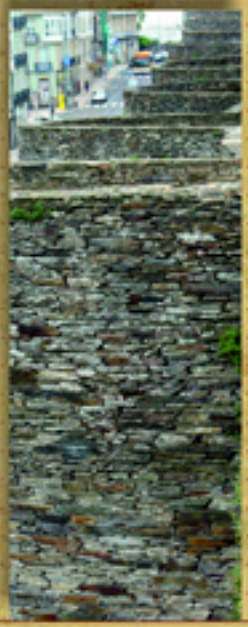
XII CONGRESO GALEGO DE ESTATÍSTICA E INVESTIGACIÓN DE OPERACIÓNS

22, 23 e 24 de OUTUBRO 2015




http://xiicongreso.sgapeio.es

ACTAS

XII CONGRESO GALEGO DE ESTATÍSTICA E INVESTIGACIÓN DE OPERACIÓNS

Lugo, 22-23-24 de OUTUBRO de 2015

ACTAS

XII Congreso Galego de

Estatística e Investigación de Operacións

Lugo, 22-23-24 de outubro de 2015

ACTAS

XII Congreso Galego de Estatística e Investigación de Operacións

Lugo, 22-23-24 de outubro de 2015

Ginzo Villamayor, M.J.; Alonso Meijide, J.M. e Ramil Novo, L.A. (editores científicos)

Organiza

Departamento de Estatística e Investigación Operativa

Universidade de Santiago de Compostela e

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

Edita

SGAPEIO e Servizo de Publicacións da Deputación de Lugo

secretaria@sgapeio.es, publicacions@deputacionlugo.org

Primeira edición: outubro de 2015

I.S.B.N.: 978-84-8192-522-7

Depósito Legal: LU 187-2015

DEPUTACIÓN DE LUGO

XII Congreso Galego de Estatístico a e Investigación de Operacións

2015. Lugo 22-24 de outubro: Libro de actas /[equipo editorial, Ginzo Villamayor, María José; Alonso Meijide, José Mª y Ramil Novo, Luis Alberto (editores científicos)]. – Lugo: Universidade de Santiago de Compostela, Editor: Servizo de Publicacións da Deputación de Lugo e SGAPEIO - SGAPEIO - Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións.
424 p..
D.L.: LU 187-2015. – ISBN: 978-84-8192-522-7

Equipo editorial:
(Editores científicos)

María José Ginzo Villamayor,
mariajose.ginzo@usc.es
Departamento de Estatística e Investigación Operativa, USC

José María Alonso Meijide,
josemaria.alonso@usc.es
Departamento de Estatística e Investigación Operativa, USC

Luis Alberto Ramil Novo,
l.ramil@usc.es
Departamento de Estatística e Investigación Operativa, USC

© Servizo de Publicacións da Deputación Provincial de Lugo e SGAPEIO – Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións
ISBN: 978-84-8192-522-7
Depósito legal: LU 187-2015

Comité Organizador:

PRESIDENTE
José Manuel Colmenero Álvarez.
Instituto Nacional de Estadística, Delegación Provincial de Lugo.

SECRETARIO
Luis Alberto Ramil Novo.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela, Campus de Lugo.

VOCAIS
José Mª Alonso Meijide.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela, Campus de Lugo.

Nieves Muñoz Ferreiro.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela, Campus de Lugo.

Antonio Sampayo Flores.
Xunta de Galicia, Consellería de Traballo e Benestar, Xefatura Territorial de Lugo.

Balbina Casas Méndez.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela, Campus Vida.

María José Ginzo Villamayor.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela, Servizo de Consultoría Estatística, Campus Vida.

Mª Esther López Vizcaíno.
Instituto Galego de Estatística. Presidenta da Sociedade Galega de Estatística e Investigación de Operacións.

Comité Científico:

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

Rubén Fernández Casal.
Departamento de Matemáticas, área de Estatística e Investigación Operativa, Universidade da Coruña.

Mª Gloria Fiestras Janeiro.
Departamento de Estatística e Investigación Operativa, Universidade de Vigo.

Ignacio García Jurado.
Departamento de Matemáticas, área de Estatística e Investigación Operativa, Universidade da Coruña.

Julio González Díaz.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela.

Wenceslao González Manteiga.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela.

Manuel Antonio Presedo Quindimil.
Departamento de Matemáticas, área de Estatística e Investigación Operativa, Universidade da Coruña

César Andrés Sánchez Sellero.
Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela.

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

Benvida ao XII CONGRESO DA SGAPEIO:

Como Presidentes da SGAPEO e do Comité Organizador do XII Congreso Galego de Estatística e Investigación de Operacións, é un pracer para nós darvos a benvida a Lugo para a celebración deste Congreso, onde os estatísticos galegos poderemos compartir experiencias con colegas que veñen do resto de España, do noso país veciño Portugal e de América Latina. É a segunda vez que se celebra en Lugo, despois do III Congreso de 1997, hai xa 18 anos. Nesta ocasión foi posible pola elección do Campus de Lugo polo Departamento de Estatística e Investigación Operativa da USC, ao que a SGAPEIO encargou a organización deste Congreso, e a colaboración da Deputación de Lugo, ao cedernos o seu magnífico Pazo de San Marcos como sede do mesmo.

A SGAPEIO é unha sociedade independente e sen ánimo de lucro, que ten por obxectivo promover a estatística e a investigación operativa en Galicia desde diferentes puntos de vista: na investigación, no ensino, na empresa e nas Administracións Públicas. Para iso a sociedade difunde a súa información a través do seu boletín trimestral INFORMEST, da páxina web e das noticias semanais e organiza cursos, congresos e premios. Con respecto aos premios, hai que resaltar que neste Congreso entregaranse dous Premios ás mellores comunicacións presentadas por investigadores novos, de acordo co obxectivo da SGAPEIO de promover a investigación no campo da estatística e a investigación de operacións.

Nesta ocasión, ademais das conferencias, cursos, mesas redondas e talleres, así como do importante número de comunicacións orais e en póster, programáronse algunhas actividades novidosas para un Congreso destas características, como a exposición itinerante sobre estatística "Explorística", en colaboración coa Sociedade Portuguesa de Estatística, os cursos online previos á sesión final dos mesmos no propio Congreso ou o panel de cata de viños e a consecuente análise estatística dos resultados.

Como nos anteriores Congresos, o programa científico compleméntase cun programa de actividades sociais que nesta ocasión inclúe unha visita á Ribeira Sacra, contribuíndo ao coñecemento deste fermoso enclave e á promoción da súa declaración como Patrimonio da Humanidade, a visita guiada aos importantes restos arqueolóxicos da Domus do Mitreo e un concerto que, baixo o título de "Resonancias Aritméticas", foi programado especificamente para este Congreso.

Finalmente, queremos expresar o noso profundo agradecemento a todas as institucións e empresas que patrocinaron este Congreso, así como tamén aos demais colaboradores e a todos os congresistas pola súa asistencia.

Desde a SGAPEIO e desde o comité organizador, desexamos que este encontro sexa frutífero para contribuír á difusión e a promoción da Estatística e a Investigación de Operacións como ferramenta para un mellor coñecemento da realidade que nos rodea, e ao mesmo tempo para fortalecer os lazos dos que participamos no Congreso.

Esther López Vizcaíno
Presidenta da SGAPEIO

José M. Colmenero Álvarez
Presidente do Comité Organizador

Patrocinadores:



DEPUTACIÓN DE LUGO



Concello de Lugo







Colaboran:

SAGAZH

Prelo

RIBEIRA SACRA
Consello Regulador Denominación de Orixe

Rúa do Comercio 6-8, 27400 Monforte de Lemos - Lugo - España
Telf.: +34 982 41 09 68 Fax.: +34 982 41 12 65
info@ribeirasacra.org www.ribeirasacra.org

Ribeira Sacra
CONSORCIO DE TURISMO

BODEGAS REGINA VIARUM
RV

Eduardo Ochoa, fotógrafo

Agradecementos:

A celebración do XII Congreso Galego de Estatística e Investigación de Operacións foi posible grazas as seguintes institucións e empresas patrocinadoras,

- DEPUTACIÓN DE LUGO
- CONCELLO DE LUGO
- CETAL
- TRAGSA
- MODESTYA

O Comité Organizador quere facer constar o seu especial agradecemento a todas elas. O seu apoio foi fundamental para poder levar a cabo as diversas actividades do Congreso.

O Comité Organizador quere tamén facer constar a súa gratitude ás seguintes institucións, empresas e persoas colaboradoras,

- SAGATH
- PRELO
- CONSELLO REGULADOR DA DENOMINACIÓN DE ORIXE RIBEIRA SACRA
- CONSORCIO DE TURISMO DA RIBEIRA SACRA
- BODEGA REGINA VIARUM
- Eduardo Ochoa, FOTÓGRAFÍA
- Pilar Cordeiro, SERVIZOS DE TRADUCCIÓN

O Comité Organizador desexa ademais agradecer o apoio prestado para a cata de viños do Congreso ao Consello Regulador da D.O. Valdeorras e a bodega José Antonio García de Valtuille de Abajo (León). Agradece tamén a colaboración prestada polas seguintes persoas participantes no panel de cata: Luis Buitrón Amaro, Elisabet Castro Vázquez, Francisco Costas Iglesias, Juan Ángel Fernández Martínez, Juan Bosco Ferreiro Darriba, Óscar González Rizo, Mercedes González Rodríguez, Bruno Lovelle Fernández, Juan M. Mesas Mesas, Ana Paula Losada García, Alberto Pérez Ferreiro, Mayra Ratón Fernández, Daniel Romay Piñeiro e Xesús Verao Pérez. O agradecemento faise extensivo ás seguintes persoas polo seu apoio técnico na cata, no deseño e no tratamento dos datos: Mª Belén García Gómez, María del Mar López García e Raquel Lorences Touzón

Para finalizar, o noso agradecemento ás persoas que colaboraron na exposición Explorística no Congreso, Ana Bouzas Lorenzo (coordinadora) e Noelia Bazarra García, María Calvo Torres e Iria Cañete Quián.

Índice:

Conferencias plenarias

Optimización bajo incertidumbre: un vistazo a la programación estocástica.
González Díaz, J. 2

La estadística en una empresa cervecera. Del dato al conocimiento.
López Múñiz, J. 3

Extracting informative features in functional datasets: statistical methods with applications in spectrometry.
Vieu, P. 4

La sensometría, la estadística del gusto.
Bécue Bertaut, M. 5

Comunicacións orais

Sesión 1 de Investigación de Operacións

Otimização do pessoal técnico do serviço de manutenção da TAP.
Santos, J. e Cavique, L. 7

Optimización del proceso de llenado de placas PCR en secuenciación Sanger.
Carpente Rodríguez., M.L.; Cerdeira Pena, A.; Lorenzo Freire, S., y Places, A.S. 15

On the set of extreme core allocations for minimum cost spanning tree problems.
Trudeau C. and Vidal Puga, J. 21

A monotonic and merge-proof rule in minimum cost spanning tree situations.
Gómez Rúa, M. and Vidal Puga, J. 33

Comunicacións orais

Sesión 1 de Estatística

Aspetos geométricos na escolha da janela para o estimador no núcleo.
Oliveira, P.E. 40

A survey on statistical choice of extreme value models.
Neves Figueiredo, M.M. 41

Diagnóstico de modelos de clases latentes en tablas con frecuencias pequeñas o nulas: efecto de sumar a las frecuencias de los patrones de respuestas.
Araya Alpízar, C. 48

Nonparametric incidence and latency estimation in mixture cure models.

López Cheda, A.; Cao Abad, R.; Jácome Pumar, M.A. and Van Keilegom, I. 51

Bandwidth selection for the presmoothed logrank test.
Jácome Pumar, M.A. and López de Ullibarri, I..... 63

Comunicacións orais
Sesión de Ensinanzas Medias

Simulación.
Rodríguez-Gigirey Pérez, M.A..... 75

LOMCE, hacia un cambio de estrategias en el aula.
Casado Barrio, M.J. 78

Carreiras populares, aprendizaxe estatística e valores sociais.
Pujales Martínez, X.E..... 84

Propostas de traballos estatísticos para as aulas de secundaria.
Rodríguez-Moldes Rey, C...... 90

Comunicacións orais
Sesión de novos investigadores - premios SGAPEIO

Problemas de asignación en extinción de incendios forestales: Una aplicación al control de los recursos aéreos.
Rodríguez Veiga, J.; Gómez Costa, I.; Ginzo Villamayor, MJ. y Casas Méndez, B..... 98

Programación lineal sucesiva sin región de confianza.
González Rueda, A.; González Díaz, J. y Pérez Fernández de Córdoba, M. 103

Optimización de la gestión forestal de rodales regulares.
Arias Rodil, M.; Diéguez Aranda, U. y Vázquez Méndez, M.E. 114

Functional prediction with applications to the Electricity Market.
Raña Míguez, P.; Vilar Fernández, J. and Aneiros Pérez, G. 121

ARTICLERANK: una nueva forma de ordenar artículos.
Ameijeiras Alonso, J.; Crujeiras Casais, R.M. y Rodríguez Casal, A. 123

Predicting using quantile regression methods.
Conde Amboage, M.; González Manteiga, W. and Sánchez Sellero, C. 135

Comunicacións orais
Sesión 2 de Investigación de Operacións

Strategy-proofness and monotonicity for land rental agreements.
Valencia Toledo, A. and Vidal Puga, J. 148

Coalition formation in a multi-agent inventory problem with two acquisition costs.
Saavedra Nieves, A.; García Jurado, I. and Fiestras Janeiro, M.G. 160

Computational procedure for a wide family of mixed coalitional values.
Giménez, J.M. and Puente, M.A. 161

Implementación de funciones generatrices para la obtención de índices con configuración de coaliciones.
Rodríguez Veiga, J.; Novoa Flores, G.I. y Casas Méndez, B. 170

Juegos cooperativos e índices de cohesión: El valor de Choquet-Owen.
Fiestras Janeiro, M.G; Gallardo, J.M.; Jiménez Losada, A. y Mosquera Rodríguez, M.A..... 174

Comunicacións orais

Sesión 2 de Estatística

Un estudio comparativo de la estimación noparamétrica de la función de intensidad en procesos puntuales sin marcas.
Borrajo García, M.I.; González Manteiga, W. y Martínez Miranda, M.D. 176

Inferencia sobre el variograma mediante técnicas bootstrap en procesos espaciales estacionarios y no estacionarios.
Castillo Páez, S.; Fernández Casal, R. y García Soidán, P. 182

Bootstrap ponderado para la bondad de ajuste del error en modelos de regresión no paramétrica.
Rivas Martínez, G.y Jiménez Gamero, M.D. 186

Tests for the equality of conditional variance functions in nonparametric regression.
Pardo Fernández, J.C.; Jiménez Gamero, M.D. and El Ghouch, A. 191

