pérez, A. y Rovira, C. (2013) Estimaciones de población estacional: metodología de cálculo para los municipios catalanes. http://www.idescat.cat/es/poblacio/estacional.html

Ayuntamiento de Madrid (2015), La población flotante de la ciudad de Madrid. Barómetro de economía de la ciudad de Madrid, núm. 45, páx. 85-100

Wood, S. N. (2006) Generalized Additive Models: an introduction with R. Chapman & Hall/CRC

Faraway, J. J. (2006) Extending the linear model with R. Chapman & Hall/CRC

Crawley, M. J. (2007) The R Book. John Wiley & Sons, Ltd.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

AUTO-DUALIDADE EN PROBLEMAS DE ALUGUER DE TERREOS

Alfredo Valencia-Toledo[1] e Juan Vidal-Puga[1]

[1]Universidade de Vigo

RESUMO

Supoñamos que existe unha empresa que necesita alugar certa cantidade de terras de varios propietarios. Estes propietarios posúen, en total, máis terreo do que precisa a empresa. A forma na que os propietarios alugan os seus terreos pode considerarse equivalente a un problema de bancarrota onde os fondos a repartir corresponden ao terreo requirido pola empresa e cada demanda é a cantidade de terreo dispoñible por cada propietario. Neste artigo, exploramos a idea de auto-dualidade en problemas de bancarrota nestas situacións de aluguer de terras. Auto-dualidade di que se debe seguir o mesmo procedemento de reparto tanto se o que se están a repartir son as terras a alugar como se o que se están a repartir son as terras non alugadas. Presentamos unha completa caracterización da familia de regras de reparto que verifican esta idea de auto-dualidade. En particular, a auto-dualidade é suficiente para asegurar que a cantidade de terreo debe ser repartida proporcionalmente. Isto non ocorre, por exemplo, en problemas de bancarrota, onde outras regras de reparto, aparte da proporcional, verifican auto-dualidade. Tamén caracterizamos unha única regra que verifica auto-dualidade e outras propiedades razoables.

SUMMARY

Assume there is a single tenant who needs to rent a particular amount of land from several lessors. Moreover, lessors own more than this particular amount. The way lessors rent their land can be seen as equivalent to a bankruptcy problem where the estate is the required land and each claim is the amount of land available from each lessor. In this paper, we explore the idea of self-duality in bankruptcy problems to these land rental situations. Roughly speaking, self-duality says that it is equivalent to share the rented land than sharing the land that is not rented. We provide a complete characterization of the family of rules that satisfy this self-duality. In particular, self-duality is enough to assure that the amount of land is shared under the proportional rule, as opposed to bankruptcy problems, where other properties are also required. We also provide the characterization of a unique rule satisfying self-duality and other reasonable properties.

Palabras e frases chave: Self-duality, land rental problems, bankruptcy problems, proportional rule.

1. INTRODUCIÓN

Duality is an important concept in mathematics which usually implies the involution operation, i.e. given x and its dual y, then the dual of y is x. Even though duality has different interpretations depending on the area, we focus on the interpretation in bankruptcy problems (O'Neill, 1982), which is consequently extended to a land rental problem context.

A land rental problem involves several lessors, and a single tenant, who needs to rent land, but less than all land that is owned by the lessors. The main question is knowing how much land each lessor should rent and the price per unit of land (Valencia-Toledo and Vidal-Puga, 2015).

In order to set the concept of self-duality in land rental problems, let us introduce bankruptcy problems as follows: these are situations where a group of agents (N) have some claims ($c \in \Re_+^N$) that sum up more than the particular amount of land demanded by the tenant (E), which is a perfectly divisible estate to share among them, i.e. $\sum_{i \in N} c_i > E$. A bankruptcy rule is a way to share the total estate among the claimants. For each claimant, her share can be seen either as an award of part of her claim or as a loss of the other part.

The self-duality principle in bankruptcy problems was firstly studied by Aumann (1985). It states that the criterion used to grant awards should coincide with the one used to grant losses. Many important rules in the context of bankruptcy problems satisfy self-duality, for example, the proportional rule (Young, 1988, Herrero and Villar, 2001, 2003), the Talmut rule (Chun, 1988), the Contested Garment principle (Dagan, 1996), and the reverse Talmud rule (van den Brink, 2013). For a survey, we refer to Thomson (2003, 2015) and Moulin (2002).

We extend the idea of self-duality to land rental problems. The lessors can be seen as claimants in a bankruptcy problem. Lessors' claims are the land they own, and the estate is the amount of land that the tenant needs. Moreover, land renting problems have other elements to take into account, as for example the fact that the tenant is also an agent[1]. There is a revenue, or reservation price, that every lessor obtains by each unit of non-rented land, which is denoted by r. As opposed, the revenue that the tenant can get by each unit of rented land is denoted by R. We always assume that there exists benefit of cooperation, i.e. $R > r$.

We apply the concept of self-duality to the land share. We can think in terms of how to share the amount of land that is not rented instead of the amount of rented land, since they are equivalent problems.

In order to study self-duality, let us define the dual problem of a land rental problem as follows. The agents and the amount of available land remain unchanged. Meanwhile, the dual of the amount of land that the tenant needs is given by $E^d = c(N) - E$. The dual of the reservation price per unit of land (denoted by r^d) satisfies $r^d E^d = rE$. This is done so that the revenue for the lessors are equal in both the original problem and its dual problem. Also, the dual of the revenue per unit of land for the tenant (denoted by R^d) satisfies $R^d E^d = RE$. This means that the revenue for the tenant are equal in both the original problem and its dual problem.

We provide the complete characterization of the family of rules that satisfy self-duality. Under these rules, the amount of land is shared by the proportional rule. As opposed, in bankruptcy problems, it is necessary at least to consider some additional properties apart from self-duality.

In order to characterize a unique rule as a main result, we introduce two reasonable properties of the literature.

Price independence property says that the price does not depend on the distribution among the lessors of the available land. This property is related to the idea of land reassignment. If it does not hold, lessors could redistribute their land in order to increase the price and end up better off than they were initially. The property assures that even if they redistribute their land, so that the aggregate land does not change with respect to the initial land rental problem, the price does not change either. Consequently, the tenant and lessors generate and define a new land rental problem in which at least a pair of lessors changes her available land. Then, a proposal is said to be price independence if they obtain the same price in both land rental problems.

Standard for two-person property was introduced by Hart and Mas-Colell (1989). The idea is that whether there are two agents, the total benefit is shared as follows: the corresponding to each one plus half of the benefit of cooperation that they generate. We make an extension and study this principle in land rental problems. This means that whenever there are a tenant and a unique lessor, they share equally the final benefit of cooperation.

Properties that provide conditions on how the share should be for the 2-person case have played an important role in the axiomatization of many solutions in different contexts. For example, in cooperative games (Hart and Mas-Colell, 1989, van den Brink and Funaki, 2009, Huettner, 2015, van den Brink et al., 2016), in bankruptcy problems (Thomson, 2008), and in land renting problems (Valencia-Toledo and Vidal-Puga, 2015).

The remainder of this paper is organized as follows. In Section 2, we present the model. In Section 3, we define self-duality and we provide a complete characterization of the family of rules

[1] Cska and Herings (2017) also study an extension of bankruptcy problems in which the estate holder is an explicit agent.

that satisfy self-duality. Also, we present charaterizations of a subfamily of rules, adding price independence, and a unique rule, adding standard for two. In Section 5, we give some concluding remarks.

2. NOTATION AND MODEL

We denote the set of nonnegative real numbers as $\Re_+$. We denote the set of positive real numbers as $\Re_{++}$. Let $\{1, 2, \ldots\}$ be the set of potential lessors and let $\mathcal{N}$ be the set of their non-empty finite subsets. Let $N = \{1, 2, \ldots, n\} \in \mathcal{N}$ be a generic set of lessors, and let S be a generic subset of N. Let $\Re^S$ be the $|S|$-dimensional Euclidean space, whose coordinates are indexed by the elements of S. Given $y \in \Re^S$, we write $y(S) = \sum_{i \in S} y_i$. Given $x, y \in \Re^S$, we write $x \leq y$ when $x_i \leq y_i$ for all $i \in S$. Given $x \in \Re^N$ and $S \subset N$, we denote the restriction of x on S as x_S i.e. $x_S = (x_i)_{i \in S}$. Moreover, the vectors $(0, \ldots, 0) \in \Re^S$ and $(1, \ldots, 1) \in \Re^S$ are denoted by 0_S and 1_S, respectively.

A *land rental problem* is a tuple $L = (N_0, c, r, K, E)$ where $N_0 = \{0\} \cup N$ is the set of agents with 0 the unique tenant and N the set of lessors, $c \in \Re^N_{++}$ is the vector whose coordinates represent the amount of available land for each lessor, $r \in \Re_{++}$ is the reservation price per unit of land for the lessors, $E \in]0, c(N)[$ is the amount of land that the tenant needs to rent, and $K \in \Re_{++}$ is the revenue that the tenant can obtain from that amount of rented land. In order to assure that there exists benefit of cooperation, we assume $K > rE$. We define $R = \frac{K}{E}$ as the revenue per unit of land that the tenant gets with her activity. For convenience, we denote (N, c, r, R, E) instead of (N_0, c, r, K, E). Let $\mathcal{L}$ be the set of all land rental problems.

An efficient feasible agreement is a pair $(p, x) \in \Re_+ \times \Re^N_+$ where p denotes the price per unit of land, and $x \leq c$ and $x(N) = E$, so that x_i denotes the land rented by lessor $i \in N$. We denote A^L as the set of feasible agreements on a land rental problem L. Let $\mathcal{A} = \bigcup_{L \in \mathcal{L}} A^L$ be the set of all potential feasible agreements.

Given $(p, x) \in A^L$, the utility for tenant and each lessor $i \in N$ are $u^L_0(p, x) = (R - p)E$ and $u^L_i(p, x) = (p - r)x_i$, respectively. Notice that with a slight notation abuse and for simplicity, we write $u^L_j(p, x)$ instead of $u^L_j((p, x))$ for all $j \in N_0$.

We define a *rule* as a function $\psi : \mathcal{L} \longrightarrow \mathcal{A}$ that assigns to each problem $L = (N, c, r, R, E) \in \mathcal{L}$ a pair $\left(p^\psi(L), x^\psi(L)\right) = \psi(L) \in A^L$ satisfying the following conditions:

(i) *Scale invariance* says that the final price and the amount of rented land are independent of changes of scale i.e. $p^\psi\left(N, \beta c, \frac{\alpha}{\beta} r, \frac{\alpha}{\beta} R, \beta E\right) = \frac{\alpha}{\beta} p^\psi\left(N, c, r, R, E\right)$ and $x^\psi\left(N, \beta c, \frac{\alpha}{\beta} r, \frac{\alpha}{\beta} R, \beta E\right) = \beta x^\psi\left(N, c, r, R, E\right)$ for all $\alpha, \beta > 0$.

(ii) *Strict individual rationality* says that the price is greater than r and less than R, i.e. $r < p^\psi\left(N, c, r, R, E\right) < R$, in order to assure the benefit of the lessors and the tenant respectively.

(iii) *Continuity* says that the price and the amount of land are continuous functions.

Let Ψ be the set of all rules ψ.

3. A CHARACTERIZATION RESULT WITH SELF-DUALITY

Given a land rental problem $L = (N, c, r, R, E)$, we define its dual as $L^d = \left(N, c, r^d, R^d, E^d\right) \in \mathcal{L}$ where $E^d = c(N) - E$ and r^d and R^d are such that $r^d E^d = rE$ and $R^d E^d = RE$, respectively. Hence,

$$r^d = \frac{rE}{c(N) - E}$$

and

$$R^d = \frac{RE}{c(N) - E}.$$

We define self-duality as follows.

Definition A rule ψ satisfies self-duality if for all $L = (N, c, r, R, E) \in \mathcal{L}$, $u^L\left(p^\psi(L), x^\psi(L)\right) = u^{L^d}\left(p^\psi\left(L^d\right), x^\psi\left(L^d\right)\right)$ and $x^\psi(L) = c - x^\psi\left(L^d\right)$.

In order to characterize the rules that satisfy self-duality, we define

$$\Theta = \left\{ (e,t) \in \left]0, \frac{1}{2}\right] \times]0,1[\right\}$$

and

$$\Delta^{\mathcal{N}} = \left\{ \hat{c} \in \Re_{++}^{N} : N \in \mathcal{N}, \hat{c}(N) = 1 \right\}.$$

Let $\mathcal{F}$ be the set of continuous functions $f : \Theta \times \Delta^{\mathcal{N}} \to \Re$ with $f(e,t,\hat{c}) \in]t,1[$ for all $(e,t) \in \Theta$ and $\hat{c} \in \Delta^{\mathcal{N}}$.

In Proposition 1, we state the characterization result.

Proposition 1 A rule ψ satisfies self-duality iff there exists $f \in \mathcal{F}$ such that for all $L = (N,c,r,R,E) \in \mathcal{L}$, the price is given by

$$p^{\psi}(L) = \begin{cases} \left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) & \text{if } c(N) > 2E \\ \left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) & \text{if } c(N) \leq 2E, \end{cases}$$

and the amount of land is given by $x^{\psi}(L) = \frac{E}{c(N)}c$.

Proof ($\Rightarrow$) Let ψ be a rule that satisfies self-duality. Let $L = (N,c,r,R,E) \in \mathcal{L}$. Preliminarily, we prove that the dual price satisfies $p^{\psi}(L^d) = \frac{E}{c(N)-E}p^{\psi}(L)$. By self-duality, the tenant's utility is $\left(R - p^{\psi}(L)\right)E = \left(R^d - p^{\psi}(L^d)\right)E^d$. This is equivalent to write $\left(R - p^{\psi}(L)\right)E = \left(\frac{RE}{c(N)-E} - p^{\psi}(L^d)\right)(c(N) - E)$.

Now, we prove that there exists a function $f \in \mathcal{F}$ such that the price is given as $p^{\psi}(L) = Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ if $c(N) > 2E$ and $p^{\psi}(L) = Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ if $c(N) \leq 2E$.

By scale invariance, $p^{\psi}(N,c,r,R,E) = Rp^{\psi}\left(N, \frac{c}{c(N)}, \frac{r}{R}, 1, \frac{E}{c(N)}\right)$. We define $f(e,t,\hat{c}) = p^{\psi}(N,\hat{c},t,1,e)$ for all $(e,t,\hat{c}) \in \Theta \times \Delta^{\mathcal{N}}$. By strict individual rationality, $t < f(e,t,\hat{c}) < 1$. Since the price (p^{ψ}) is continue function, f is continue. Hence, $f \in \mathcal{F}$.

On the one hand, when $c(N) > 2E$,

$$p^{\psi}(L) = Rp^{\psi}\left(N, \frac{c}{c(N)}, \frac{r}{R}, 1, \frac{E}{c(N)}\right) = Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right).$$

On the other hand, when $c(N) \leq 2E$,

$$\begin{aligned} p^{\psi}(L) &= \frac{c(N)-E}{E}p^{\psi}\left(L^d\right) \\ &= \frac{c(N)-E}{E}p^{\psi}\left(N, c, \frac{rE}{c(N)-E}, \frac{RE}{c(N)-E}, c(N)-E\right) \\ &= \frac{c(N)-E}{E}\frac{RE}{c(N)-E}p^{\psi}\left(N, \frac{c}{c(N)}, \frac{r}{R}, 1, \frac{c(N)-E}{c(N)}\right) \\ &= Rp^{\psi}\left(N, \frac{c}{c(N)}, \frac{r}{R}, 1, 1 - \frac{E}{c(N)}\right) \\ &= Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right). \end{aligned}$$

Finally, we prove that the amount of land is given by $x^{\psi}(L) = \frac{E}{c(N)}c$. By self-duality, for all $i \in N$, $\left(p^{\psi}(L) - r\right)x_i^{\psi}(L) = \left(p^{\psi}\left(L^d\right) - r^d\right)x_i^{\psi}\left(L^d\right)$. By the above expression and the last part of self-duality ($x^{\psi}(L) = c - x^{\psi}(L^d)$), it is equivalent to write $\left(p^{\psi}(L) - r\right)x_i^{\psi}(L) = \left(\frac{E}{c(N)-E}p^{\psi}(L) - \frac{rE}{c(N)-E}\right)\left(c_i - x_i^{\psi}(L)\right)$.

Since $p^{\psi}(L) > r$, the above expression is simplified to $(c(N) - E)\,x_i^{\psi}(L) = E\left(c_i - x_i^{\psi}(L)\right)$, from where we deduce $x_i^{\psi}(L) = \frac{Ec_i}{c(N)}$ for all $i \in N$.

($\Leftarrow$) Let ψ be a rule given by $p^{\psi}(L) = Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ if $c(N) > 2E$ and $p^{\psi}(L) = Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ if $c(N) \leq 2E$, for some $f \in \mathcal{F}$, and $x^{\psi}(L) = \frac{E}{c(N)}c$.

We prove that ψ satisfies self-duality. Since the definition of self-duality has two parts, we proceed as follows.

i We check first that $x^{\psi}(L) = c - x^{\psi}(L^d)$. Since the land is shared proportionally, i.e. $x^{\psi}(L) = \frac{E}{c(N)}c$, it is checked straightforward.

ii We check that $u^{L}\left(\psi(L)\right) = u^{L^d}\left(\psi(L^d)\right)$.