Comunicacións orais

Sesión 3 de Estatística

Librería TTS en R
Meneses Freire, A.; Naya Fernández, S. y Tarrío Saavedra, J. 197

QCR: propuesta de una nueva librería en R para el control de calidad.
Naya Fernández, S.; Flores, M.; Fernández Casal, R. y Tarrío Saavedra, J. 201

Propuesta de un método para el control de calidad de la eficiencia energética basado en el estudio de datos funcionales.
Oviedo de la Fuente, M.; Tarrío Saavedra, J.; Zaragoza, S.; Febrero Bande, M. y Naya Fernández, S. 206

Clasificación de fondos marinos a partir de datos acústico.
Tarrío Saavedra, J.; Sánchez Carnero, N. y Prieto Aneiros, A. 211

Comunicacións orais

Sesión 4 de Estatística

Un potente aroma de estadística
Bécue Bertaut, M. y Sánchez Espigares, J.A. 219

Variación de la densidad de la madera a lo largo del fuste en Eucalyptus globulus. Influencia de los incendios forestales.
Riesco Muñoz, G. 220

Avaliação do Desempenho Estatístico de um Novo Método Combinado para obter os Instantes.
Coelho do Carmo, M.A.; Infante, P. e Mendes, J.M. 230

Modelos mixtos de Poisson: aplicación a datos de incendios forestales en Galicia.
Boubeta Martínez, M., Lombardía Cortiña, M.J. Marey Pérez, M.F. y Morales, D. 232

Comunicacións orais

Sesión 5 de Estatística

Construcción de un indicador sintético de coyuntura económica en Galicia.
López Vizcaíno, M.E.; Sánchez Fernández P. y Iglesias Patiño, C. 234

Indicador Mensal de Actividade Económica da Região Autónoma dos Açores (IAE – Açores.)
Monjardino, I.C. 242

Estatística, turismo e sustentabilidade um estudo de caso aplicado a uma região portuguesa.
Oliveira, E. e Pires Manso, J.R. 247

Jultilocalización e estatisticas rexionais.
Santiso Blanco, J. 260

Modelos temporales lineales mixtos: Aplicación a datos de la encuesta de condiciones de vida.
Esteban, M.D.; Morales, D. y Pérez, A. 261

Estimation of employed people by economic activity in Galicia.
Lombardía Cortiña, M.J.; López Vizcaíno, M.E. and Rueda, C. 262

Comunicacións orais

Sesión 6 de Estatística

Impacto do programa galego de detección precoz do cancro de mama (PGDPCM) sobre a mortalidade por cancro de mama en Galicia. Período 1980-2012.
Naveira Barbeito, G.; Pollán Santamaría, M.; Santiago Pérez, M.I.; Fernández Llanes, A.B.; Zubizarreta Aberdi, R. e Malvar Pintos, A. 266

Modelo logístico multinomial: identificação de factores associados à mudança na adesão à terapêutica anti-hipertensora.
Mouriño, H.; Rosa, S. e Fernandes, M. 273

Restricting the weights and the judgment values in dea models.
Alberca, P. and Santos, J. 275

A method for processing perceptual dialectology data.
Calaza Díaz, L.; Suárez Quintas, S.; Crujeiras Casais, R.M.; Rodríguez Casal, A.; Sousa Fernández, X. e Ríos Viqueira, J.R. 282

Modelización onomástica.
Ginzo Villamayor, M.J. e Crujeiras Casais, R.M. 292

Comunicacións en forma de póster

Os titulados universitarios no Sistema Universitario de Galicia (SUG) traballan en Galicia nos dous primeiros anos posteriores á súa titulación?
Fernández Montes, M.C.; López Pereira, J.E.; Morrondo Pelayo, M.P.; Faraldo Roca, P.; Crujeiras Casais, R.M.; Martínez Calvo, A.; Lombardía Cortiña, M.J.; Naya Fernández, S., Matías Fernández, J.M.; Carreira Figueiras, V. e Fernández de Castro, B.M. 301

A inserción laboral dos titulados en Másteres no Sistema Universitario de Galicia 2007-2008, 2008-2009 e 2009-2010.
Fernández Montes, M.C.; López Pereira, J.E.; Morrondo Pelayo, M.P.; de Miguel Domínguez, J.C.; Rodríguez González, D.; Valcárcel Aguiar, B.; Vecino Aguirre, A.; Pereira Andrade, L.P. e Guisado González, M. 306

Análisis de la visión de la prensa viguesa sobre la exclusión social en la primera fase de la crisis (2009-2010).
Gómez López, C. y García Soidán, P. 309

La internacionalización como factor de supervivencia de las empresas del área de Vigo.
Meijide Vecino, M.; Vaamonde Liste, A.; Luaces Pazos, R. y Sánchez Bello, P. 319

Metodologías para la autoevaluación, gestión y análisis del perfil de riesgo basado en ORSA.
Segura Gisbert, J. y Muñoz Murgui, F. 324

Reducción de la demora diagnóstica tras la introducción de la vía rápida de acceso a la colonoscopia.
López Calviño, B.; Pita Fernández, S.; González Santamaría, P.; Seijo Bestilleiro, R.; Pértega Díaz, S.; y Seoane Pillado, T. 326

Demora diagnóstica y pronóstico en pacientes con cáncer colorrectal.
Pértega Díaz, S.; Pita Fernández, S.; González Saez, L.; López Calviño, B.; Seoane Pillado, T.; Rodríguez Camacho, E. y González Santamaría, P. 330

Desarrollo de un score de predicción de mortalidad en pacientes VIH con neumonía por Pneumocystis *JIROVECII.*
Rodríguez Osorio, I.; Mena, A.; Pértega, S.; Meijide, H.; Castro Iglesias, A.; Pedreira, J.; Poveda, E.; López Calviño, B. y Seoane Pillado, T. 332

Nomograma y validez de un modelo para predecir desnutrición en pacientes en lista de trasplante hepático.
López Calviño, B.; García Rodríguez, M.T.; Pita Fernández, S.; Pértega Díaz, S.; Piñón Villar, M.C. y Seoane Pillado, T. 334

Modelos de riesgos competitivos para determinar factores pronósticos asociados a la mortalidad por infección en pacientes trasplantados renales.
López Calviño, B.; Pérez Fontán, M.; Pita Fernández, S.; Pértega Díaz, S. y Seoane Pillado, T. 338

CVRCALC: a cardiovascular risk calculator.
Seoane Pillado, T.; Pita Fernández, S.; Hervada Vidal, X.; Pértega Díaz, S.; López Calviño, B y *Rodríguez Muíños, M.A.* 341

Saúde mental em tempos de crise: análise dos cuidados prestados no agrupamento de centros de saúde arco ribeirinho, área metropolitana de Lisboa.
Tavares, D.; Mouriño, H.; Durval, M. e Carvalho das Neves, J. 343

Modelización de la degradación de smart biomateriales para uso odontológico.
Robles Bykbaev, Y.; Naya Fernández, S. y Tarrío Saavedra, J. 345

On the Hill estimator: a comparison of methods.

Ferreira, M. and Rebelo, M. 350

Weighted estimation of conditional mean function with truncated, censored and dependent data.
Liang, H.Y. and Iglesias Pérez, M.C. 356

A comparative simulation study about the model checks for nonparametric regression with missing data.
Cotos Yáñez, T.R.; Pérez González, A. and González Manteiga, W. 364

Análisis estadístico de la evaluación sensorial de un medicamento mediante un panel de catadores.
García Quiroga, M.; Muñoz Ferreiro, M.N.; Romero Rodríguez, M.A. y Vázquez Ordériz, M.L. 371

Estadísticas sobre productos gallegos con dop e igp: conocimiento, consumo y percepción.
Vázquez Pita, M.; García Luque, J.; Muñoz Ferreiro, M.N.; Romero Rodríguez, M.A. y Vázquez Ordériz, M.L. 372

Encuesta sobre compra y consumo de productos pesqueros frescos y congelados en Lugo: análisis de resultados mediante métodos estadísticos exactos.
López García, M.M.; Ramil Novo, L.A.; Romero Rodríguez, M.A. y Vázquez Ordériz, M.L. 373

Aplicación del análisis factorial múltiple para la estimación del riesgo de incendios en Galicia.
Maestro Requena, M.G. y Marey Pérez, M.F. 374

Spatial pattern analysis of wildfires in Galicia.
Marey Pérez, M.F.; González Manteiga, W; Saavedra Nieves, P. and Fuentes Santos, I. 382

Pesquisa e validação do mpp global na análise da fiabilidade estrutural dos compósitos.
Hoffbauer, L.N. e António, C.C. 384

Diseño de una técnica de monitorización.
Vázquez Méndez, M.E.; Alvarez Vázquez, L.J.; Casal, G. y Martínez, A. 392

Study and characterization of the multinomial values.
Domènech, M. Giménez, J.M. and Puente, M.A. 398

El valor de Shapley para Modelos Cooperativos Bi objetivo.
Meca, A.; Kimms, A. y Kozeletskyi, I. 400

Conferencias plenarias

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

OPTIMIZACIÓN BAJO INCERTIDUMBRE: UN VISTAZO A LA PROGRAMACIÓN ESTOCÁSTICA

Julio González Díaz[1]

[1] Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela

RESUMEN

Prácticamente todos los problemas de optimización que surgen en el mundo real están sujetos a algún tipo de incertidumbre. En muchas ocasiones es razonable reemplazar los elementos estocásticos por aproximaciones deterministas, ya sea porque dicha incertidumbre es pequeña o porque afecta a parámetros no fundamentales del modelo. Sin embargo, muchas otras veces la resolución efectiva del problema pasa por trabajar con modelos de optimización que tengan en cuenta la incertidumbre de modo explícito.

Es precisamente en este contexto en el que surge la programación estocástica, y en esta charla llevaremos a cabo una introducción informal a los mismos. Partiremos de una serie de ejemplos ilustrativos de las diferencias entre modelos deterministas y estocásticos, que permitirán obtener una primera impresión de los nuevos retos que supone la optimización bajo incertidumbre, tanto teóricos como prácticos.

El núcleo de la presentación consistirá en recorrer las distintas capas en las que hay que trabajar de cara a una resolución exitosa de este tipo de problemas: comprensión de la versión determinista del problema y de las técnicas existentes para su resolución, modelado de la incertidumbre que rodea la versión estocástica, inclusión de medidas de riesgo en la función objetivo, desarrollo de metodologías para resolver el problema estocástico y estudio de la estabilidad de las mismas.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

LA ESTADÍSTICA EN UNA EMPRESA CERVECERA. DEL DATO AL CONOCIMIENTO

Jorge López Muñiz[1]

[1] Estrella Galicia, Hijos de Rivera, S.A.U.

RESUMEN

"Data Scientist": La profesión más sexi del siglo XXI según la revista Harvard Business Review.
El área de Inteligencia de Negocio se ha venido desarrollando en Estrella Galicia durante la última década, facilitando la toma de decisiones a la dirección de la compañía y apoyando el crecimiento de una empresa que ha triplicado su tamaño en este período.
En esta ponencia se tratará de compartir esta experiencia apasionante vivida en primera persona: cuál era la situación de partida, proyectos abordados, las dificultades encontradas en el camino y cómo se ha llegado al punto actual. Se presentarán diferentes actividades realizadas relacionadas con las ciencias y técnicas estadísticas desde una perspectiva global y de Negocio, siempre con el foco en la aportación de conocimiento a la organización.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

EXTRACTING INFORMATIVE FEATURES IN FUNCTIONAL DATASETS: STATISTICAL METHODS WITH APPLICATIONS IN SPECTROMETRY

Philippe Vieu1[1]

[1]Institut de Mathmatiques, Toulouse, France

ABSTRACT

In this talk we will look at regression problem with continuous covariates χ and scalar response Y. The main goal of the talk is to emphasize on exploratory questions through the search of what are the specific features in the functional covariate χ which are acting on the response Y. When the covariate is a function

$$\chi = \chi(t), t \in [a, b]$$

this question may take various forms, like:

- Which derivative(s) of χ is(are) the most informative for explaining Y?
- Is there some subinterval $[c, d] \subset [a, b]$ containing all informations for predicting Y?
- Can the curve χ be summarized by a small number of discretized values?

The presentation will be mainly methodological with the presentation of various approaches which have been developed in the last few years in Functional Data Analysis. The methodologies will be illustrated by means of various real datasets coming from spectrometrics in which χ will be the absorbance channel of some element (meat, orange juice, sugar, ...) as function of the spectrum wavelengths and in which Y will be some chemical componant of the element (fatness, moisture, ...) to be predicted from the spectrum. Asymptotics issues/results will only be briefly mentionned.

The methods to be presented will be linked with various topics of Statistics being of great current actuality (nonparametric statistics, semiparametric statistics, goodness of fit test, variable selection, ...). The main general point that will be supported in this talk is the fact that most of the modern statistical techniques (previously introduced for finite dimensional problems) may have (pending to suitable functional adaptations ...) great interest in Functional Data Analysis.

LA SENSOMETRÍA, LA ESTADÍSTICA DEL GUSTO

Mónica Bécue Bertaut[1]

[1] Departamente Estadística i Invetigació Operativa
Fundació Miguel Augustí
Universitat Politècnica de Catalunya

RESUMEN

La sensometría es el área científica que aplica métodos estadísticos a los problemas de la ciencia sensorial y del consumidor. Dicha área, muy dinámica, evoluciona constante y rápidamente tanto en sus objetivos, dominios de aplicación y métodos. En un primer tiempo, se deseaba determinar las propiedades organolépticas de los productos alimenticios mediante jurados de panelistas entrenados. Luego, mejor conocer a los consumidores no entrenados, desde sus percepciones hasta sus preferencias, cambiantes costumbres y, más recientemente, sentimientos, emociones y deseos se convirtió en un tema prioritario. Tanto los campos de aplicación, los productos como los hábitos y demandas de los consumidores se diversificaron. Actualmente el análisis sensorial y la sensometría conciernen no sólo la industria agroalimentaria, sino también la industria cosmética, textil, automóvil, telefónica, publicidad, etc. La recogida de datos concierne ahora todos los sentidos; en particular, se debe resaltar los trabajos interesantes efectuados sobre los sonidos. Esta evolución en tan diversas direcciones ha estimulado la búsqueda de nuevos métodos sensoriales y sensométricos.

Los métodos estadísticos para analizar los datos sensoriales y los datos contextuales son muy variados. Podemos reagruparlos en métodos descriptivos cuantitativos, cualitativos y afectivos.

Entre los enfoques cualitativos, la descripción mediante comentarios libres aporta una información espontánea irreemplazable. Además, ha surgido un gran interés para el análisis conjunto de datos sensoriales y datos objetivos como la composición química de los productos, la composición del suelo, el clima, etc., lo que impulsó la utilización y desarrollo de métodos estadísticos para el análisis de tablas múltiples complejas.