First, $c(N) > 2E$. For every $i \in N$, $u_i^{L}\left(\psi(L)\right) = \left(p^{\psi}(L) - r\right)x_i^{\psi}(L)$. Since $c(N) > 2E$, for every $i \in N$, we have

$$u_i^{L}\left(\psi(L)\right) = \left[Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]x_i^{\psi}(L).$$

Furthermore, for every $i \in N$, we have $u_i^{L^d}\left(\psi(L^d)\right) = \left(p^{\psi}(L^d) - r^d\right)x_i^{\psi}(L^d)$. Since $c(N) > 2E$ or $c(N) - 2E > 0$, $2E^d = 2(c(N) - E) = c(N) + c(N) - 2E > c(N)$, i.e. $c(N) < 2E^d$, so that $u_i^{L^d}\left(\psi(L^d)\right) =$

$$= \left[R^d f\left(1 - \frac{E^d}{c(N)}, \frac{r^d}{R^d}, \frac{c}{c(N)}\right) - r^d\right]\left[c_i - x_i^{\psi}(L)\right]$$

$$= \left[\frac{RE}{c(N) - E}f\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - \frac{Er}{c(N) - E}\right]\left[c_i - \frac{Ec_i}{c(N)}\right]$$

$$= \frac{E}{c(N) - E}\left[Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]\frac{(c(N) - E)c_i}{c(N)}$$

$$= \left[Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]\frac{Ec_i}{c(N)}$$

$$= \left[Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]x_i^{\psi}(L).$$

Second, $c(N) < 2E$. For every $i \in N$, $u_i^{L}\left(\psi(L)\right) = \left(p^{\psi}(L) - r\right)x_i^{\psi}(L)$. Since $c(N) < 2E$, for every $i \in N$, we have

$$u_i^{L}\left(\psi(L)\right) = \left[Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]x_i^{\psi}(L).$$

Furthermore, for every $i \in N$, we have $u_i^{L^d}\left(\psi(L^d)\right) = \left(p^{\psi}(L^d) - r^d\right)x_i^{\psi}(L^d)$. Since $c(N) < 2E$ or $c(N) - 2E < 0$, $2E^d = 2(c(N) - E) = c(N) + c(N) - 2E < c(N)$, i.e. $c(N) > 2E^d$, so that $u_i^{L^d}\left(\psi(L^d)\right) =$

$$= \left[R^d f\left(\frac{E^d}{c(N)}, \frac{r^d}{R^d}, \frac{c}{c(N)}\right) - \frac{E}{c(N) - E}r\right]\left[c_i - x_i^{\psi}(L)\right]$$

$$= \left[\frac{RE}{c(N) - E}f\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - \frac{rE}{c(N) - E}\right]\left[c_i - \frac{c_i E}{c(N)}\right]$$

$$= \frac{E}{c(N) - E}\left[Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]\frac{(c(N) - E)c_i}{c(N)}$$

$$= \left[Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]\frac{Ec_i}{c(N)}$$

$$= \left[Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right]x_i^{\psi}(L).$$

Third, $c(N) = 2E$, for every $i \in N$, we have $u_i^L(\psi(L)) =$

$$= \left(p^\psi(L) - r\right) x_i^\psi(L)$$

$$= \left[Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - r\right] \frac{Ec_i}{c(N)}$$

$$= \left[Rf\left(1 - \frac{E}{2E}, \frac{r}{R}, \frac{c}{2E}\right) - r\right] \frac{Ec_i}{2E}$$

$$= \left[\frac{RE}{2E-E} f\left(1 - \frac{2E-E}{2E}, \frac{r}{R}, \frac{c}{2E}\right) - \frac{rE}{2E-E}\right] \left[c_i - \frac{c_i E}{2E}\right]$$

$$= \left[\frac{RE}{c(N)-E} f\left(1 - \frac{c(N)-E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right) - \frac{rE}{c(N)-E}\right] \left[c_i - \frac{c_i E}{c(N)}\right]$$

$$= \left[R^d f\left(1 - \frac{E^d}{c(N)}, \frac{r^d}{R^d}, \frac{c}{c(N)}\right) - r^d\right] \left[c_i - x_i^\psi(L)\right]$$

Since $c(N) = 2E$, $2E^d = 2(c(N) - E) = 2(2E - E) = 2E = c(N)$, i.e. $c(N) = 2E^d$. Then, the above expression is equal to $u_i^L(\psi(L)) = \left(p^\psi(L^d) - r^d\right) x_i^\psi(L^d)$. Hence, $u_i^L(\psi(L)) = u_i^{L^d}\left(\psi(L^d)\right)$.

By efficiency, for the utility of tenant, the proof is analogous as before. Q.E.D.

The rule associated to $f \in \mathcal{F}$ is denoted as ψ^f.

The next property states that the price does no depend on the distribution of the land.

We formalize the property as follows.

Definition A rule ψ satisfies price independence if for all $L = (N, c, r, R, E) \in \mathcal{L}$ and $c' \in \Re_{++}^N$ such that $c'(N) = c(N)$, we have $p^\psi(L') = p^\psi(L)$ where $L' = (N, c', r, R, E)$.

In order to characterize the rules that satisfy self-duality and price independence, we define $\mathcal{G}$ as the set of continuous functions $g : \Theta \to \Re$ with $g(e,t) \in]t, 1[$ for all $(e,t) \in \Theta$.

In Proposition 2, we state the corresponding characterization result.

Proposition 2 A rule ψ satisfies self-duality and price independence iff there exists $g \in \mathcal{G}$ such that for all $L = (N, c, r, R, E) \in \mathcal{L}$, the price is given by

$$p^\psi(L) = \begin{cases} \left(\frac{E}{c(N)}, \frac{r}{R}\right), & \text{if } c(N) > 2E \\ \left(1 - \frac{E}{c(N)}, \frac{r}{R}\right), & \text{if } c(N) \leq 2E, \end{cases}$$

and the amount of land is given by $x^\psi(L) = \frac{E}{c(N)} c$.

Proof ($\Rightarrow$) Let ψ be a rule that satisfies self-duality and price independence. Let $L = (N, c, r, R, E) \in \mathcal{L}$. Under self-duality, by Proposition there exits $f \in \mathcal{F}$ such that $p^\psi(L) = Rf\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ if $c(N) > 2E$ and $p^\psi(L) = Rf\left(1 - \frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ if $c(N) \leq 2E$, and $x^\psi(L) = \frac{E}{c(N)} c$. Under price independence, for every $c' \in \Re_{++}^N$ with $c'(N) = c(N)$, we have $p^\psi(L') = p^\psi(L)$. We prove the existence of $g \in \mathcal{G}$ so that $p^\psi(L) = Rg\left(\frac{E}{c(N)}, \frac{r}{R}\right)$ if $c(N) > 2E$ and $p^\psi(L) = Rg\left(1 - \frac{E}{c(N)}, \frac{r}{R}\right)$ if $c(N) \leq 2E$. First, assume that $c(N) > 2E$. Since $c(N) = c'(N)$,

$$\left(\tfrac{E}{c'(N)}, \tfrac{r}{R}, \tfrac{c'}{c'(N)}\right) = \left(\tfrac{E}{c(N)}, \tfrac{r}{R}, \tfrac{c'}{c(N)}\right).$$

Then, $f\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c'}{c(N)}\right) = f\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$. Consequently, f does not depend on $\frac{c}{c(N)}$, and hence $g\left(\frac{E}{c(N)}, \frac{r}{R}\right) = f\left(\frac{E}{c(N)}, \frac{r}{R}, \frac{c}{c(N)}\right)$ is well-defined. Second, assume now that $c(N) \leq 2E$. Since $c'(N) = c(N)$,

$$Rf\left(1 - \tfrac{E}{c'(N)}, \tfrac{r}{R}, \tfrac{c'}{c'(N)}\right) = Rf\left(1 - \tfrac{E}{c(N)}, \tfrac{r}{R}, \tfrac{c'}{c(N)}\right).$$

Then, $f\left(1-\frac{E}{c(N)},\frac{r}{R},\frac{c'}{c(N)}\right)=f\left(1-\frac{E}{c(N)},\frac{r}{R},\frac{c}{c(N)}\right)$. Consequently, f does not depend on $\frac{c}{c(N)}$ and hence $g\left(1-\frac{E}{c(N)},\frac{r}{R}\right)=f\left(1-\frac{E}{c(N)},\frac{r}{R},\frac{c}{c(N)}\right)$ is well-defined.

($\Leftarrow$) Let ψ be a rule given by, for all $L\in\mathcal{L}$, $p^{\psi}(L)=Rg\left(\frac{E}{c(N)},\frac{r}{R}\right)$ if $c(N)>2E$ and $p^{\psi}(L)=Rg\left(1-\frac{E}{c(N)},\frac{r}{R}\right)$ if $c(N)\leq 2E$ for some $g\in\mathcal{G}$, and $x^{\psi}(L)=\frac{E}{c(N)}c$. Let $f\in\mathcal{F}$ defined as $f(e,t,\hat{c})=g(e,t)$ for all $(e,t)\in\Theta$ and $\hat{c}\in\Delta^{\mathcal{N}}$. Then, we write $p^{\psi}(L)=Rf\left(\frac{E}{c(N)},\frac{r}{R},\frac{c}{c(N)}\right)$ if $c(N)>2E$ and $p^{\psi}(L)=Rf\left(1-\frac{E}{c(N)},\frac{r}{R},\frac{c}{c(N)}\right)$ if $c(N)\leq 2E$. By Proposition , ψ satisfies self-duality. Let $L'\in\mathcal{L}$ be given as in the definition of price independence. We prove $p^{\psi}(L')=p^{\psi}(L)$. First, assume that $c(N)>2E$. Since $c'(N)=c(N)$. Then, we deduce

$$\begin{aligned}p^{\psi}(L') &= Rg\left(\frac{E}{c'(N)},\frac{r}{R}\right)\\ &= Rg\left(\frac{E}{c(N)},\frac{r}{R}\right)=p^{\psi}(L).\end{aligned}$$

Second, assume that $c(N)\leq 2E$. Since $c'(N)=c(N)$. Then, we deduce

$$\begin{aligned}p^{\psi}(L') &= Rg\left(1-\frac{E}{c'(N)},\frac{r}{R}\right)\\ &= Rg\left(1-\frac{E}{c(N)},\frac{r}{R}\right)=p^{\psi}(L).\end{aligned}$$

Q.E.D.