Se presentan aquí los principales métodos sensométricos mediante varios ejemplos en el campo de la industria agroalimentaria, haciendo énfasis en los métodos holísticos y las tareas de verbalización libre.

Comunicacións orais

Sesión 1 de Investigación de Operacións

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Otimização do pessoal técnico do serviço de manutenção da TAP

Jorge Santos [1], Luís Cavique [2]

[1] Departamento de Matemática, CIMA, ECT, Universidade de Évora, jmas@uevora.pt
[2] SIFT, DCeT, Universidade Aberta, BioISI-MAS/LabMAg da FCUL, luis.cavique@uab.pt

RESUMO

Na TAP-AMT (airline maintenance technicians) durante o período de 24 horas de laboração existem intervalos de tempo com excesso de pessoal e outros intervalos com pessoal em falta, que muito dificulta a atividade do pessoal técnico do serviço de manutenção. Pretende-se desenvolver um modelo matemático, que de acordo com períodos de exploração comercial, permita criar um gráfico de cargas versus o pessoal disponível. Na solução encontrada as disponibilidades são sempre superiores às cargas de trabalho e não se apresentam diferenças significativas que indiciem desperdício

Palavras e frases chave: Otimização, Horários de pessoal, Escalas de pessoal

1. INTRODUÇÃO

No 101th European Study Group with Industry, que teve lugar no Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, de 5 a 9 de maio de 2014, foi apresentado pela transportadora aérea TAP um problema que consistia na otimização dos horários das suas equipas de manutenção. Neste trabalho apresentamos uma das soluções propostas para o problema.

Na secção 2 é apresentado o problema proposto pela TAP, na secção 3 é realizada uma breve revisão bibliográfica, de seguida, na secção 4 é proposto um modelo. Na secção 5 são apresentados os resultados computacionais obtidos e na secção 6 são discutidas as conclusões.

2. PROBLEMA TAP

Existem vários requisitos a respeitar neste estudo para a TAP-AMT (airline maintenance technicians).

Adequar a dotação de pessoal técnico ao serviço da Manutenção de Linha (ML) à exploração comercial do operador em Lisboa (LIS), considerando inclusive as possibilidades de alteração na estruturação da equipa ML.

Definir os horários dos turnos das equipas de produção da ML em LIS, de acordo com as rotações de aeronaves e quantitativos de pessoal técnico, enquadrada na regulamentação interna aplicável, baseado no acordo de empresa.

Criação de um Modelo Matemático, de acordo com períodos de exploração comercial, que permita projetar um gráfico de cargas versus o pessoal disponível.

2.1. Regras para criação de serviços

Existem um conjunto de regras contratuais que se podem resumir da seguinte forma:

i) Cada colaborador trabalha 7,5 horas no máximo, em regime de retribuição sem valor extraordinário; contudo, após 5 horas consecutivas de trabalho este deve interromper por 1 hora (daí dizermos que os turnos têm 8,5 horas); acresce que horários que se verifiquem com entradas antes das 07:00 no período da noite, determinam ½ hora de interrupção para pequeno-almoço.

As interrupções acima, quer a de refeição quer a de pequeno-almoço, podem não se concretizar por interesse da empresa; quando assim acontece, esses períodos são pagos em regime de horas extraordinárias, e o descanso depois de 5 horas de trabalho é efetuado; como nota deixamos a informação

que na estação de Verão do ano anterior a equipa de reforço entrou por horário às 06:00, tendo a empresa assumido os custos de não interrupção de pequeno-almoço.

ii) Entre períodos de trabalho, os colaboradores têm de descansar um mínimo de 12 horas.

iii) Os turnos de trabalho, cada 1 composto por uma equipa de base, são os seguintes:

- 1º turno, entrada às 00:00 e saída às 08:30 (têm uma interrupção de 1 hora para cear, nas horas de funcionamento da cantina, entre as 01:00 e as 04:00);
- 2º turno, entrada às 08:00 e saída às 16:30 (têm uma interrupção de 1 hora para almoço, nas horas de funcionamento da cantina, entre as 11:00 e as 15:00);
- 3º turno, entrada às 16:00 e saída às 00:30 (têm uma interrupção de 1 hora para jantar, nas horas de funcionamento da cantina, entre as 18:30 e as 21:30);
- Turno de reforço, é a distribuição de 1 equipa pelo 2º e 3º turno, nos quantitativos que se entenderem adequados (por exemplo, neste momento têm 75% de uma equipa para acréscimo ao 2º turno e os 25% remanescentes para acréscimo ao 3º turno);

iv) O horário de trabalho é cada equipa laborar 4 dias e folgar 2, na seguinte sequência de rotação pelos turnos de trabalho :

- 2º turno (vulgo manhãs) 4 dias; folgam 2 dias;
- 1º turno (vulgo noites) 4 dias; folgam 2 dias;
- Reforço (parte de 1 equipa vai para manhãs parte vai para as tardes) 4 dias; folgam 2 dias;
- 3º turno (vulgo manhãs) 4 dias; folgam 2 dias

2.2. Objetivo do trabalho

A Figura 2.1 mostra o problema que se pretende ser resolvido. A zona azul mostra as disponibilidades do pessoal. A linha a vermelho representa a média das restantes linhas, as quais indicam as necessidades dos vários dias da semana. Durante as 5, 6 e 12 horas existe um excesso de carga de trabalho relativamente às disponibilidades.

O objetivo passa por criar um Modelo Matemático, que de acordo com períodos de exploração comercial permita projetar um gráfico de cargas versus pessoal disponível.

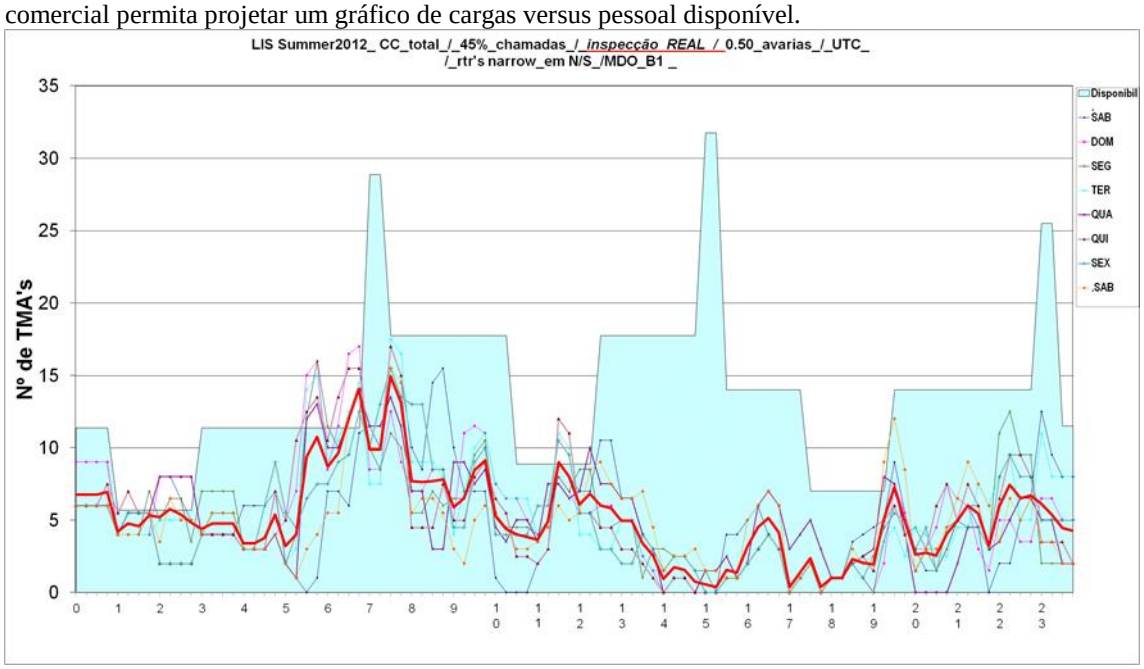


Figura 2.1 Gráfico de cargas e disponibilidades

3. REVISÃO BIBLIOGRÁFICA

A importância do tema de escalas e horário de pessoal quer em termos de investigação, quer em termos de aplicação prática está patente na revisão bibliográfica de Ernst et al. [2004].

No caso particular nos transportes públicos, a partir de 1975 foram realizadas seis conferências de "Computer-Aided Transit Scheduling": [Bodin e Bergman 1975], [Wren 1981], [Rousseau 1985], [Daduna e Wren 1988], [Desroches e Rousseau 1992] e [Daduna, Branco e Paixão 1995]. A compilação dos trabalhos mais relevantes podem ser encontrados em [Cavique 1994].

Na área de "Computer-Aided Transit Scheduling" distinguem-se três problemas diferentes:

- horário diário dos veículos (“vehicle scheduling”)
- horário diário do pessoal (“crew scheduling”)
- plano semanal, escalas ou turnos do pessoal (“crew rostering”)

Os modelos de planeamento de transportes têm em geral três componentes que podem ser estudadas sequencialmente: os horários dos veículos, os horários das tripulações e as escalas das tripulações. No horário dos veículos procura-se minimizar o número de veículos necessários para cobrir o horário ao público, sujeito a restrições tais como a autonomia dos veículos, percurso a realizar e a garagem a utilizar. O horário das tripulações é definido com vista a minimizar do número de tripulações (i.e. serviços necessários) para cobrir o horário dos veículos num dia em particular, sujeito a regras contratuais. A geração de escalas de tripulações corresponde à rotação dos serviços ao longo de uma semana, mês ou outro período mais longo, para que em cada dia estejam disponíveis os serviços necessários para assegurar o horário.

Para a resolução do problema da TAP-AMT, o problema dos veículos não se coloca pelo que nos vamos concentrar nos últimos dois.

3.1. Escalas de Pessoal

Existem vários sinónimos para as escalas de pessoal: plano semanal, turnos do pessoal, “crew rostering” ou “tour scheduling”.

O problema de geração de escalas consiste na afetação diária de um serviço a um recurso, existindo a obrigatoriedade de garantir a realização de todos os serviços para cada dia, num horizonte temporal específico. A representação matricial da escala inclui duas entradas, dia da semana e recurso que contêm os serviços.

Na abordagem do problema é objetivo criar escalas cíclicas e com variação regular, i.e. com cargas uniformes de trabalho, Figura 3.1, utilizando o menor número de efetivos necessários e respeitando as regras contratuais.

	A	**B**	**C**	**D**
Seg	1		2	
Ter		1		2
Qua	2		1	
Qui		2		1
Sex	1		2	
Sab		1		2
Dom	2		1	

Figura 3.1 - Escala com a duração de uma semana, utilizando os recursos <A,B,C,D> e cobrindo dois serviços <1,2>

No caso da TAP-AMT, dadas as 6 equipas de trabalhadores de manutenção aérea (A, B, C, D, E, F) é necessário cobrir 4 equipas para cobrir o trabalho diário, mantendo duas de folga. Ver Figura 3.2.

	1	**2**	**3**	**4**	**folga**
1	E	D	A	B	C F
2	E	D	A	B	C F
3	F	D	A	C	B E
4	F	D	A	C	B E
5	F	E	B	C	A D
6	F	E	B	C	A D
7	A	E	B	D	C F
1	A	E	B	D	C F
2	A	F	C	D	B E
3	A	F	C	D	B E
4	B	F	C	E	A D
5	B	F	C	E	A D
6	B	A	D	E	C F
7	B	A	D	E	C F

Figura 3.2- Escalas da TAP-AMT

A representação matricial da escala é ligeiramente diferente da anterior, inclui duas entradas, dia da semana e serviço, e contém os recursos no seu interior.

Dado que estamos a trabalhar com escalas, onde se pretende equilibrar o esforço nos serviços, surge o conceito de padrão de rotação ou ciclo trabalho - descanso. O padrão da TAP-AMT repete-se por vários meses até se voltar a encontrar a configuração inicial de "ED, AB – CF". A solução da TAP-AMT parece adequada às necessidades, pelo que não carece de ser otimizada.

3.2. Horário de pessoal

Existem vários sinónimos para sequenciamento de pessoal: horário do pessoal ou "crew scheduling".

Antes de descrever as restrições do problema é conveniente definir alguns termos usuais, por forma a uniformizar a nomenclatura utilizada na literatura especializada.

* Período de trabalho - com duração máxima de 5 horas e sujeito a restrições contractuais;

* Período de descanso ("meal break") - com duração de 1 hora para descanso e para tomar uma refeição;

* Serviço - é composto por dois períodos de trabalho, separados por um período de descanso, tendo de satisfazer regras contratuais.

Dos algoritmos desenvolvidos para o problema do sequenciamento de pessoal, implementados com sucesso distinguem-se: a Heurística do tipo de "Run-Cutting", o Método de Emparelhamento e os procedimentos baseados em formulações do Problema de Cobertura de um Conjunto.

A heurística mais comum para o problema de sequenciamento de serviços é a heurística de "run-cutting", inspirada no método manual de afetação de pessoal. O algoritmo gera um serviço admissível ("run") sobre o conjunto de peças de trabalho, cobrindo de seguida o horário com o novo serviço gerado. O processo repete-se até que não seja possível gerar mais nenhum serviço admissível para o horário a cobrir, ou que exista cobertura total. A tradução de "run-cutting" será então o corte sucessivo do horário com serviços, ou mais sucintamente cobrir-com-serviços.

Na abordagem do problema de sequenciamento de tripulações utilizando o método de emparelhamento, é usual a divisão em dois sub-problemas: (i) criar um conjunto de tramos que respeitem as normas contratuais; (ii) emparelhar dois ou mais tramos formando um serviço, respeitando também as regras contratuais.

O método utilizando a formulação de partição / cobertura de um conjunto decompõe-se em duas partes: uma primeira relativa à geração de um conjunto grande de serviços, teoricamente todos os serviços admissíveis e uma segunda de aplicação do problema de partição / cobertura de um conjunto. Em seguida é apresentada a formulação do problema de partição de um conjunto.

$$\text{Min} \quad \sum_{j \in S} c_j x_j$$

s.a.

$$(1) \quad \sum_{j \in S} a_{ij} x_j = 1 \qquad \forall i \in I$$

$$(2) \quad x_j \in \{0,1\} \qquad \forall j \in S$$

com S={serviços admissíveis}

$$x_j = \begin{cases} 1 \text{ se o serviço é admissível} \\ 0 \text{ caso contrário} \end{cases}$$

$$a_{ij} = \begin{cases} 1 \text{ se a peça de trabalho i é cobertura pelo serviço j} \\ 0 \text{ caso contrário} \end{cases}$$

A matriz [aij] contém um conjunto de combinações de peças de trabalho que verificam as normas contratuais, i.e. um conjunto de serviços admissíveis. O número de linhas, I, é dado pelo número de peças de trabalho e o número de colunas, S, corresponde aos serviços gerados. Para o problema de cobertura a restrição (1) é relaxada utilizando o sinal maior ou igual, garantindo que todas as peças de trabalho são cobertas pelo menos uma vez, podendo originar sobre-coberturas i.e. disponibilizar dois ou mais serviços para o mesmo conjunto de peças de trabalho. O problema de cobertura de um conjunto é formulado como :

$$\text{Min} \sum_{j \in S} c_j x_j$$

s.a.

$$(1) \sum_{j \in S} a_{ij} x_j \geq 1 \qquad \forall i \in I$$

$$(2) x_j \in \{0,1\} \qquad \forall j \in S$$

Uma revisão bibliográfica mais detalhada pode se encontrada em Cavique [1994].

4. MODELO PROPOSTO

Para cada hora existe um valor de carga de trabalho em Horas.homem (H.h) que é necessário garantir. Nesta abordagem iremos "discretisar" as variáveis em horas em que a carga de trabalho deve ser sempre maior que os funcionários disponíveis.

Considerando o abstencionismo cada equipa tem 10,7 elementos. Ao fim de 3 horas 1/3 da equipa vai descansar; ao fim de 4 horas o segundo terço vai descansar e ao fim de 5 horas o último terço entra em período de descanso. Na Figura 4.1. é apresentado um serviço de 8 horas de uma equipa em H.h. O custo associado a esta equipa foi considerado de 4, já que as equipas se podem dividir em quatro partes iguais. Neste estudo iremos considerar equipas inteiras e equipas parciais com ¼ dos elementos.

Horas	0	1	2	3	4	5	6	7	cj
x_1	10,7	10,7	10,7	7,1	7,1	7,1	10,7	10,7	4

Figura 4.1: Serviço de 8 horas de uma equipa em H.h

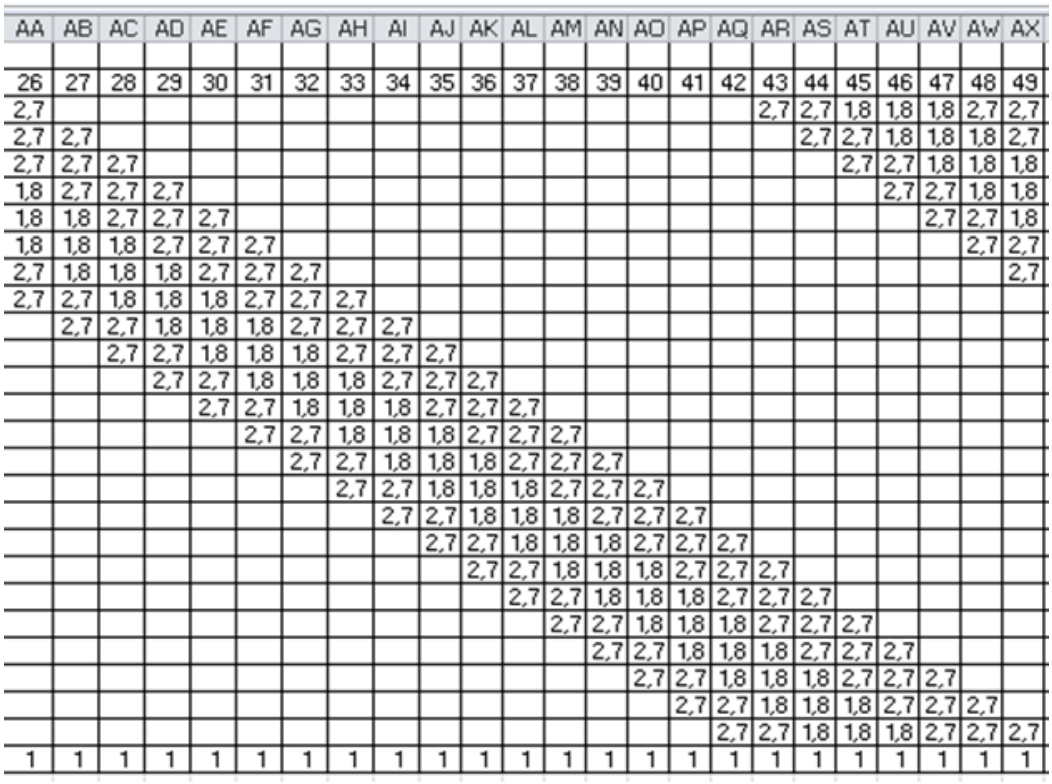
Com base no serviço criado vamos gerar novas colunas com hora de início diferente. Na Figura 4.2 apresentamos a geração de serviços inteiros.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Xj																								
2	horas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
3	0	10,7																	10,7	10,7	7,1	7,1	7,1	10,7	10,7
4	1	10,7	10,7																	10,7	10,7	7,1	7,1	7,1	10,7
5	2	10,7	10,7	10,7																	10,7	10,7	7,1	7,1	7,1
6	3	7,1	10,7	10,7	10,7																	10,7	10,7	7,1	7,1
7	4	7,1	7,1	10,7	10,7	10,7																	10,7	10,7	7,1
8	5	7,1	7,1	7,1	10,7	10,7	10,7																	10,7	10,7
9	6	10,7	7,1	7,1	7,1	10,7	10,7	10,7																	10,7
10	7	10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7																
11	8		10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7															
12	9			10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7														
13	10				10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7													
14	11					10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7												
15	12						10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7											
16	13							10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7										
17	14								10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7									
18	15									10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7								
19	16										10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7							
20	17											10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7						
21	18												10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7					
22	19													10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7				
23	20														10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7			
24	21															10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7		
25	22																10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7	
26	23																	10,7	10,7	7,1	7,1	7,1	10,7	10,7	10,7
27	cj	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

Figura 4.2. – Geração de serviços inteiros.

O processo pode ser repetido para a geração de serviços parciais com um custo unitário. As equipas contam com 10,7 horas para 1/4 dos elementos e são respeitadas as mesmas regras para o período de descanso.

Utilizar o modelo de cobertura referido anteriormente, utilizando para matriz $[a_{i,j}]$ os vários serviços gerados em H.h.



AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49
2,7																	2,7	2,7	1,8	1,8	1,8	2,7	2,7
2,7	2,7																	2,7	2,7	1,8	1,8	1,8	2,7
2,7	2,7	2,7																	2,7	2,7	1,8	1,8	1,8
1,8	2,7	2,7	2,7																	2,7	2,7	1,8	1,8
1,8	1,8	2,7	2,7	2,7																	2,7	2,7	1,8
1,8	1,8	1,8	2,7	2,7	2,7																	2,7	2,7
2,7	1,8	1,8	1,8	2,7	2,7	2,7																	2,7
2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7																
	2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7															
		2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7														
			2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7													
				2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7												
					2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7											
						2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7										
							2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7									
								2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7								
									2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7							
										2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7						
											2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7					
												2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7				
													2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7			
														2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7		
															2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7	
																2,7	2,7	1,8	1,8	1,8	2,7	2,7	2,7
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figura 4.3. – Geração de serviços parciais

$$Min \sum_{j \in S} c_j . x_j$$

$$s.a.\ (1) \sum_{j \in S} a_{i,j} . x_j \geq carga \qquad \forall i \in I$$

$$(2) \quad x_j \in \{0,1\} \quad \forall j \in S$$

Impomos como restrições adicionais que os serviços inteiros não devem exceder o valor 3 e os serviços parciais não devem exceder o valor 4. Na figura 4.4 apresentamos a formulação no Solver do Excel.

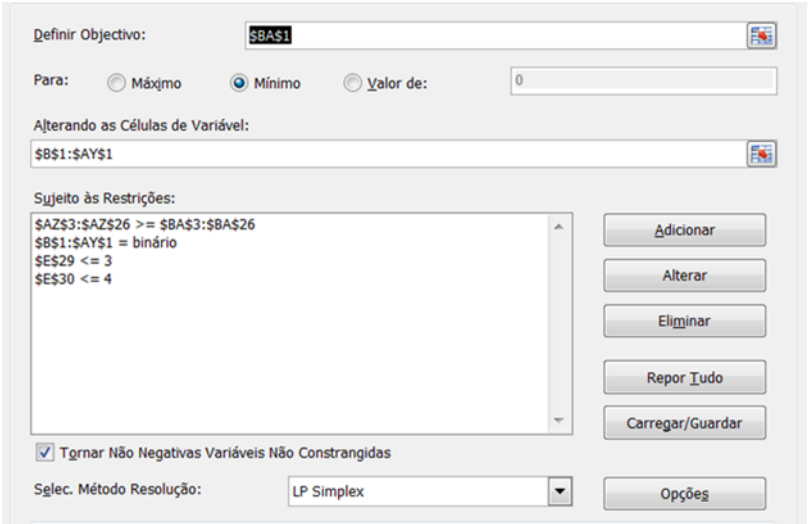


Figura 4.4. – Implementação no Solver do Excel

5. RESULTADOS COMPUTACIONAIS

Os resultados computacionais encontrados apontam para a utilização de 3 serviços inteiros e 2 serviços parciais, ver figura 5.1. Os serviços inteiros têm início as 5:00, 13:00 e 21:00 e são consecutivos, pelo que existe um período de 30 minutos que permite a passagem de testemunho.

X_i	I1	I2	I3	P1	P2	F.O.=	**14**
horas						disponíveis	carga
0			7,1	2,7		9,8	8,6
1			7,1	2,7		9,8	4,8
2			7,1	2,7		9,8	2,0
3			10,7	1,8		12,5	4,0
4			10,7	1,8		12,5	5,9
5	10,7			1,8		12,5	12,1
6	10,7			2,7		13,4	6,2
7	10,7			2,7	2,7	16,1	13,5
8	7,1				2,7	9,8	7,6
9	7,1				2,7	9,8	5,3
10	7,1				1,8	8,9	8,1
11	10,7				1,8	12,5	11,7
12	10,7				1,8	12,5	7,1
13		10,7			2,7	13,4	4,4
14		10,7			2,7	13,4	2,4
15		10,7				10,7	5,1
16		7,1				7,1	6,3
17		7,1				7,1	4,7
18		7,1				7,1	2,1
19		10,7				10,7	8,4
20		10,7				10,7	2,1
21			10,7			10,7	4,6
22			10,7			10,7	8,6
23			10,7			10,7	6,5
c_j	4	4	4	1	1		

Figura 5.1 – Soluções encontradas.

O custo da função objetivo, F.O., é de 14, inferior ao valor esperado de 16 que corresponde à utilização das 4 equipas. Com esta solução existem 2 soluções parciais que ficam disponíveis.

Graficamente na Figura 5.2 apresentamos a carga de trabalho e às disponibilidades encontradas. As disponibilidades são sempre superiores às cargas, não apresentando contudo diferenças significativas que indiciem desperdício.

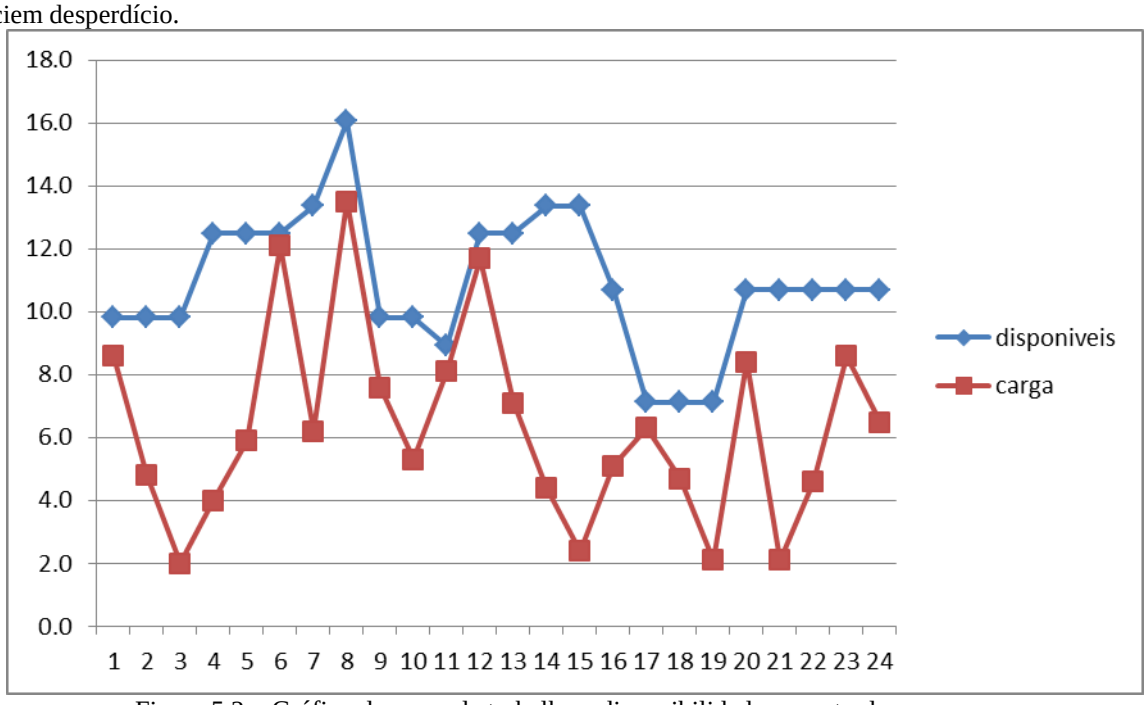


Figura 5.2 – Gráfico de carga de trabalho e disponibilidade encontradas

6. CONCLUSÕES

Dado o problema da TAP- AMT com vista a cobrir a falta de disponibilidade de pessoal técnico do serviço de manutenção durante o período de 24 horas de laboração, bem como de minimizar o excesso de pessoal, foram considerados dois sub-problemas: escalas de pessoal e o horário do pessoal.

A solução das escalas da TAP-AMT parece adequada às necessidades, pelo que não carece de ser otimizada. Para o problema dos horários de pessoal foi desenvolvido um modelo matemático com a formulação do problema de cobertura.

De acordo com períodos de exploração comercial, foi comparada a carga de trabalho com o pessoal disponível, não tendo sido encontradas falhas. As disponibilidades são sempre superiores às cargas de trabalho. Para os excessos de pessoal significativos são aproveitados pelos trabalhadores para formação on-line, pelo que não representam desperdício.

Depois de encontradas as soluções descritas, têm decorrido reuniões com os responsáveis dos serviços da TAP-AMT, por forma a avaliar os serviços gerados e em custear novos serviços em busca de novas soluções.

REFERENCIAS

Bodin L. e D. Bergman eds (1975), Workshop on Automated Techniques for Scheduling of Vehicle Operator for Urban Public Transportation Services (Preprint).

Cavique L. (1994), dissertação de mestrado Sequenciamento de pessoal tripulante, Investigação Operacional e Engenharia de Sistemas, IST-UTL.

Daduna J.R., Branco e Paixão J.P. (1995), Computer-Aided Scheduling of Public Transport, Springer-Verlag.