The rule corresponding to $g\in\mathcal{G}$ is denoted (with a slight abuse of notation) as ψ^{g}.

Let $\mathcal{L}^{2}\subset\mathcal{L}$ be the set of all rental problems with a unique lessor. A rule is standard for two if the tenant and the unique lessor share equally the benefit of cooperation. We formalise this property as follows.

Definition A rule ψ satisfies standard for two if for all $L\in\mathcal{L}^{2}$, $u_0^{L}\left(\psi(L)\right)=u_i^{L}\left(\psi(L)\right)$ where $N=\{i\}$.

In Theorem 1, we state the characterization of the unique rule that satisfies self-duality, consistency and standard for two.

Theorem 1 A rule ψ satisfies self-duality, price independence and standard for two iff for all $L\in\mathcal{L}$, $p^{\psi}(L)=\frac{R+r}{2}$ and $x^{\psi}(L)=\frac{E}{c(N)}c$.

Proof ($\Rightarrow$) Let ψ be a rule that satisfies self-duality, price independence and standard for two. Let $L=(N,c,r,K,E)\in\mathcal{L}$. By Proposition **??**, there exists $g\in\mathcal{G}$ such that $p^{\psi}(L)=Rg\left(\frac{E}{c(N)},\frac{r}{R}\right)$ if $c(N)>2E$ and $p^{\psi}(L)=Rg\left(1-\frac{E}{c(N)},\frac{r}{R}\right)$ if $c(N)\leq 2E$, and $x^{\psi}(L)=\frac{E}{c(N)}c$. We need to prove that $Rg\left(\frac{E}{c(N)},\frac{r}{R}\right)=\frac{R+r}{2}$ if $c(N)>2E$ and $Rg\left(1-\frac{E}{c(N)},\frac{r}{R}\right)=\frac{R+r}{2}$ if $c(N)\leq 2E$, or equivalently $g\left(e,t\right)=\frac{1+t}{2}$ if $\frac{1}{2}>e$ and $g\left(1-e,t\right)=\frac{1+t}{2}$ if $\frac{1}{2}\leq e$ for all $(e,t)\in\Theta$. Let $e=\frac{E}{c(N)}\in\left]0,\frac{1}{2}\right]$ and $t=\frac{r}{R}\in[0,1[$. By standard for two, for every $i\in N$, $u_0^{L'}\left(\psi(L')\right)=u_i^{L'}\left(\psi(L')\right)$. Assume first $\frac{1}{2}>e$. By efficiency, we have $\left(R-Rg\left(e,t\right)\right)E=\left(Rg\left(e,t\right)-r\right)E$. Then, we deduce that $g\left(e,t\right)=\frac{1+t}{2}$. Assume now $\frac{1}{2}\leq e$. By efficiency, we have $\left(R-Rg\left(1-e,t\right)\right)E=\left(Rg\left(1-e,t\right)-r\right)E$. Then, we deduce that $g\left(1-e,t\right)=\frac{1+t}{2}$.

($\Leftarrow$) Let ψ a rule given by $p^{\psi}(L)=\frac{R+r}{2}$ and $x^{\psi}(L)=\frac{E}{c(N)}c$. It is straightforward to check that $\psi=\psi^{g}$ for $g\left(e,t\right)=\frac{1+t}{2}$ for all $(e,t)\in\Theta$. By the Proposition 2, ψ satisfies self-duality and price dependence. Then, we prove that ψ satisfies standard for two. It is equivalent to prove that $u_0^{L}\left(\psi(L)\right)=u_i^{L}\left(\psi(L)\right)$. The left side of the equality is equal to $u_0^{L}\left(\psi(L)\right)=\left(R-\frac{R+r}{2}\right)E=\frac{R-r}{2}E$. Analogously, the right side of the equality is equal to $u_i^{L}\left(\psi(L)\right)=\left(\frac{R+r}{2}-r\right)E=\frac{R-r}{2}E$. Q.E.D.

We denote by ψ^{2} the unique rule which is characterized in Theorem 1.

We present the discussion about the independence of properties as follows. Whether we remove one the properties, we obtain more rules. First, let us consider the rule given by $p^{\psi}(L)=\frac{R+\alpha r}{1+\alpha}$ for $\alpha\in]0,1[$ and $x^{\psi}(L)=\frac{E}{c(N)}c$, which satisfies self-duality and price independence, but fails standard for two. Second, let us consider the rule given by $p^{\psi}(L)=r+\frac{R-r}{2c(N)}\max_{i\in N}c_i$ and $x^{\psi}(L)=\frac{E}{c(N)}c$,

which satisfies self-duality and standard for two, but fails price independence. Finally, let us consider the rule given by $p^{\psi}(L) = \frac{R+r}{2}$ and for every $i \in N$, $x_i^{\psi}(L) = \min\{c_i, \lambda\}$ where λ solves $\sum_{i \in N} \min\{c_i, \lambda\} = E$, which is known as CEA rule in bankruptcy problems. This rule satisfies standard for two and price independence, but fails self-duality. Therefore, these three properties are independent.

4. CONCLUDING REMARKS

In this paper, we have studied the duality concept in land rental problems. We have presented the self-duality concept based on the idea of self-duality in bankruptcy problems. In bankruptcy problems, for every agent, the final outcome coincides with the amount obtained from the share of the estate. But, in land rental problems, we have considered that for every agent, the utility (final outcome) depends on the share of the rented land (estate) and the price by each unit of land.

The proportional land share is only obtained by self-duality in land rental problems. As opposed, for example, in bankruptcy problems, it is necessary at least to consider some other properties. Furthermore, we have characterized the family of rules that satisfy self-duality. A rule in this family is denoted as ψ^f. We have characterized the subfamily of rules that satisfies self-duality and price independence. A rule in this subfamily is denoted as ψ^g . Finally, We have characterized the unique rule that satisfies self-duality, price independence and standard for two, which is denoted as ψ^2. We also provided the discussion about independence of properties.

The following table shows a summary of these results.

Property	ψ^f	ψ^g	ψ^2
Self-duality	Y	Y	Y
Price independence	-	Y	Y
Standard for two	-	-	Y

Táboa 1: Properties of rules in land renting problems.

REFERENCIAS

Aumann, R.J. and Maschler, M. (1985) Game theoretic analysis of a bankruptcy problem from the Talmud. Journal of Economic Theory, 36(2), 195–213.

Chun, Y. (1988) The proportional solution for rights problems.MathematicalSocial Sciences, 15(3), 231–246.

Csóka, P. and Herings, P. J.-J. (2017). Bankruptcy games where the estate is a player. In Moretti, S., editor, Abstracts of 13th European (formerly Spain-Italy-Netherlands) Meeting on Game Theory (SING13), page 42. Paris-Dauphine University.

Dagan, N. (1996). New characterizations of old bankruptcy rules.SocialChoice and Welfare, 13(1):51–59.

Hart, S. and Mas-Colell, A. (1989). Potential, value, and consistency.Econometrica, 57(3):589–614.

Herrero, C. (2003). Equal awards vs. equal losses: Duality in bankruptcy. InSertel, M. and Koray, S., editors,Advances in Economic Design, Studiesin Economic Design, pages 413–426. Springer Berlin Heidelberg.

Herrero, C. and Villar, A. (2001) The three musketeers: four classicalsolutions to bankruptcy problems.Mathematical Social Sciences, 42(3), 307–328.

Herrero, C. and Villar, A. (2002). Sustainability in bankruptcy problems.Top, 10(2), 261–273.

Huettner, F. (2015) A proportional value for cooperative games with acoalition structure.Theory and Decision, 78(2), 273–287.

Moulin, H. (2002). Axiomatic cost and surplus sharing. In Arrow, K., Sen,A., and Suzumura, K., editors,Handbook of Social Choice and Welfare,volume I, chapter 6, pages 289–357. North-Holland, Amsterdam.

ONeill, B. (1982) A problem of rights arbitration from the Talmud.Mathematical Social Sciences, 2(4), 345–371.

Thomson, W. (2003). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: a survey.Mathematical Social Sciences, 45(3), 249– 297.

Thomson, W. (2008). Two families of rules for the adjudication of conflicting claims. Social Choice and Welfare, 31(4), 667–692.

Thomson, W. (2015). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: An update.Mathematical Social Sciences, 74:41 –59.14

Valencia-Toledo, A. and Vidal-Puga, J. (2015). Non-manipulable rules forland rental problems. Working paper, Universidade de Vigo.

van den Brink, R., Chun, Y., Funaki, Y., and Park, B. (2016). Consistency,population solidarity, and egalitarian solutions for tu-games.Theory andDecision, 81(3), 427–447.

van den Brink, R. and Funaki, Y. (2009) Axiomatizations of a classof equal surplus sharing solutions for tu-games.Theory and Decision,67(3), 303–340.

van den Brink, R., Funaki, Y., and van der Laan, G. (2013). Characterizationof the reverse Talmud bankruptcy rule by exemption and exclusionproperties.European Journal of Operational Research, 228(2), 413–417.