Daduna J.R. e A.Wren eds (1988), Computer-Aided Transit Scheduling ,Springer-Verlag.

Desroches M. e J.M.Rousseau eds (1992), Computer-Aided Transit Scheduling, Springer-Verlag.

Ernst A.T., H. Jiang, M. Krishnamoorthy, D. Sier (2004), Staff scheduling and rostering: A review of applications, methods and models, European Journal of Operational Research, vol. 153, pp. 3–27.

Rousseau J.M. eds (1985), Computer Scheduling of Public Transport 2, North Holland.

Wren A. eds (1981), Computer Scheduling of Public Transport, North Holland 1981.Nash, J. (1950) Equilibrium points in n-person games. Proceedings of the National Academy of the USA,

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

Optimización del proceso de llenado de placas PCR en secuenciación Sanger

Carpente R., M. Luisa[1], Cerdeira-Pena, Ana[2], Lorenzo-Freire, Silvia[1] y Places, Ángeles S.[2]

[1] Grupo MODES, Departamento de Matemáticas, Universidade da Coruña
[2] Laboratorio de Bases de Datos, Departamento de Ciencias de la Computación, Universidade da Coruña

RESUMEN

En este trabajo se resuelve el problema de optimización relativo al proceso de llenado de placas PCR en secuenciación Sanger. El objetivo de este problema de optimización es obtener la disposición de los fragmentos de muestras de ADN en placas PCR, minimizando el número de placas PCR a utilizar y maximizando el porcentaje de llenado de éstas. Para ello, se han de tener en cuenta una serie de restricciones impuestas por el funcionamiento de las máquinas en que las placas son procesadas (los termocicladores) , así como derivadas de las propias reacciones PCR.

Para resolver este problema de optimización se ha hecho uso de herramientas heurísticas, en concreto, se ha utilizado el método de *simulated annealing*. De esta forma, se ha creado una librería software, llamada *simPCR*, fácilmente integrable en otras aplicaciones, y que, actualmente, ya está siendo utilizada con éxito en entornos productivos reales de laboratorios de análisis clínicos genéticos.

Palabras y frases clave: proceso PCR, optimización, algoritmos heurísticos, simulated annealing.

1. INTRODUCCIÓN

El proceso de *secuenciación* es aquel a través del cual es posible determinar el orden de los nucleótidos (A, C, G, y T) en un segmento de ADN. En otras palabras, es el proceso capaz de leer nuestros genes. Sin embargo, su implicación va más allá de una simple lectura, ya que además de permitir conocer nuestra información genética, la *secuenciación* abre la posibilidad de predecir la predisposición o riesgo a padecer ciertas enfermedades, mediante la identificación de variantes genéticas respecto a un ADN de referencia, pero también incluso a adaptar algunos tratamientos en función de nuestro propio genoma.

El conocido como método *Sanger* es una de las principales técnicas de referencia para la secuenciación del ADN, el cual se compone, a grandes rasgos, de las siguientes fases:

- Extracción de ADN genómico a partir de muestras (sangre, saliva, tejido, etc.)
- Dilución del ADN genómico extraído
- Aplicación de la reacción en cadena de la polimerasa (*proceso PCR*)
- Purificación de productos PCR
- Secuenciación *forward* y *reverse*

Todas ellas juegan un papel fundamental en el proceso global de secuenciación Sanger, si bien uno de los pasos críticos lo encontramos en el denominado *proceso PCR*. Se trata de un método de amplificación del ADN, cuyo objetivo es obtener múltiples copias de un fragmento o conjunto de fragmentos de ADN. Para ello, los fragmentos son dispuestos en placas (también conocidas como *placas PCR*), e introducidas en un termociclador que recrea los ciclos de temperaturas necesarios para producir

una reacción en cadena de la polimerasa. El problema radica en que las posiciones de dichas placas no se pueden disponer libremente, ya que están sujetas a limitaciones derivadas del propio funcionamiento de los programas de termociclador y de las temperaturas características de las reacciones PCR.

Obtener la distribución óptima de los fragmentos en las placas PCR de manera eficiente, en base a una serie de condiciones impuestas, supone de entrada un ahorro económico (el coste de procesamiento por placa puede ser elevado, por lo que se debe intentar utilizar el mínimo número necesario de ellas y maximizar su porcentaje de llenado, en la medida de lo posible), pero también un no menos despreciable ahorro de tiempo (el número de termocicladores de que disponen los laboratorios es limitado, y el actual proceso de distribución de fragmentos en placas, en muchos de ellos, incurre en tiempos prohibitivos, generando el retraso de muchos estudios y listas de espera inadmisibles), lo que permitiría, por extensión, un incremento del beneficio económico, dando cabida a nuevos proyectos/clientes.

En este trabajo se plantea la solución al problema de optimización del proceso de llenado de placas PCR en secuenciación Sanger, haciendo uso de herramientas heurísticas.

2. EL PROBLEMA DE OPTIMIZACIÓN

2.1. Descripción del problema

El objetivo de este trabajo es obtener la distribución de los fragmentos de muestras de ADN en placas PCR, minimizando el número de placas PCR utilizadas y maximizando el porcentaje de llenado de dichas placas. Para ello, se dispone de los siguientes datos de entrada:

- Datos de entrada relativos a los fragmentos de ADN procedentes de muestras biológicas:

  Código individual: número identificativo correspondiente al par muestra-fragmento concreto.
  Código grupal: todos los pares muestra-fragmento que hagan referencia al mismo fragmento de ADN, se considera que pertenecen a un mismo grupo; ya que las condiciones PCR están asociadas a fragmentos, independientemente de la muestra origen de la que procedan. Es condición indispensable que, si una placa PCR alberga muestras de un fragmento (grupo) determinado, en dicha placa exista una posición reservada para albergar el denominado *control*[1] de grupo.
  Temperatura: temperatura a la que se debe someter el par muestra-fragmento durante el proceso PCR.
  Prioridad: parámetro opcional para establecer la prioridad en el procesamiento de un par muestra-fragmento. La prioridad más baja se identifica con el valor 1, siguiendo las restantes una secuencia ascendente.

- Datos de entrada relativos a la configuración de las placas PCR:

  Número de filas por placa.
  Número de columnas por placa.
  Número de franjas por placa: una franja incluye varias columnas, y delimita las zonas de la placa que pueden compartir una misma temperatura.
  Diferencia máxima de temperatura entre franjas de una placa: la diferencia de temperatura entre franjas estará restringida por este valor.

En cuanto a las restricciones del problema, además de las condiciones dadas por el propio conjunto de muestras de entrada, también hay que considerar otras restricciones impuestas por el propio funcionamiento de los termocicladores. De esta forma, habría que tener en cuenta las siguientes restricciones:

[1] Necesario para validar la corrección del proceso PCR sobre las muestras correspondientes a dicho fragmento (grupo).

- Todas las posiciones de las columnas correspondientes a una franja (de una placa PCR) deben tener la misma temperatura.
- La diferencia de temperatura entre franjas adyacentes de una misma placa nunca debe superar la diferencia máxima permitida.
- Toda placa que contenga muestras correspondientes a un fragmento (grupo) determinado, debe albergar también una posición en ella en la que se aloje el control asociado (cuya temperatura será la definida por el propio fragmento).

La solución que devuelva el problema de optimización consistirá en una lista de placas PCR que deben introducirse en el termociclador para su procesamiento, indicando, para cada posición física de la placa, el identificador del par muestra-fragmento o bien el código de grupo ubicado en dicha placa. En el segundo caso, el identificador del grupo indica que, en esa posición de la placa, estará contenido el control de grupo del fragmento asociado al mismo.

Para obtener dicha solución se ha considerado el método de *simulated annealing*, introducido por Kirkpatrick et al. (1983). Este método establece una analogía entre el proceso de recocido de ciertos materiales, como el acero, y ciertos problemas de optimización. Una revisión detallada de la aplicación de este método a los problemas de optimización puede encontrarse en los trabajos de Aarts et al. (1997) o Eglese (1990). El algoritmo de *simulated annealing* parte de una solución inicial y luego se mueve en el espacio de soluciones, siguiendo la filosofía del proceso de recocido de acero y cerámicas.

2.2. Obtención de la solución inicial

El procedimiento de obtención de la solución inicial consiste en ordenar los fragmentos de muestras de ADN de acuerdo a su prioridad y, para una misma prioridad, por temperatura. A continuación, siguiendo dicho orden, cada par muestra-fragmento se ubicará en una placa de acuerdo al siguiente procedimiento:

1) Se recorre la lista de placas existentes y ordenadas de acuerdo a su orden inicial.

   a) Para una placa concreta, se comprueban las franjas con pares muestra-fragmento y/o controles de grupo ya asignados (i.e. no vacías). En cada franja evaluada, cuya temperatura coincida con la del par a colocar se observa:

      i) Si el grupo (fragmento) correspondiente al par ya tiene presencia en la placa, se comprueba si existe espacio en la franja para únicamente albergar ese par; añadiéndose en caso afirmativo.
      ii) Si el grupo (fragmento) no tiene presencia todavía en la placa, se comprueba si existe espacio en la franja para alojar el par muestra-fragmento en cuestión, así como el control de grupo asociado, añadiéndose ambos si se dan las condiciones.
      iii) De no haber sitio suficiente en la franja seleccionada, para albergar, según el caso, el par muestra-fragmento y, eventualmente, el control de grupo correspondiente, se pasa a evaluar la siguiente franja.

   b) Una vez evaluadas todas las franjas con algún par muestra-fragmento y/o control de grupo ya asignados, se repite el mismo proceso, pero con las franjas actualmente vacías. Cabe tener presente que en el momento que a una franja vacía se le asigna un par muestra-fragmento y/o control de grupo, dicha franja pasa a tener asociada automáticamente, una temperatura igual a la del par / control alojado. Por lo tanto, en esta ocasión, además de comprobar si se debe añadir o no el control de grupo (según el fragmento tenga o no ya presencia en la placa), se debe garantizar siempre, antes de efectuar la asignación, que se respeta la diferencia máxima de temperatura permitida entre franjas adyacentes de una placa. De no ser así, se pasará a revisar la siguiente franja vacía.

   c) En el caso de que no haya sido posible ubicar el par en ninguna de las franjas de la placa de estudio, se repite el mismo proceso con la siguiente placa de la lista.

2) En última instancia, si la colocación del par muestra-fragmento no es factible en ninguna de las placas existentes, se procederá a crear una nueva placa, siguiendo, a continuación, el mismo procedimiento general ya descrito para una placa cualquiera.

Al término de este proceso, se obtiene una primera solución de partida que servirá de entrada para el algoritmo de optimización.

2.4. Mejora de la solución inicial

En este paso obtendremos la solución final del problema de optimización. Utilizaremos un algoritmo heurístico, basado en el procedimiento de *simulated annealing*, y que partirá de la solución inicial generada anteriormente.

Para implementar este método es necesario definir el problema en términos de un espacio de soluciones y una función objetivo definida sobre aquel, para efectuar la búsqueda de la mejor aproximación al óptimo. Tal y como se señaló en secciones anteriores, el espacio de búsqueda lo definen y delimitan las restricciones que hay que contemplar a lo largo del proceso, mientras que la función objetivo viene determinada, por un lado, por la minimización del número de placas de la solución y, por otro, por la maximización de sus respectivos niveles de ocupación.

En base a ello, el procedimiento general de este algoritmo parte de una solución inicial, y continúa con la realización de movimientos a través del espacio de soluciones, seleccionando, en cada paso, una nueva solución vecina de la solución actual, y comparando la diferencia de valor (coste), entre ambas en relación a la función objetivo. Si la nueva solución alcanzada, proporciona una mejora de la función objetivo, sustituirá a la solución actual. En caso contrario, podrá ser seleccionada todavía con una determinada probabilidad, dada por $e^{-\beta/T}$, donde β es el incremento de valor (i.e. empeoramiento en coste de la función objetivo) y T es un parámetro de control correspondiente a la temperatura, en analogía con el proceso de *physical annealing*. Este procedimiento se repite de manera iterativa desde una temperatura, T_{max}, de partida, durante diferentes fases, en las que dicha temperatura va disminuyendo en base a un parámetro de enfriamiento, α, hasta alcanzar un mínimo, T_{min}. El proceso global, a su vez, se repite durante un número fijo de iteraciones. A lo largo de todo este proceso se mantiene siempre actualizada la mejor solución alcanzada, devolviéndose ésta, tras su finalización, como mejor aproximación al óptimo.

Uno de los elementos más importantes en el éxito de un algoritmo de estas características lo encontramos en el diseño de los movimientos que se llevan a cabo para generar una nueva solución a partir de otra (i.e. para moverse de una solución a otra a través del espacio de búsqueda). En el caso que nos ocupa, los movimientos creados para la optimización del problema han sido dos, los cuales son aplicados en base a una probabilidad. Si se da la situación en que ninguno de los dos movimientos es posible (i.e. ninguno da lugar a una solución factible desde una solución actual), el algoritmo se dará también por finalizado.

Se describe, a continuación, el funcionamiento general de los dos movimientos implementados:

- *Intercambio de franjas*: permuta de franjas escogidas al azar entre placas diferentes, de forma factible.
- *Reagrupamiento*: agrupamiento de conjuntos de pares muestra-fragmento correspondientes a un mismo grupo (fragmento), dispersos en franjas de placas diferentes, mediante su adición y/o posible intercambio con otros conjuntos de pares muestra-fragmento (según el espacio disponible).

Una vez finalizado el proceso de optimización, la solución final es sometida a un post-procesado para determinar la localización concreta de cada par muestra-fragmento y control de grupo en las posiciones físicas de una placa PCR. Es importante mencionar que la heurística trabaja con asignaciones a nivel de franja dentro de las placas, por tanto es en este punto cuando las posiciones finales (i.e. fila y

columna dentro de una placa) se fijan. Para ello, se tiene en cuenta que, en el espacio definido por franjas con igual temperatura dentro de una placa, todos los pares muestra-fragmento correspondientes a un mismo grupo deben ir consecutivos, siendo finalizados por el control de grupo asociado.

2.4. Ejemplo

Tal y como se ha comentado en secciones previas, la algoritmia desarrollada para abordar el problema planteado constituye el núcleo de una librería que hemos dado en llamar *simPCR*. Desarrollada en C# usando Visual Studio 2013, su integración en una tercera aplicación que precise hacer uso de ella, es tan simple como añadir como referencia la *.dll* compilada de la misma. De hecho, a día de hoy, esta librería está siendo utilizada con éxito, en entornos productivos reales de laboratorios de análisis clínicos genéticos. Diferentes validaciones experimentales realizadas, precisamente, en dichos entornos han permitido contrastar el buen comportamiento de nuestros algoritmos frente a soluciones previas alcanzadas con software alternativo (como LabWare), constatando que simPCR logra mejorar sustancialmente los resultados obtenidos, con una reducción de los tiempos de ejecución en varios órdenes de magnitud (de horas, a segundos), y una disminución considerable del número de placas en volúmenes de datos de entrada importantes.