Young, H. (1988) Distributive justice in taxation.Journal of EconomicTheory, 44(2), 321–335.

XIII Congreso Galego de Estatística e Investigación de Operacións
Ferrol, 26, 27 e 28 de outubro de 2017

IMPUTACIÓN DE DATOS DE PRECIPITACIÓN DIEZMINUTALES EN LA RED DE ESTACIONES DE METEOGALICIA

Vidal Paz, José[1], Rodríguez-Gómez, Benigno[2]

[1]Universidade da Coruña, Grupo de Arquitectura de Computadores (GAC), Facultad de Informática, Elviña, 15071 A Coruña, España (jose.vidal.paz@udc.es)

[2]Universidade da Coruña, Departamento de Enxeñaría Industrial, Escuela Universitaria Politécnica, Serantes, 15405 Ferrol, España (benigno.rodriguez@udc.es)

RESUMEN

La precipitación es una variable meteorológica importante en aplicaciones de ingeniería civil, hidrología y medioambientales. Para determinados trabajos de investigación es necesario contar con registros de datos diezminutales, pero en ocasiones las series temporales presentan datos erróneos y otros están ausentes, siendo necesario eliminar los erróneos y estimar a continuación valores coherentes para todos los datos faltantes. En este estudio se ha llevado a cabo la comparación entre dos métodos de imputación múltiple, *missForest* y *MICE*, para rellenar las series de precipitaciones en una estación meteorológica de Galicia a partir de los datos de sus estaciones vecinas más próximas. Se ha trabajado con 7 estaciones y los datos diezminutales de 10 años. Se propone una metodología para la evaluación de la bondad de ajuste y la precisión de las estimaciones proporcionadas por ambos métodos de imputación, consistente en crear series de datos incompletas a partir de patrones de pérdida de datos reales para cada mes e imputar los datos faltantes en distintas circunstancias. Esos resultados muestran que la precipitación está desfasada en las distintas estaciones debido a su situación geográfica pudiendo mejorarse la imputación con un desplazamiento de las series temporales, y por otro lado el método de *missForest* proporcionó mejores resultados en la imputación de esta variable meteorológica que *MICE*.

Palabras y frases clave: Imputación, datos faltantes, precipitación, meteorología. (Máximo de seis).

1. INTRODUCIÓN

En el año 2000 se crea Meteogalicia, como un organismo dependiente de la Consellería de Medioambiente, Territorio e Infraestructuras de la Xunta de Galicia fruto de un convenio con la Universidad de Santiago de Compostela. Actualmente este organismo gestiona una extensa red de estaciones meteorológicas (≈100 estaciones) y agrometeorológicas (≈50 estaciones) en Galicia, y dispone de un sitio web (http://www.meteogalicia.es) donde se pueden encontrar, no sólo las predicciones para los próximos días, sino también los datos históricos de las medidas tomadas con los distintos sensores de cada estación.

En general las bases de datos públicas (o compartidas) como la disponible en el sitio web de Meteogalicia tienen datos faltantes. Los investigadores y usuarios finales de esos datos públicos suelen disponer de distintas herramientas estadísticas y programas informáticos con los que realizar inferencias estadísticas a partir de esa información, pero la validez de esas inferencias puede quedar en entredicho por el hecho de que esas herramientas y programas suelen estar diseñados para trabajar con conjuntos completos de datos.

Dado que los constructores de esas bases de datos públicas y los usuarios finales suelen ser entidades diferentes, éstos últimos no tienen información suficiente que les indique en qué circunstancias o cuáles son las razones por las que se han perdido parte de los datos, y aunque dispusieran

de los recursos adecuados, no podrían aplicar ningún método que les permitiese reconstruir fielmente la información perdida. Es bajo estas circunstancias cuando tiene sentido aplicar métodos estadísticos que permitan rellenar, también denominado "*imputar*", los datos faltantes.

En las aplicaciones de ingeniería civil, hidrología y medioambientales, como pueden ser el diseño de redes de saneamiento o cálculo de caudales en cuencas hidrográficas, existe la necesidad de disponer de datos diezminutales de precipitación que nos permitan estimar las curvas intensidad-duración-frecuencia (IDF) y realizar su ajuste a las distribuciones de Gumbel y GEV, elaborar hietogramas sintéticos, etc ...

Los primeros métodos que se empleaban para rellenar las series temporales simplemente calculaban la media de la precipitación de las estaciones vecinas para obtener una estimación del dato faltante en la estación objetivo, como el *método de la razón normal*, propuesto por Paulhus and Kohler (1952).

Estudios posteriores tratan de interpolar la precipitación utilizando distintos métodos, como el análisis por discrimante múltiple (Young 1992), redes neuronales (Coulibaly and Evora 2007), regresión (Lo Presti 2010), modelo espacio-temporal (Feng 2014, Carvalho 2016). En todos los casos nos hemos encontrado con métodos que rellenan datos diarios o mensuales, pero no se ha encontrado a ningún autor que trate de completar los datos diezminutales, posiblemente por la dificultad que entraña manejar la alta variabilidad que en esta escala temporal presenta la lluvia.

Por otro lado, los métodos de imputación anteriores se pueden considerar como de *imputación simple*, puesto que reemplazan cada dato faltante por un único dato estimado. Esto puede llevar a la pérdida de variabilidad en las muestras. Para solventar este problema, Rubin (1987) propuso el método de *imputación múltiple*. En este método, cada dato faltante se puede reemplazar por un conjunto de valores plausibles, dando lugar a varios conjuntos de datos completos.

Podemos encontrar varios métodos que implementan el procedimiento de imputación múltiple. Nosotros vamos a comparar los resultados obtenidos mediante el método de *missForest* (Steckhoven 2012), que está basado en el algoritmo *Random Forest* (Breiman 2001), con los obtenidos por el método *MICE* (van Buuren and Oudshoorn 1999), que es un método de imputación multivariable que usa ecuaciones encadenadas. Estos algoritmos están disponibles como bibliotecas en el paquete estadístico R.

2. ÁREA DE ESTUDIO Y DATOS

El área de estudio va a cubrir la comunidad autónoma de Galicia. En la Figura 1 podemos ver la distribución de la red de estaciones meteorológicas y agrometeorológicas de Meteogalicia.

Se han obtenido los datos de la variable **precipitación** de las 150 estaciones, desde el 1 de enero de 2000 hasta la actualidad, sumando un total de 74.367.306 registros.

Para este estudio concreto, hemos escogido como estación objetivo una estación del interior de la provincia de Lugo, la estación ID#10046, y como estaciones vecinas las estaciones ID#10045, ID#10050, ID#10055, ID#10088, ID#10092 y ID#10118, que rodean completamente a la primera y tienen datos suficientes para realizar las imputaciones. La Tabla 1 nos indica el nombre e identificación de cada estación, así como la fecha de alta de sus pluviómetros.

ID	Nombre	Ayuntamiento	Fecha de alta
10045	Mabegondo	Abegondo	18/01/2000
10046	**Marco da Curra**	**Monfero**	**03/02/2000**
10050	CIS Ferrol	Ferrol	01/01/2000
10055	Guitiriz	Guitiriz	01/01/2001
10088	Fragavella	Abadín	16/10/2003
10092	Punta Candieira	Cedeira	05/08/2004
10118	Burela	Burela	05/12/2005

Tabla 1: Estaciones meteorológicas usadas en el estudio.

En la Figura 2 se muestran la situación geográfica de estas estaciones:

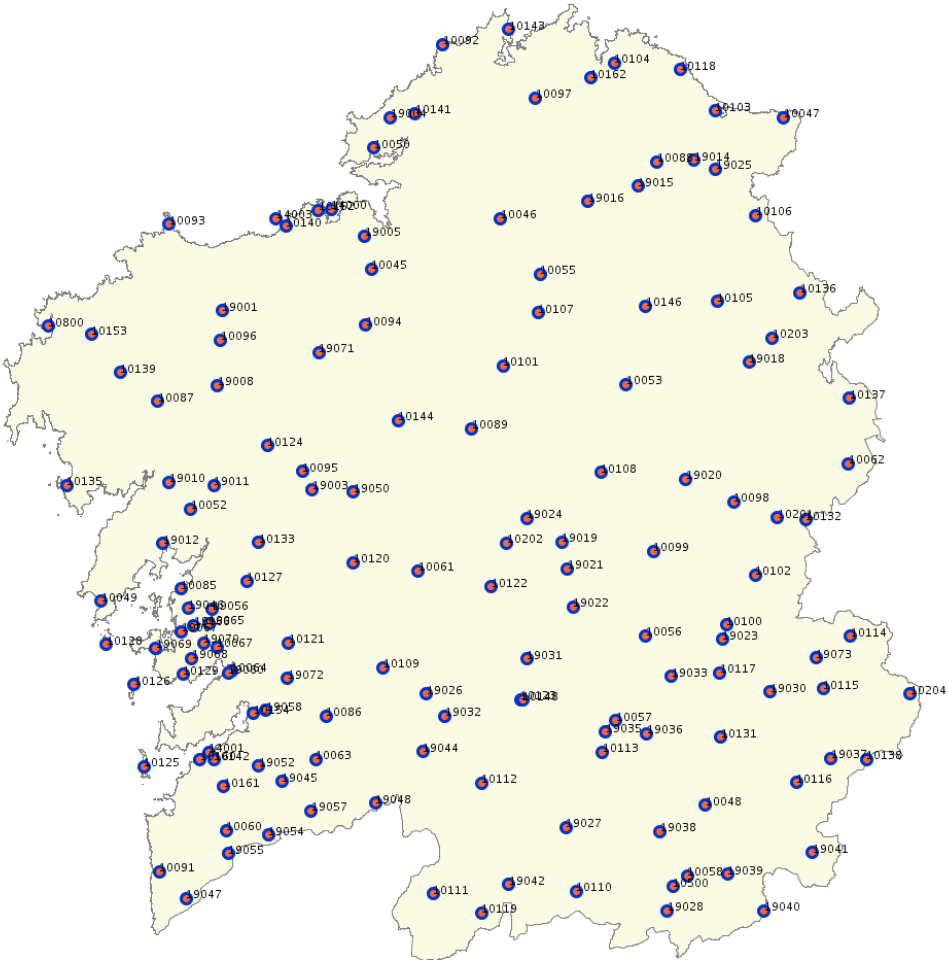


Figura 1: Red de estaciones de Meteogalicia.

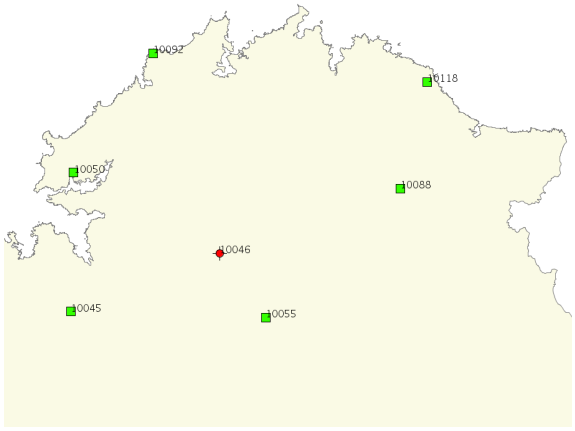


Figura 2: Estaciones objetivo y vecinas.

3. METODOLOGÍA

Vamos a realizar el proceso de imputación para cada mes concreto, dada la estacionalidad de la variable meteorológica **precipitación**, con valores más altos en los meses invierno y más bajos en los meses de verano (Eischeid et al. 2000). Por este motivo, los patrones de pérdida de datos de un mes/año concreto se aplicarán en años diferentes al mismo mes del calendario.

La metodología propuesta para realizar la imputación de los datos faltantes es la siguiente:

1. Seleccionar varias estaciones de Galicia que posean características diferentes, por ejemplo, una estación situada en la costa, otra en el interior, otra en altura en las montañas, ... A estas estaciones las llamaremos estaciones **objetivo**.
2. Para cada una de las estaciones objetivo, seleccionar varias estaciones vecinas por proximidad, intentando que éstas rodeen a la estación objetivo y que dispongan de un número de datos suficiente. A estas estaciones las llamaremos estaciones **candidatas**.
3. Para cada mes de todos los años de los que se disponga de registros:
   (a) Leer los datos diezminutales del mes correspondiente para la estación objetivo y las estaciones candidatas.
   (b) Si ese mes está completo para la estación objetivo, ese mes pasa a formar parte del conjunto de datos con los que podemos experimentar. Si ese mes está incompleto para la estación objetivo, debemos guardar su patrón de pérdida de datos, es decir, en qué posición están situados sus NAs (valor usado para indicar un dato faltante).
   (c) Una vez leídos los datos de ese mes para todos los años en los que disponemos de registros, tendremos dos conjuntos de datos: un conjunto de meses completos y un conjunto de meses con patrones de pérdida.
      i. A cada mes del conjunto completo para la estación objetivo le vamos a aplicar uno de los patrones de pérdida, generando así un mes con pérdidas, que llamaremos mes **experimental**.
      ii. Para ese mes experimental debemos elegir, de entre las estaciones candidatas para imputar, a aquellas que van a ser las **predictoras**, basándonos en los siguientes criterios:
         - Correlación con la estación objetivo.
         - Datos diezminutales completos para ese mes.
      iii. Imputaremos con las bibliotecas **missForest** y **MICE** del paquete estadístico R usando distintas opciones y distinto número de predictoras.
      iv. Calculamos la correlación cruzada entre la estación objetivo y cada una de las candidatas, y realizaremos un **desplazamiento temporal** de las series para mejorar la correlación.
      v. Repetiremos la imputación como se comentó en el paso (iii).
      vi. Realizaremos un **pre-procesamiento** al mes experimental rellenando los NAs con 0s cuando aparezcan ceros en todas las predictoras para esa posición concreta.
      vii. Volveremos a realizar la imputación como se comentó en el paso (iii).
      viii. Compararemos todos los conjuntos completos generados con el mes completo original de la estación objetivo, calculando el error absoluto medio (MAE), la raíz del error cuadrático medio ($RMSE$), el error relativo (RE), el coeficiente de determinacin (R^2) y el índice de concordancia de Willmot (d).
      ix. Repetimos el proceso completo desde el paso (i) para el siguiente patrón de pérdida.
4. Finalmente estudiaremos los resultados obtenidos.

4. RESULTADOS Y DISCUSIÓN

Como ya hemos comentado previamente, hemos escogido como estación objetivo una estación meteorológica del interior, la estación de Marco da Curra (Monfero). En primer lugar se ha llevado a cabo un control de calidad de los datos, identificando los datos erróneos y procediendo a su eliminación, dado que no contamos con la información necesaria para poder corregirlos. A continuación se han analizado los datos de esta estación, observando que tiene datos faltantes en los meses que indica la Figura 3. Como se puede observar, aunque esta estación dispone de datos de lluvia desde el 03/03/2000, el análisis se ha hecho desde el 01/01/2006; esto es debido a que la fecha de alta del pluviómetro de una de las estaciones candidatas es el 05/12/2005.

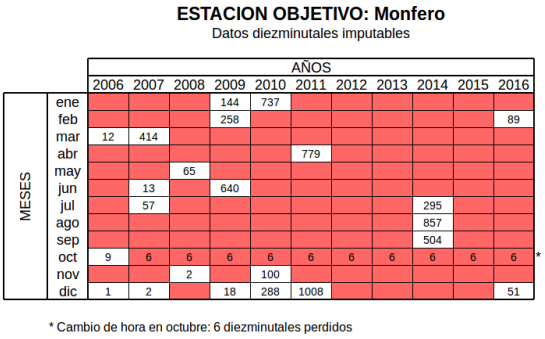


Figura 3: Datos faltantes en la estación de Marco da Curra.

La Figura 3 nos muestra que para el mes de enero nos faltan 144 datos en el año 2009 y 737 datos en el año 2010, así que procedemos a guardar estos dos patrones de pérdida, tal como muestran las Figuras 4 y 5.

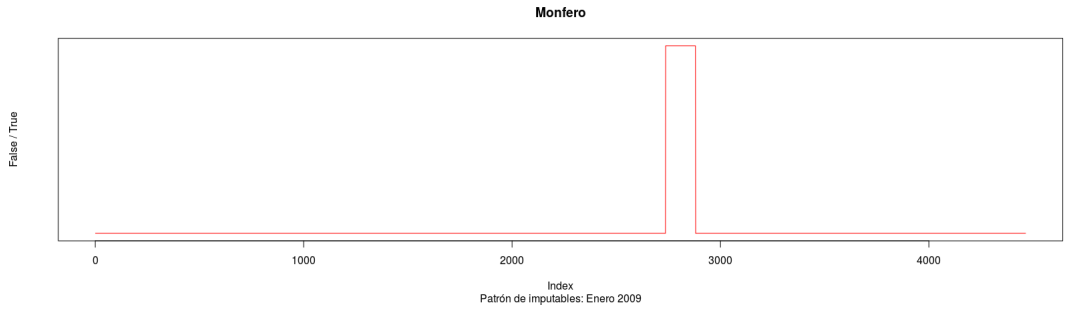


Figura 4: Patrón de pérdida del mes de enero de 2009.

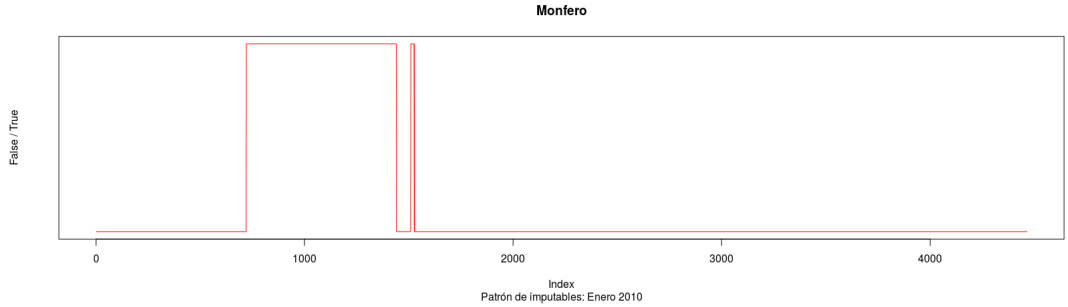


Figura 5: Patrón de pérdida del mes de enero de 2010.