A continuación, y a modo ilustrativo, la Figura 1 muestra un sencillo ejemplo de las salidas inicial y optimizada correspondientes a un conjunto de fragmentos de ADN procedentes de 151 muestras diferentes, cuya distribución debía realizarse en placas PCR de 8 filas y 6 franjas (de 2 columnas cada una de ellas) con una diferencia máxima de temperatura entre franjas permitida de 5º. Como se puede observar, aunque la solución inicial alcanzada logra distribuir todas los pares muestra-fragmento en 3 placas diferentes con porcentajes de llenado semejantes; es después, tras aplicar el proceso de optimización heurístico, cuando se consigue una mejora de la distribución final, completando el llenado de una de las placas y prácticamente el de otra, restando una menor parte en la tercera. Aparte del tiempo empleado (~5 seg.), la redistribución obtenida ofrece ventajas adicionales, permitiendo que la tercera placa pueda incluso ser utilizada en un proceso de llenado posterior con nuevas muestras, y conseguir con ello, a mayores, una optimización indirecta del número de recursos necesarios (placas PCR) en procesos de llenado que tengan lugar en cadena.

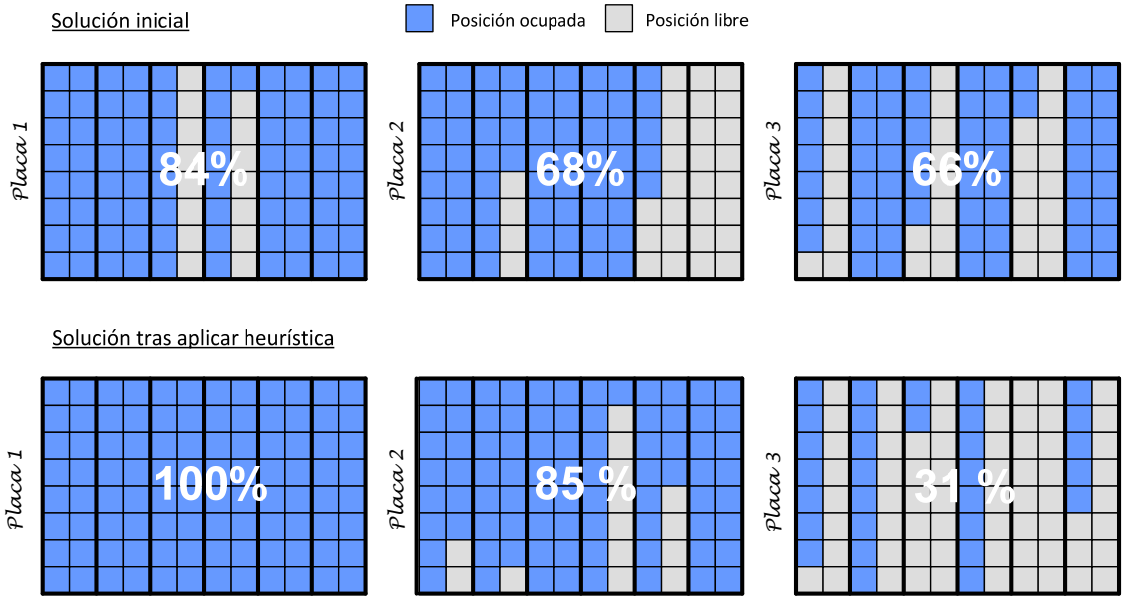


Figura 1: Placas PCR resultantes y porcentaje de llenado de las mismas, tras aplicar la solución inicial, y después del proceso de optimización heurístico sobre un conjunto de muestras de entrada.

REFERENCIAS

Aarts, E.H.L, Korst, J.H.M. y van Laarhoven, P.J.M. (1997) Simulated annealing. In Aarts, E., Lenstra, J.K. (eds) Local search in combinatorial optimization. Wiley, New York, 91-120.

Eglese, R.W. (1990) Simulated annealing: a tool for operational research. European Journal of Operational Research, 46, 271-281.

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

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

ON THE SET OF EXTREME CORE ALLOCATIONS FOR MINIMUM COST SPANNING TREE PROBLEMS

Christian Trudeau[1] and Juan Vidal-Puga[2]

[1]Department of Economics, University of Windsor.
[2]Research Group of Economic Analysis and Departamento de Estatistica e IO, Universidade de Vigo.

ABSTRACT

Minimum cost spanning tree problems connect agents efficiently to a source when agents are located at different points and the cost of using an edge is fixed. We propose a method, based on the concept of marginal games, to generate all extreme points of the corresponding core. We show that three of the most famous solutions to share the cost of mcst problems, the Bird, folk and cycle-complete solutions, are closely related to our method.

Keywords: Minimum cost spanning tree problems; extreme core allocations, reduced game, Bird solution, folk solution, cycle-complete solution.

1. INTRODUCTION

Minimum cost spanning tree (mcst) problems model a situation where agents are located at different points and need to be connected to a source in order to obtain a good or information. Agents do not care if they are connected directly to the source or indirectly through other agents who are. The cost to build a link between two agents or an agent and the source is a fixed number, meaning that the cost is the same whether one or ten agents use that particular link. Mcst problems can be used to model various real-life problems, from telephone and cable TV to water supply networks.

The core of mcst problems has been and early focus of attention, with Bird (1976) and Granot and Huberman (1981) showing that it is always non-empty and Granot and Huberman (1984) providing an algorithm to generate multiple core allocations. We present an improvement over these results by providing a method that allows to obtain the full set of extreme core allocations.

The method is based on the concept of marginal games (Núñez and Rafels , 1998), where we assign an agent her marginal cost to join the grand coalition, remove her from the problem and update the stand-alone costs of the remaining coalitions: they can either keep their original stand-alone cost or the stand-alone cost of them with the departing player, net of her cost share. This reduction is itself a special case of the Davis-Maschler reduction (Davis and Maschler, 1965). Given an ordering of the agents, repeating the process until all players are removed allows to find an extreme core allocation.

This method or very similar ones have been implemented for the assignment problem (Núñez and Rafels, 2003) and shortest path problems (Bahel and Trudeau, 2014), among others. In the non-cooperative setting, a similar approach consists in ordering buyers according to a given permutation and letting them buy goods in that order (Prez-Castrillo and Sotomayor (2002), VidalPuga (2004)).

The method does not work as well on all problems. Nez and Rafels (1998) provide sufficient conditions for the method to always generate extreme core allocations. Under stricter conditions, we obtain the full set of extreme core allocations[1]. These conditions are not satisfied by mcst

[1]A related approach is that of Tijs et al. (2011), which also looks for extreme core allocations given some lexicographic order. However, their approach is explicitly based on the core constraints and not on marginal games.

problems. We are still able to prove that the method generates the full set of extreme core allocations, using a representation of marginal games as minimum cost spanning tree problems with priced nodes. This new problem is a generalization of both mcst problems and Steiner tree problems (Hwang and Richards (1992), Skorin-Kapov (1995)).

By taking the average of these extreme core allocations for all permutations, we obtain a very natural cost sharing solution, that corresponds to the barycenter of the core.[2] If the game is concave, it also corresponds to the Shapley value. We show that our procedure is very close to three well-known cost-sharing solutions for mcst problems.

First, if we only consider permutations that correspond to the order in which we connect agents in an optimal network configuration, we obtain directly the Bird solution (Bird ,1976). The Bird solution was the first solution to be shown to always be in the core and it is known for its simplicity, as we may assign cost at the same time as we construct an optimal tree.

Secondly, we show that for elementary problems (where all costs are either 0 or 1), our solution corresponds to the cycle-complete solution (Trudeau, 2012). The cycle-complete solution is obtained by modifying the cost of some links before taking the Shapley value of the corresponding cost game: we reduce the cost of edge (i, j) if there exists a cycle that goes through nodes i and j and such that its most expensive edge is cheaper than the direct edge (i, j). The modification is enough to make the corresponding cost game concave, and thus the Shapley value stable.

Thirdly, we show that for elementary problems, if we modify the game to make it monotonically increasing (so that an agent never reduces the cost when joining a coalition) before applying our solution, we obtain the folk solution (Feltkamp et al. (1994), Bergantiños and Vidal-Puga (2007)). The folk solution is obtained in the same way as the cycle-complete solution, but we look at paths instead of cycles. An interpretation of our result is that the folk solution is the barycenter of the non-negative core (the set of stable allocations such that no agent is subsidized).

Given that for the cycle-complete and folk solutions, we obtain cost shares for a general problem by decomposing it in a series of elementary problems, the results on their correspondence with our solution do not hold in general. However, our method provides a new way to extend the cycle-complete and folk solution from elementary problems to more general mcst problems.

The paper is divided as follows: Section 2 defines the minimum cost spanning tree problems. In Section 3 we describe our method and show that it generates extreme core allocations. We show that it generates the full set of extreme core allocations in Section 4. Links with popular cost sharing solutions are explored in Section 5. Section 6 contains some discussions.

2. THE MODEL

A (cost sharing) game is a pair (N, C) where $N = 1, /dots, n$ is a nonempty, finite set of agents, and C is a characteristic function that assigns to each nonempty coalition $S \subseteq N$ a nonnegative cost $C(N) \in \mathbb{R}_+$ that represents the price agents in S should pay in order to receive a service. In particular, we assume that the agents in N need to be connected to a source, denoted by 0. Let $N_0 = N \cup \{0\}$. For any set Z, define Z^p as the set of all non-ordered pairs (i, j) of elements of Z. In our context, any element (i, j) of Z^p represents the edge between nodes i and j. Let $c = (c_e)_{e \in N_0^p}$ be a vector in $\mathbb{R}_+^{N_0^p}$ with $N_0^p = (N_0)^p$ and c_e representing the cost of edge e. Let Γ be the set of all cost vectors. Since c assigns cost to all edges e, we often abuse language and call c a cost matrix. A minimum cost spanning tree problem is a triple $(0, N, c)$. Since 0 and N do not change, we omit them in the following and simply identify a mcst problem $(0, N, c)$ by its cost matrix c.

A cycle p_{ll} is a set of $K \geq 3$ edges (i_k, i_{k+1}), with $k \in \{0, \dots, K-1\}$ and such that $i_0 = i_K = l$ and $i_1, \dots, i_{K-1}$ distinct and different than l. A path p_{lm} between l and m is a set of K edges (i_k, i_{k+1}), with $k \in \{0, \dots, K-1\}$, containing no cycle and such that $i_0 = l$ and $i_K = m$.Let $P_{lm}(N_0)$ be the set of all such paths between nodes l and m.

A spanning tree is a non-orientated graph without cycles that connects all elements of N_0. A spanning tree t is identified by the set of its edges. Its associated cost is $c(t) = \sum_{e \in t} c_e$.

We call mcst a spanning tree that has a minimal cost. Note that the mcst might not be unique. Let $t^*(c)$ be a mcst and $T^*(c)$ be the set of all mcst for the cost matrix c. Let $C(N, c)$ be the cost of a mcst. Let c^S be the restriction of the cost matrix c to the coalition $S_0 \subseteq N_0$. Let $C(S, c)$ be the

[2]This is an abuse of language, as some permutations might yield identical extreme core allocations.

cost of the mcst of the problem (S, c^S). Given these definitions, we say that C is the stand-alone cost function associated with c.

An allocation is a vector $x \in \mathbb{R}^N$ such that $\sum_{i\in N} x_i = C(N)$. For any $S \subseteq N$, let $x(S) = \sum_{i\in S} x_i$. Given $S \subseteq N$ and $x \in \mathbb{R}^N$, we denote as $x_S \in \mathbb{R}^S$ the restriction of x to $\mathbb{R}^S$.

For any cost matrix c, the associated cost game is given by (N, C) with $C(S) = C(S, c)$ for all $S \subseteq N$. We then say that C is a mcst game. We define the set of stable allocations as $Core(C)$. Formally, an allocation $x \in Core(C)$ if $x(S) \leq C(S)$ for all $S \subseteq N$.

3. A METHOD TO FIND EXTREME CORE ALLOCATIONS

We provide a method that allows to find extreme allocations of $Core(C)$. It is based on the concept of marginal games (Núñez and Rafels, 1998). We take an ordering of the agents and allocate to the last of them her marginal cost to join the grand coalition. We then update the stand-alone cost of remaining coalitions by giving them the option, instead of using their own stand-alone cost, to use their stand-alone cost with the removed player, net of her allocation. We then repeat the process by removing the next-to-last player, and so on.

Formally, for all $i \in N$ and all C, let $b_i^C = C(N) - C(N \setminus \{i\})$. For all $S \subseteq N$, the ith-marginal game $(N \setminus \{i\}, C^i)$ is given by $C^i(\emptyset) = 0$ and $C^i(S) = \min\left\{C\left(S \cup \{i\}\right) - b_i^C, C(S)\right\}$ for all $\emptyset \neq S \subseteq N \setminus \{i\}$.

We label $C^{ij}(S)$ to mean $\left(C^i\right)^j(S)$, $C^{ijk}(S)$ to mean $\left(C^{ij}\right)^k(S)$ and so on.

Let $\Pi(N)$ be the set of permutations of N. Let $\pi = (\pi_1, \ldots, \pi_n) \in \Pi$. We define the reduced marginal cost vector of C related to permutation π, denoted as $y^{r\pi}(C)$, or simply $y^{r\pi}$, as follows:

$$\begin{aligned}
y^{r\pi}_{\pi_n} &= C(N) - C(N \setminus \{\pi_n\}) = b^C_{\pi_n} \\
y^{r\pi}_{\pi_{n-1}} &= C^{\pi_n}(N \setminus \{\pi_n\}) - C^{\pi_n}(N \setminus \{\pi_{n-1}, \pi_n\}) = b^{C^{\pi_n}}_{\pi_{n-1}} \\
&\vdots \\
y^{r\pi}_{\pi_2} &= C^{\pi_n\pi_{n-1}\ldots\pi_{n-3}}(\{\pi_1, \pi_2\}) - C^{\pi_n\pi_{n-1}\ldots\pi_{n-3}}(\pi_1) = b^{C^{\pi_n\pi_{n-1}\cdots\pi_{n-3}}}_{\pi_2} \\
y^{r\pi}_{\pi_1} &= C^{\pi_n\pi_{n-1}\ldots\pi_{n-2}}(\{\pi_1\}) = b^{C^{\pi_n\pi_{n-1}\cdots\pi_{n-2}}}_{\pi_1}.
\end{aligned}$$

Núñez and Rafels (1998) provide a sufficient condition for $y^{r\pi}$ to be an extreme point of the core, and in fact for the set $(y^{r\pi})_{\pi\in\Pi}$ to be the set of extreme points of the core. The sufficient condition is that of almost-concavity of the cost game: $C(S)+C(T) \geq C(S\cup T)+C(S\cap T)$ for all $S, T \subset N$ such that $S \cup T \neq N$. We thus have all concavity conditions except those involving the grand coalition. We show with the following example that this condition is not satisfied by games generated by mcst problems.