A partir de estos patrones creamos los meses de enero experimentales. En la Figura 6 podemos ver estos patrones aplicados al mes de enero de 2008.

En el siguiente paso procedemos a realizar la imputación en estos conjuntos de datos experimentales, obteniendo los resultados que muestran las Figuras 7 y 8. En la parte superior de las Figuras se muestran las gráficas de los datos originales con el patrón de pérdida superpuesto, y las siguientes gráficas nos muestran los datos originales e imputados con missForest y con MICE.

Este paso de imputación lo repetimos eligiendo distinto número de estaciones predictoras entre las estaciones objetivo (en este caso se ha imputado con 1, 2, 4 y 6 predictoras, no encontrando diferencias significativas). Esta elección de predictoras se ha hecho basándonos en la correlación entre los datos de la estación objetivo y las candidatas.

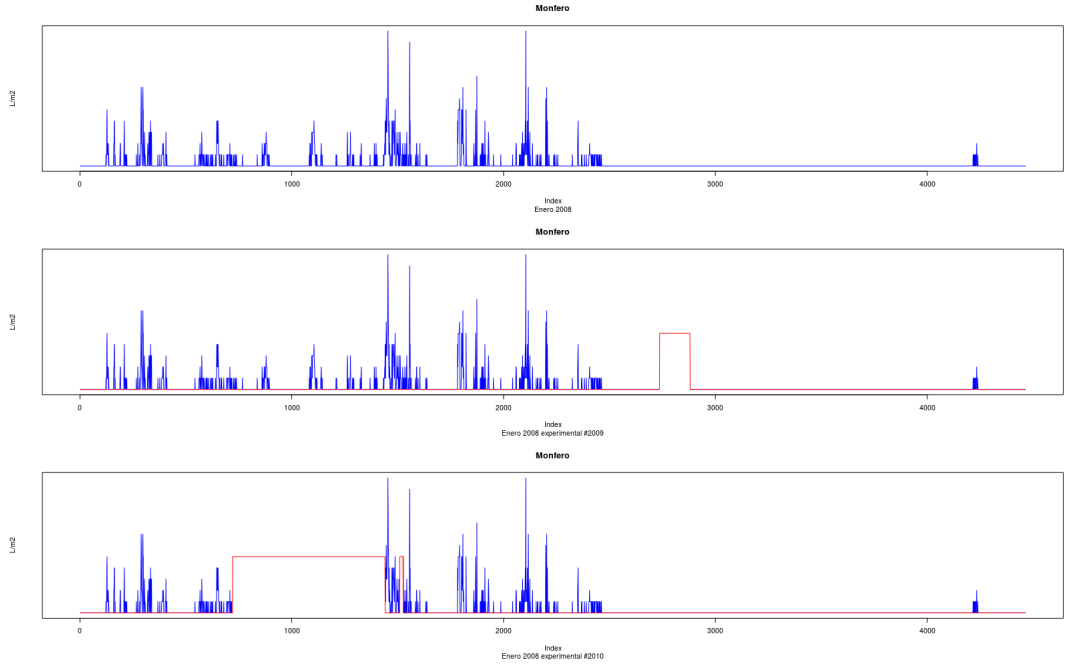


Figura 6: Meses experimentales para enero de 2008.

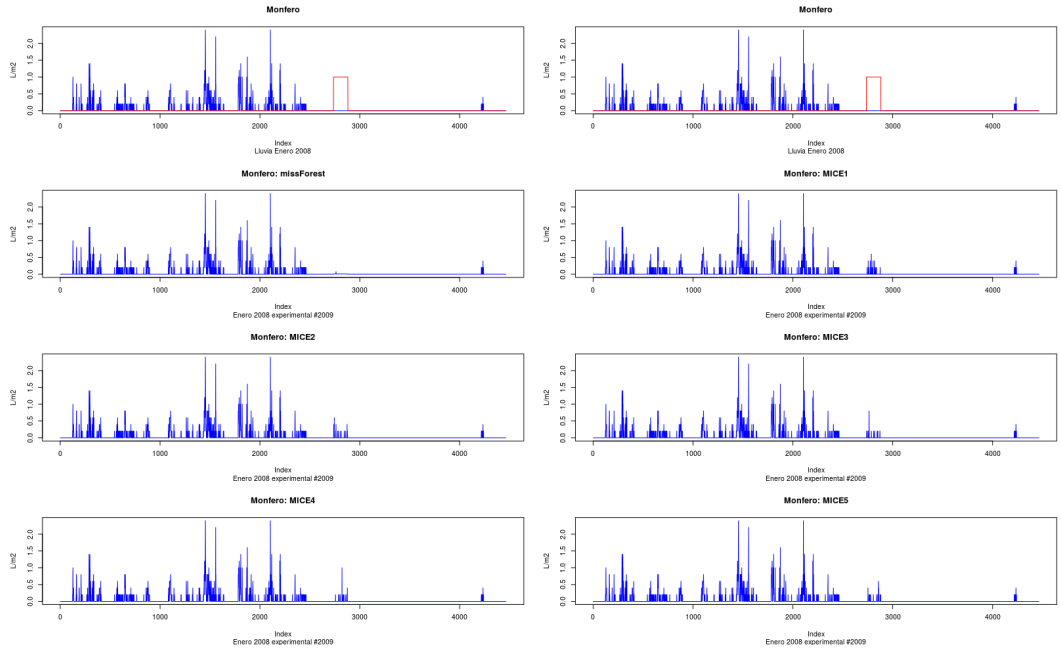


Figura 7: Imputación del mes experimental con missForest y MICE (patrón 2009).

En las Tablas 2 y 3 se indican los índices estadísticos obtenidos para el método de *missForest* y para cada conjunto del método de *MICE* en el mes de enero de 2008.

Teniendo en cuenta que la lluvia en Galicia se suele producir en su mayor medida como consecuencia del paso de frentes asociados a borrascas que se originan en el Océano Atlántico, y que traen consigo el desplazamiento de las nubes en sentido de Oeste a Este, las precipitaciones recogidas en estaciones separadas entre sí por varios Km. pueden estar desfasadas temporalmente. Bajo esta premisa se ha procedido a calcular la correlación cruzada entre la estación objetivo y las candidatas, para realizar un desplazamiento temporal de las series con el objetivo de buscar la

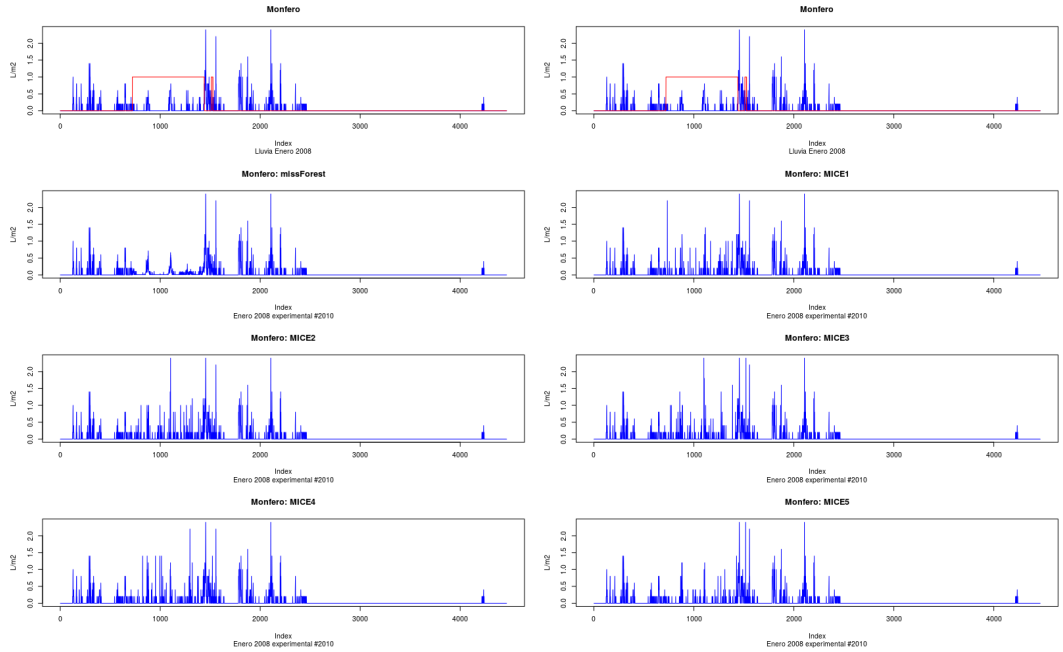


Figura 8: Imputación del mes experimental con missForest y MICE (patrón 2010).

Método	*N^o conjunto*	MAE	RMSE	RE	R^2	d
missForest		2,61e-04	1,83e-03	4,29e-02	1,00e+00	1,00e+00
MICE	N=1	7,17e-04	1,40e-02	3,28e-01	9,92e-01	9,98e-01
MICE	N=2	8,51e-04	1,67e-02	3,90e-01	9,89e-01	9,97e-01
MICE	N=3	6,27e-04	1,52e-02	3,57e-01	9,91e-01	9,98e-01
MICE	N=4	6,72e-04	1,82e-02	4,26e-01	9,87e-01	9,97e-01
MICE	N=5	4,48e-04	1,12e-02	2,62e-01	9,97e-01	9,99e-01

Tabla 2: Comparación entre métodos basada en cinco índices estadísticos (patrón 2009).