Example 1 *Let $N = \{1, 2, 3, 4\}$ and c be as described in the following (i horizontally, j vertically) and illustrated in Figure 1:*

c_{ij}	*1*	*2*	*3*	*4*
0	*1*	*6*	*5*	*5*
1		*6*	*4*	*2*
2			*5*	*5*
3				*5*

It is easy to see that $C(\{2\}) = 6$, $C(\{2, 3\}) = C(\{2, 4\}) = 10$ and $C(\{2, 3, 4\}) = 15$. We thus have that $C(\{2, 3\}) + C(\{2, 4\}) < C(\{2, 3, 4\}) + C(\{2\})$, which contradicts the almost-concavity condition.

It turns out that we can interpret a reduced game as a particular type of source connection problem. A *mcst problem with priced nodes* is a tupla (N, P, y, c) where $P \subseteq N$ are nodes that do not need to be connected and $y \in \mathbb{R}^P$ is the vector whose coordinates are the prices that nodes in P pay (or receive, when negative) if they are actually connected. Nodes in P are called *priced nodes*. Hence, the cost of (N, P, y, c) is defined as

$$C(N, P, y, c) = \min_{T\subseteq P}\{C((N \setminus P) \cup T) - y(T)\}$$

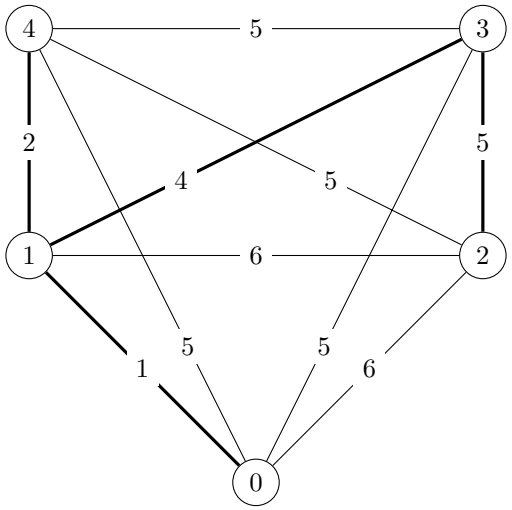


Figura 1: Example of a minimum cost spanning tree problem.

and the cost of a subset $S \subseteq N \setminus P$ is given by

$$C(S, P, y, c) = \min_{T \subseteq P} \{C(S \cup T) - y(T)\}.$$

In particular, $C(N, P, y, c) = C(N \setminus P, P, y, c)$.

As usual, the *core* of a mcst problem with priced nodes (N, P, y, c) is the set of allocations $x \in \mathbb{R}^{N \setminus P}$ that satisfy $x_i(N \setminus P) = C(N, P, y, c)$ and $x_i(S) \leq C(S, P, y, c)$ for all $S \subseteq N \setminus P$.

Notice that mcst problems with priced nodes generalize both mcst problems (when $P = \emptyset$) and minimum cost Steiner tree problems (when $y_i = 0$ for all $i \in P$).

For simplicity, when $y \in \mathbb{R}^N$, we write (N, P, y, c) instead of (N, P, y_P, c). Given $\pi \in \Pi(N)$ and $i \in N$, let $P^{\pi i} = \{\pi_i, \ldots, \pi_n\}$ be the set of nodes that come after π_i (including π_i) in the order π. It is clear that $C^{\pi_n \cdots \pi_i}(S) = C\left(S, P^{\pi i}, y^{r\pi}, c\right)$ for all $S \subseteq N \setminus P^{\pi i}$.

This correspondence allows us to prove that $y^{r\pi}$ belongs to the core.

Theorem 1 *For any mcst game (N, C) and any permutation $\pi \in \Pi(N)$, $y^{r\pi} \in Core(C)$.*

Proof. Given that $C^{\pi_n \cdots \pi_i}(S) = C\left(S, P^{\pi i}, y^{r\pi}, c\right)$ for all $S \subseteq N \setminus P^{\pi i}$, it suffices to prove that $y^{r\pi}_{N \setminus P^{\pi i}}$ belongs to the core of $(N, P^{\pi i}, y^{r\pi}, c)$.

Let $S \subseteq N \setminus P^{\pi i}$ and let t^S be a tree that connects all the nodes in S to the source using a set $T^S \subseteq P^{\pi i}$ of priced nodes. We need to prove $y^{r\pi}(S) \leq \sum_{e \in t^S} c_e - y^{r\pi}\left(T^S\right)$ or, equivalently,

$$y^{r\pi}\left(S \cup T^S\right) \leq \sum_{e \in t^S} c_e. \tag{1}$$

Assume there exists $t^* \in T^*(c)$ that is also an optimal tree in $(N, N, y^{r\pi}, c)$. Hence

$$\sum_{e \in t^*} c_e - y^{r\pi}(N) \leq \sum_{e \in t^S} c_e - y^{r\pi}\left(S \cup T^S\right). \tag{2}$$

Since $\sum_{e \in t^*} c_e = y^{r\pi}(N)$, we have that (1) and (2) coincide.

We still need to prove that there exists $t^* \in T^*(c)$ that is also an optimal tree in $(N, N, y^{r\pi}, c)$. We will prove a stronger result: that *any* $t^* \in T^*(c)$ is an optimal tree in *any* $(N, P^{\pi i}, y^{r\pi}, c)$. Assume w.l.o.g. $\pi = 1 \ldots n$. For notational convenience, let $P^{\pi 0} = \emptyset$. We proceed by induction on i. For $i = 0$, the result holds trivially. Assume the result holds for $i - 1 \geq 0$. Let $t^* \in T^*(c)$. We have to prove that t^* is an optimal tree on $(N, P^{\pi i}, y^{r\pi}, c)$, i.e. for each tree t that connects all the nodes in $(N \setminus P^{\pi i}) \cup T$ to the source, with $T \subseteq P^{\pi i}$, it holds that

$$\sum_{e \in t^*} c_e - y^{r\pi}\left(P^{\pi i}\right) \leq \sum_{e \in t} c_e - y^{r\pi}(T). \tag{3}$$

By induction hypothesis, t^* is an optimal tree in $\left(N, P^{\pi i-1}, y^{r\pi}, c\right)$. Assume first $i \in T$. Then, $\sum_{e\in t^*} c_e - y^{r\pi}\left(P^{\pi i-1}\right) \leq \sum_{e\in t} c_e - y^{r\pi}\left(T \setminus \{i\}\right)$, which is equivalent to (3). Assume now $i \notin T$. Hence

$$\begin{aligned}\sum_{e\in t^*} c_e - y^{r\pi}\left(P^{\pi i}\right) &= \sum_{e\in t^*} c_e - y^{r\pi}\left(P^{\pi i-1}\right) - y_i^{r\pi} \\ &= C\left(N, P^{\pi i-1}, y^{r\pi}, c\right) - y_i^{r\pi} \\ &= C\left(N \setminus P^{\pi i-1}, P^{\pi i-1}, y^{r\pi}, c\right) - y_i^{r\pi}.\end{aligned}$$

By definition, $y_i^{r\pi} = C\left(N \setminus P^{\pi i-1}, P^{\pi i-1}, y^{r\pi}, c\right) - C\left(N \setminus P^{\pi i}, P^{\pi i-1}, y^{r\pi}, c\right)$ and hence the above expression equals $C\left(N \setminus P^{\pi i}, P^{\pi i-1}, y^{r\pi}, c\right)$ which, by definition, is less or equal than $\sum_{e\in t} c_e - y^{r\pi}(T)$. ■

4. THE SET OF EXTREME CORE ALLOCATIONS

We now have shown that our procedure generates allocations in the core. By definition, they are also extreme core allocations, defined as follows:

Definition 1 *An allocation $y \in Core(C)$ is an extreme core allocation if there does not exist $y', y'' \in Core(C)$ and $\lambda \in (0,1)$ such that $y = \lambda y' + (1-\lambda)y''$.*

We define the set of extreme core allocations as $ExtCore(C)$. We next show that the core allocations defined in the previous section actually constitute the whole set of extreme core allocations. We proceed by first showing that for any coalition, we attain the maximal allocation compatible with the core. We need to prove two intermediary results before getting to the main result of this section.

First, we define the cost function $\widehat{C}$ and show that it has the same core as C and that a coalition S cannot receive more than $\widehat{C}(S)$ in any core allocation.

For all $c \in \Gamma$, let $(N, \widehat{C})$ be defined in the following way:

Step 0: $\widehat{C}(S) = C(S)$ for all S such that $|S| \geq n-1$.

Step k: $\widehat{C}(S) = \min\left\{C(S), \min_{T\subset N\setminus S} \widehat{C}(S\cup T) + \widehat{C}(N\setminus T) - \widehat{C}(N)\right\}$ for all S such that $|S| = n-1-k$, for $k = 1, \ldots, n-2$.

Step $n-1$: $\widehat{C}(\emptyset) = 0$.

Hence, $\widehat{C}(N) = C(N)$ and $\widehat{C}(S) = \min_{T\subset N\setminus S}\left\{\widehat{C}(S\cup T) + \widehat{C}(N\setminus T) - C(N)\right\}$ for all $S \subset N$.

To compute the alternative stand-alone cost $\widehat{C}(S)$, we let S pick a partner T and compute the sum of the costs of S with T and $N \setminus T$, to which we substract the cost of the grand coalition.

Lemma 1 *For all $c \in \Gamma$, $Core(\widehat{C}) = Core(C)$.*

Proof. Given that $\widehat{C}(S) \leq C(S)$ for all $S \subset N$ and $\widehat{C}(N) = C(N)$, it is obvious that $Core(\widehat{C}) \subseteq Core(C)$. We show that $Core(C) \subseteq Core(\widehat{C})$. Suppose, on the contrary, that there exists $y \in Core(C)$ and $y \notin Core(\widehat{C})$. Then, there exists $S \subset N$ such that $\widehat{C}(S) < y(S) \leq C(S)$. Stated otherwise, there exists $\varnothing \neq T \subset N \setminus S$ such that

$$\widehat{C}(S) = \widehat{C}(S\cup T) + \widehat{C}(N\setminus T) - \widehat{C}(N) < y(S) \leq C(S).$$

Suppose first that $|S| = n-2$. Then,

$$\widehat{C}(S) = C(S\cup T) + C(N\setminus T) - C(N) < y(S) \leq C(S) \tag{4}$$

as $|S\cup T|, |N\setminus T| \geq n-1$. Since $y \in Core(C)$, we must have that $y(S)+y(T)+y(N\setminus(S\cup T)) = C(N)$. We have the following core conditions:

$$\begin{aligned} y(S)+y(T) &\leq C(S\cup T) \\ y(S)+y(N\setminus(S\cup T)) &\leq C(N\setminus T).\end{aligned}$$

Adding these two constraints, we obtain

$$\begin{aligned} 2y(S) + y(T) + y(N \setminus (S \cup T)) &\leq C(S \cup T) + C(N \setminus T) \\ y(S) &\leq C(S \cup T) + C(N \setminus T) - C(N) = \widehat{C}(S). \end{aligned}$$

Therefore, we have a contradiction with (4).

We have thus shown that for $y \in Core(C)$, we need $y(S) \leq \widehat{C}(S)$ if $|S| = n-2$. We can then use a recursive argument to show that if $y \in Core(C)$ implies that $y(S) \leq \widehat{C}(S)$ for all S such that $|S| \geq n-k$, it also implies that $y(S) \leq \widehat{C}(S)$ for all S such that $|S| = n-k-1$. Thus, $y(S) \leq \widehat{C}(S)$ for all $S \subseteq N$ is a necessary condition for $y \in Core(C)$ and $Core(C) \subseteq Core(\widehat{C})$. ■

Our next step is to show that for any S, there exists at least one permutation π such that $y^{r\pi}(S) = \widehat{C}(S)$. To do so, we once again use the mcst with priced nodes representation.

We know, from the proof of Theorem 1, that $y^{r\pi}_{N \setminus P^{\pi i}}$ belongs to the core of $\left(N, P^{\pi i}, y^{r\pi}, c\right)$ and, moreover, that *any* $t^* \in T^*(c)$ is an optimal tree in *any* $\left(N, P^{\pi i}, y^{r\pi}, c\right)$.

For simplicity, and since N is fixed, we write Π instead of $\Pi(N)$.

Let $t^* \in T^*(c)$. We denote $t^* = \{(i, i^*)\}_{i \in N}$, where i^* is the predecessor of node i in t^*, i.e. i^* is the adjacent node to node i in the (unique) path in t^* from node i to the source.

We then define $i \preceq_* j$ as the partial relation in N given by "i precedes j in t^*", that is, $i \preceq_* j$ iff $j \in F^{*i}$, where F^{*i} is the set of followers of node i in t^* (including node i).

Given $P \subseteq N$ and $\pi \in \Pi$, we say that π is t^*-compatible with P if the following two conditions hold:

External t^*-compatibility Nodes in $S = N \setminus P$ that follow nodes in P in t^* are first in π:

$$\left.\begin{array}{l} \pi_i \preceq_* \pi_j \\ \pi_i \in P \\ \pi_j \in S \end{array}\right\} \Longrightarrow j < i.$$

Internal t^*-compatibility Nodes inside P follow in π the partial order $\preceq_*$:

$$\left.\begin{array}{l} \pi_i \preceq_* \pi_j \\ \pi_i, \pi_j \in P \end{array}\right\} \Longrightarrow i \leq j.$$

It is clear that there always exists an order t^*-compatible with P. For example, given $P \subseteq N$, $S = N \setminus P$ and $\pi \in \Pi$, we define $\pi^{*P} \in \Pi$ inductively as follows: If $P = \emptyset$, then $\pi^{*P} = \pi$. If $P = \{i\}$, then $\pi^{*P} = [\pi_S, i]$ is the order whose first components are the nodes in S (following order π) and the last component is node i. Assume we have defined $\sigma^{*Q} \in \Pi$ for any $\sigma \in \Pi$ and $Q \subset N$ with $|Q| < |P|$. Let $p_{\pi 1} \in P$, or p_1 for short, be the first node in π satisfying $p_1 \in P$ and $p_1^* \notin P$, where p_1^* is the adjacent node to p_1 in the (unique) path in t^* from p_1 to the source. Then, we apply the induction to define $\pi^{*P} = \left[\pi_{N \setminus \{p_1\}}, p_1\right]^{*P \setminus \{p_1\}}$.

Example 2 *We revisit Example 1, with* $P = \{1, 3, 4\}$.