Método	*N^o conjunto*	MAE	RMSE	RE	R^2	d
missForest		8,87e-03	4,08e-02	9,55e-01	9,35e-01	9,83e-01
MICE	N=1	1,35e-02	8,75e-02	2,05e+00	7,46e-01	9,26e-01
MICE	N=2	1,38e-02	8,69e-02	2,03e+00	7,53e-01	9,29e-01
MICE	N=3	1,47e-02	9,80e-02	2,30e+00	7,04e-01	9,10e-01
MICE	N=4	1,28e-02	9,32e-02	2,18e+00	7,28e-01	9,19e-01
MICE	N=5	1,19e-02	7,94e-02	1,86e+00	7,81e-01	9,39e-01

Tabla 3: Comparación entre métodos basada en cinco índices estadísticos (patrón 2010).

mejor correlación entre las mismas antes de realizar la imputación.

En la Figura 9 se indica la correlación con las estaciones de Abegondo y Ferrol, y se puede comprobar que se mejora esa correlación si se realiza un desplazamiento en las estaciones predictoras. Una vez desplazadas las series se ha vuelto a realizar el proceso de imputación obteniendo una mejora en los resultados, tal como se puede ver en las Tablas 4 y 5.

Otra premisa que se ha considerado es la de sustituir los datos perdidos por ceros cuando en las estaciones vecinas no hay precipitación. La justificación para este procedimiento se basa en el hecho de que el valor más probable para esta variable meteorológica es cero. Basándonos en esta premisa, hemos realizado un pre-procesamiento de los datos sustituyendo los NAs por ceros en aquellos casos en los que todas las estaciones candidatas tienen un valor cero para esa misma

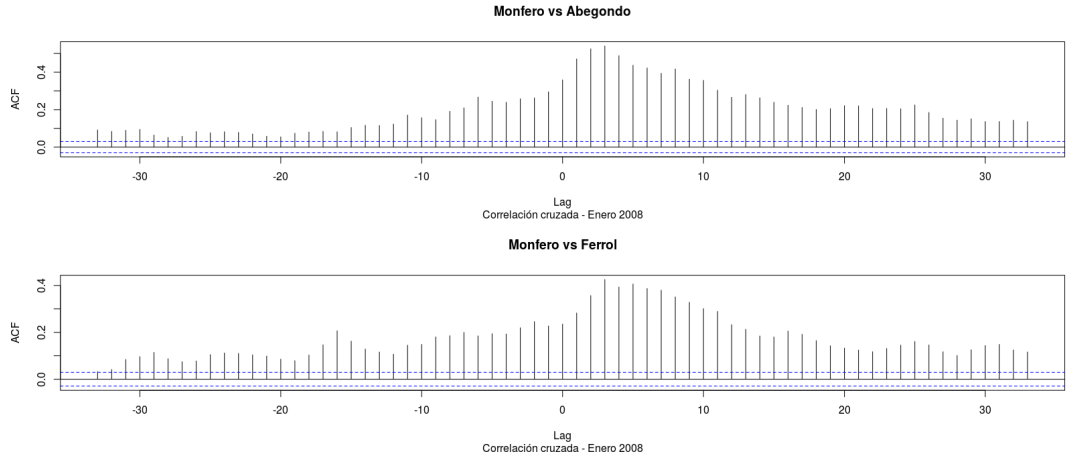


Figura 9: Correlación cruzada.

Método	*Nº conjunto*	MAE	RMSE	RE	R^2	d
missForest		2,06e-04	1,55e-03	3,62e-02	1,00e+00	1,00e+00
MICE	N=1	2,24e-04	1,23e-02	2,89e-01	9,94e-01	9,98e-01
MICE	N=2	3,58e-04	8,47e-03	1,98e-01	9,97e-01	9,99e-01
MICE	N=3	4,48e-04	2,99e-03	7,00e-02	1,00e+00	1,00e+00
MICE	N=4	1,34e-04	5,18e-03	1,21e-01	9,99e-01	1,00e+00
MICE	N=5	4,03e-04	9,93e-03	2,32e-01	1,00e+00	9,99e-01

Tabla 4: Comparación entre métodos después del desplazamiento (patrón 2009).

Método	*Nº conjunto*	MAE	RMSE	RE	R^2	d
missForest		7,64e-03	3,98e-02	9,31e-01	9,38e-01	9,84e-01
MICE	N=1	1,24e-02	7,98e-02	1,87e+00	7,91e-01	9,40e-01
MICE	N=2	9,99e-03	6,52e-02	1,53e+00	8,48e-01	9,59e-01
MICE	N=3	1,08e-02	7,06e-02	1,65e+00	8,24e-01	9,52e-01
MICE	N=4	1,03e-02	6,82e-02	1,60e+00	8,37e-01	9,55e-01
MICE	N=5	9,05e-03	6,19e-02	1,45e+00	8,61e-01	9,63e-01

Tabla 5: Comparación entre métodos después del desplazamiento (patrón 2010).

posición temporal del dato faltante, y a continuación se ha vuelto a imputar, pero no se mejoran los resultados.

Se verificó la similitud de las series temporales reconstruidas con las series originales. Se evaluó la bondad de ajuste entre la serie imputada y la serie observada mediante el coeficiente de determinación (R^2) y el índice de ajuste de Willmott (d). En todos los casos los valores del coeficiente de determinación de las series reconstruidas con missForest (1.000, 0.935, 1.000, 0.938) fueron superiores o cuando menos iguales a los obtenidos con MICE (0.997, 0.781, 1.000, 0.861). Lo mismo se puede observar para el índice de concordancia con missForest (1.000, 0.983, 1.000, 0.984) y con MICE (0.999, 0.939, 1.000, 0.963).

Como medidas de precisión se han calculado el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el error relativo (RE). Los resultados nos indican en todos los casos que missForest completó las series con un menor error que MICE.

A partir de las gráficas de las Figuras 7 y 8 podemos concluir que MICE tiende a sobreestimar los datos perdidos (no se ha calculado el sesgo). Esto se puede comprobar claramente con el primer patrón de pérdida, en el que los datos perdidos eran todos cero, aunque con el segundo patrón también se puede observar un aumento en la varianza de los cinco conjuntos de datos imputados.

Finalmente podemos observar que el desplazamiento de las series temporales previo a la imputación consigue mejorar los resultados, pero no así el pre-procesamiento con ceros. Esto confirma que existe un desfase de la lluvia entre estaciones separadas geográficamente como consecuencia del desplazamiento de las nubes asociadas a los frentes meteorológicos. Ese desfase puede ser un adelanto o un retraso en función de la situación de las estaciones y del movimiento de la lluvia.

5. CONCLUSIONES

Como se ha comentado previamente, las series de datos diezminutales de precipitación presentan con frecuencia interrupciones de registro que suponen un problema para su utilización en algunas aplicaciones de ingeniería civil e hidrología. Como solución se propone la búsqueda de métodos de imputación que mantengan las características de la serie en ausencia de pérdidas.

En los casos en los que se producen grandes bloques de datos perdidos consecutivos, una vía de imputación es recurrir a los datos de las estaciones más próximas aplicando algún método de imputación múltiple. En este artículo se han valorado dos métodos de imputación múltiple disponibles como bibliotecas en el paquete estadístico R, los algoritmos MICE y missForest.

De los resultados obtenidos podemos deducir que missForest parece más apropiado que MICE para la imputación de datos diezminutales de lluvia, aunque se recomienda realizar el estudio para estaciones situadas en otras áreas geográficas (costa, montaña, ...) y comparar los resultados con otros métodos de imputación.

REFERENCIAS

Breiman, L. (2001) Random forests. Machine Learning, 45 (1), p. 5–32

Carvalho, J.R.P.; Nakai, A.; Monteiro, J.E.B.A. (2016) Spatio-temporal modeling of data imputation for daily rainfall series in homogeneous zones. Revista brasileira de meteorología, 31 (2), p. 196–201

Coulibaly, P.; Evora, N.D. (2007) Comparison of neural network methods for infilling missing daily weather records. Journal of hydrology, 341 (1–2), p. 27–41

Eisched, J.K.; Pasteris, P.A.; Diaz, H.F.; Plantico, M.S.; Lott, N.J. (2000) Creating a serially complete, national daily time series of temperature and precipitation for the western United States. Journal of Applied Meteorology, 39 (9), p. 1580–1591

Feng, L.; Nowak, G.; O'Neill, T.J.; Welsh, A.H. (2014) CUTOFF: A spatio-temporal imputation method. Journal of hydrology, 519, p. 3591–3605

Lo Presti, R.; Barca, E.; Passarella, G. (2010) A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). Enviro. Monit. Assess, 160 (1–4), p. 1–22

Paulhus, J.L.H.; Kohler, M.A. (1952) Interpolation of missing precipitation records: U.S. Weather Bur. Monthly Weather Rev., 80, p. 129–133

Stekhoven, D.; Buhlmann, P. (2012) Missforest-non-parametric missing value imputation for mixed-type data. Bioinformatics, 28 (1), p. 112–118

Young, K.C. (1992), A 3-way model for interpolating for monthly precipitation values. Monthly Weather Review, 120 (11), p. 2561–2569.