Take $t^* = \{(0,1), (1,4), (1,3), (3,2)\}$. *External* t^**-compatibility implies that nodes* 1 *and* 3 *should go after node* 2. *Internal* t^**-compatibility implies that nodes* 3 *and* 4 *should go after node* 1. *These two conditions together give two unique* t^**-compatible orders:* $[2, 1, 4, 3]$ *and* $[2, 1, 3, 4]$.

If we choose an initial π *in which* 3 *has preference over* 4, *we obtain* $\pi^{*P} = [2, 1, 3, 4]$. *Otherwise, we obtain* $\pi^{*P} = [2, 1, 4, 3]$. *In both cases,* $y^{r\pi^{*P}} = (0, 6, 4, 2)$.

The same happens if we take $t^* = \{(0,1), (1,3), (1,4), (4,2)\}$ *instead.*

We say that π is c-compatible with P if there exists some $t \in T^*(c)$ such that π is t-compatible with P.

In the previous example, $y^{r\pi}(\{2\}) = \widehat{C}(\{2\}) = 6$ for each π c-compatible with $\{1, 3, 4\}$. We now show that this holds in general.

Theorem 2 *Given* $S \subseteq N$, *we have* $y^{r\pi}(S) = \widehat{C}(S)$ *for all* $\pi \in \Pi$ *order* c*-compatible with* $N \setminus S$.

Proof. Fix $t^* = \{(i, i^*)\}_{i \in N} \in T^*(c)$. Given $S \subseteq N$ and $P = N \setminus S$, let $\pi \in \Pi$ be an order t^*-compatible with P. We prove that the following two statements hold:

(I) Either $y^{r\pi}(S) = C(S)$ or there exists $\emptyset \neq T \subset P$ such that

(Ia) $y^{r\pi^a}(T) = y^{r\pi}(T)$ for all $\pi^a \in \Pi$ order t^*-compatible with T, and

(Ib) $y^{r\pi^b}(P \setminus T) = y^{r\pi}(P \setminus T)$ for all $\pi^b \in \Pi$ order t^*-compatible with $P \setminus T$.

(II) $y^{r\pi}(S) = \widehat{C}(S)$.

We proceed by induction on $|P|$. For $P = \emptyset$, both statements hold trivially. Assume now both statements hold when $|P| < \alpha$ and suppose $|P| = \alpha$ for some $\alpha > 0$.

We first prove statement (I). Let $p_1 = \pi_s \in P$ be the first element in P according to π (that is, $i < s$ implies $\pi_i \notin P$). Hence, for all $\pi_i \in P \setminus \{p_1\}$, we have $s < i$. Under internal t^*-compatibility, we deduce $p_1 \notin F^{*i}$ for all $i \in P \setminus \{p_1\}$.

Let $S' = S \cup \{p_1\}$ and $P' = P \setminus \{p_1\}$. By applying t^*-compatibility, it is straightforward to check that

$$y^{r\pi}_{p_1} = C\left(S', P', y^{r\pi}, c\right) - C\left(S, P', y^{r\pi}, c\right). \tag{5}$$

Let t' be an optimal tree in $(N \setminus \{p_1\}, P', y^{r\pi}, c)$. That is:

$$\begin{array}{rcl} S & \subseteq & t'(N) \\ p_1 & \notin & t'(N) \\ c(t') - y^{r\pi}(t'(N) \cap P') & = & \min\limits_{T' \subseteq P'} \{C(S \cup P') - y^{r\pi}(T')\} \end{array}$$

where $t'(N) \subseteq N \setminus \{p_1\}$ is the set of nodes that connect to the source through t'.

In case there are more than one tree satisfying the above conditions, we take t' such that $|t'(N)|$ is maximal among them.

Assume first $S = t'(N)$, which means $t'(N) \cap P' = \emptyset$ and $C(t') = C(S)$. Hence

$$\begin{array}{rcl} y^{r\pi}(S) & = & y^{r\pi}(S') - y^{r\pi}_{p_1} \\ & \overset{(5)}{=} & y^{r\pi}(S') - C(S', P', y^{r\pi}, c) + C(S, P', y^{r\pi}, c) \\ & = & y^{r\pi}(S') - y^{r\pi}(S') + c(t') = c\left(t'\right) = C(S) \end{array}$$

and so statement (I) holds.

Assume now $S \subset t'(N)$. Let $T' = t'(N) \setminus S$. It is straightforward to check that $\emptyset \neq T' \subset P$. We will prove that (I$a$) and (I$b$) hold with this T'. As a previous step, we need to prove the following Claim:

Claim A: For all $i \in N \setminus \{p_1\}$, $i^* \in S \cup T' \Rightarrow i \in S \cup T'$.

Assume, on the contrary, that there exists some $i \in N \setminus \{p_1\}$ such that $i^* \in S \cup T'$ and $i \notin S \cup T'$. That is, $i^* \in t'(N)$ and $i \notin t'(N)$.

Let $Q = \left\{j \in N \setminus T' : \tau^*_{ij} \subseteq N \setminus T'\right\}$, where τ^*_{ij} is the (unique) path form node i to node j in t^*. Thus, Q is the set of followers of node i without leaving $P \setminus T'$. Hence, $Q \subseteq P \setminus T'$. Since $p_1 \notin F^{*j}$ for all $j \in P'$, we deduce that $p_1 \notin Q$. Let $R = \{j \in N : j^* \in Q\}$ be the set of nodes in S that immediately follow some node in Q (case $R = \emptyset$ is also possible). Denote $R = \{r_1, \ldots, r_k\}$. We define the set of edges $E = \{e_1, \ldots, e_k\} \subset N^p$ inductively as follows: Let $e_1 = (i_1, i'_1) \in N^p$ be the first edge in the (unique) path in t' from r_1 to the source such that $i_1 \in F^{*r_1}$ and $i'_1 \notin F^{*r_1}$. Let $e_2 = (i_2, i'_2) \in N^p$ be the first edge in the (unique) path in t' from r_2 to the source such that $i_2 \in F^{*r_1} \cup F^{*r_2}$ and $i'_2 \notin F^{*r_1} \cup F^{*r_2}$. Let $e_3 = (i_3, i'_3) \in N^p$ be the first edge in the (unique) path in t' from r_3 to the source such that $i_3 \in F^{*r_1} \cup F^{*r_2} \cup F^{*r_3}$ and $i'_3 \notin F^{*r_1} \cup F^{*r_2} \cup F^{*r_3}$. And so on. Thus, E is the set of edges that connect R to the source in t'. Now,

$$t^1 = (t^* \setminus \{(j, j^*)\}_{j \in Q \cup R}) \cup E \tag{6}$$

is a tree that connects all the nodes in S to the source using nodes in $P \setminus Q$. Since t^* is optimal in $(S, P, y^{r\pi}, c)$, we deduce that $c(t^*) - y^{r\pi}(P) \leq c(t^1) - y^{r\pi}(P \setminus Q)$ or, equivalently,

$$c(t^*) \leq c(t^1) + y^{r\pi}(Q). \tag{7}$$

Now,

$$t^2 = (t' \setminus E) \cup \{(j, j^*)\}_{j \in Q \cup R} \tag{8}$$

is a tree that connects all the nodes in S to the source using nodes in $T' \cup Q \subseteq P'$. Since t' is optimal in $(S, P', y^{r\pi}, c)$, we deduce that

$$c(t') - y^{r\pi}(T') \leq c(t^2) - y^{r\pi}(T' \cup Q)$$

or, equivalently,

$$c(t') \leq c(t^2) - y^{r\pi}(Q). \tag{9}$$

By applying (7) and (9),

$$\begin{array}{rcl} c\left(t^2\right) & \overset{(9)}{\geq} & c(t') + y^{r\pi}(Q) \overset{(7)}{\geq} c(t') + c(t^*) - c(t^1) \\ & \overset{(6)}{=} & c(t') + c(t^*) - c(t^*) + c(\{(j, j^*)\}_{j \in Q \cup R}) - c(E) \\ & = & c\left(t'\right) + c\left(\{(j, j^*)\}_{j \in Q \cup R}\right) - c\left(E\right) \\ & \overset{(8)}{=} & c\left(t^2\right). \end{array}$$

This implies $c\left(t^2\right) = c\left(t'\right) + y^{r\pi}(Q)$, so $c\left(t^2\right) - y^{r\pi}(T' \cup Q) = c\left(t'\right) - y^{r\pi}(T')$ and hence t^2 is also an optimal tree in $(S, P', y^{r\pi}, c)$ with $|t^2(N)| = |t'(N)| + |Q|$, which contradicts that $|t'(N)|$ is maximum among these optimal trees (notice that $i \in Q \neq \emptyset$). This contradiction completes the proof of Claim A.

We can now prove that (Ia) holds with T'. Under the induction hypothesis on statement (II), we have $y^{r\pi^a}(N \setminus T') = \widehat{C}(N \setminus T')$ for all $\pi^a \in \Pi$ order t^*-compatible with T'. On the other hand, if π is t^*-compatible with T', then we can also apply the induction hypothesis on (II) to deduce that $y^{r\pi}(N \setminus T') = \widehat{C}(N \setminus T')$. Hence, it is enough to prove that π is indeed t^*-compatible with T'. We check both external and internal t^*-compatibility.

External t^*-compatibility: Let i, j such that $\pi_i \preceq_* \pi_j$, $\pi_i \in T'$ and $\pi_j \in N \setminus T'$. We have three cases: If $\pi_j \in S$, then $j < i$ because π is externally t^*-compatible with P. If $\pi_j = p_1$, then $\pi_i \preceq_* p_1$ implies $s < i$ because $\pi_i \in T' \subset P$ and $p_1 = \pi_s$ is the first element of P in π; by internal t^*-compatibility with P, we deduce $\pi_i \not\preceq_* \pi_s = \pi_j$, which is a contradiction. Finally, if $\pi_j \in P \setminus (T' \cup \{p_1\})$, then $\pi_j \notin S \cup T' \cup \{p_1\}$; under Claim A, we have $\pi_j^* \notin S \cup T'$; by applying Claim A iteratively, and since $\pi_i \preceq_* \pi_j$, we conclude that $\pi_i \notin S \cup T'$, which is a contradiction because $\pi_i \in T'$.

Internal t^*-compatibility: Let i, j such that $\pi_i \preceq_* \pi_j$ and $\pi_i, \pi_j \in T'$. Since $T' \subset P$, we have $\pi_i, \pi_j \in P$. Hence, $i \leq j$ because π is internally t^*-compatible with P.

We now prove that (Ib) holds with T'. By definition of $y^{r\pi}_{p_1}$, both t^* and t' are optimal trees in $(N, P, y^{r\pi})$. Hence, $C(N) - y^{r\pi}(P) = C(S \cup T') - y^{r\pi}(T')$. Equivalently,

$$y^{r\pi}(P \setminus T') = C(N) - C(S \cup T'). \tag{10}$$

Let $\pi^b \in \Pi$ be an order t^*-compatible with $P \setminus T'$. We check that $y^{r\pi^b}(P \setminus T') = C(N) - C(S \cup T')$. By the induction hypothesis on statement (II), we can assume that nodes in $P \setminus T'$ follow the same order as in π. In particular, p_1 is still the first node in $P \setminus T'$ under π^b. Denote $P \setminus T' = \{p_1, \ldots, p_L\}$ following order π^b, i.e. when $p_i = \pi^b_{i^b}$ and $p_j = \pi^b_{j^b}$, then $i \leq j \Leftrightarrow i^b \leq j^b$. Internal t^*-compatibility assures that $p_i \preceq_* p_j \Rightarrow i \leq j$. For each $l \in \{1, \ldots, L\}$, let $G^l = \{i \in S \cup T' : i^* = p_l\}$ be the set of nodes in $S \cup T'$ that immediately follow p_l in t^* (case $G^l = \emptyset$ is also possible), and denote $G^l = \{g^l_1, \ldots, g^l_{k_l}\}$. We define the set of edges $E^l = \{e^l_1, \ldots, e^l_{k_l}\} \subset N^p$ inductively as follows: Let $e^l_1 = (i_{l1}, i'_{l1}) \in N^p$ be the first edge in the (unique) path in t' from g^l_1 to the source such that $i_{l1} \in F^{*g^l_1}$ and $i'_{l1} \notin F^{*g^l_1}$. Let $e^l_2 = (i_{l2}, i'_{l2}) \in N^p$ be the first edge in the (unique) path in t' from g^l_2 to the source such that $i_{l2} \in F^{*g^l_1} \cup F^{*g^l_2}$ and $i'_{l2} \notin F^{*g^l_1} \cup F^{*g^l_2}$. And so on. Thus, E^l is the set

of nodes that connect each G^l to the source in t'. We now define the trees $t^0, t^1, \ldots, t^L$ inductively as $t^0 = t'$ and $t^l = \left(t^{l-1} \cup \{(i,i')\}_{i \in G^l \cup \{p_l\}}\right) \setminus E^l$ for each $l = 1, \ldots, L$. By optimality of t^* and t', each t^l is also optimal in $(N, \{p_{l+1}, \ldots, p_L\}, y^{r\pi}, c)$. Hence, we have

$$\begin{aligned}
y^{r\pi}_{p_L} &= c\left(t^L\right) - c\left(t^{L-1}\right) = C(N) - C(N \setminus \{p_L\}) \\
y^{r\pi}_{p_{L-1}} &= c\left(t^{L-1}\right) - c\left(t^{L-2}\right) = C(N \setminus \{p_L\}) - C(N \setminus \{p_L, p_{L-1}\}) \\
&\vdots \\
y^{r\pi}_{p_1} &= c\left(t^1\right) - c\left(t^0\right) = C(S \cup T' \cup \{p_1\}) - C(S \cup T')
\end{aligned}$$

from where we deduce

$$\begin{aligned}
y^{r\pi^b}_{p_L} &= C(N) - C(N \setminus \{p_L\}) \\
y^{r\pi^b}_{p_{L-1}} &= C(N \setminus \{p_L\}) - C(N \setminus \{p_L, p_{L-1}\}) \\
&\vdots \\
y^{r\pi^b}_{p_1} &= C(S \cup T' \cup \{p_1\}) - C(S \cup T')
\end{aligned}$$

and, adding up these terms, we get

$$y^{r\pi^b}(P \setminus T') = C(N) - C(S \cup T') \tag{11}$$

so that (Ib) comes from (10) and (11).

We now prove statement (II), i.e. $y^{r\pi}(S) = \widehat{C}(S)$. Since $y^{r\pi} \in Core(C) = Core\left(\widehat{C}\right)$, we have $y^{r\pi}(S) \leq \widehat{C}(S)$. When $y^{r\pi}(S) = C(S)$, we have $\widehat{C}(S) \leq y^{r\pi}(S)$ because $\widehat{C}(S) \leq C(S)$. When $y^{r\pi}(S) \neq C(S)$, under statement (I) there exits $\emptyset \neq T' \subset P$ satisfying (Ia) and (Ib). We now apply the induction hypothesis on statement (II) to deduce that for all $\pi^a \in \Pi$ order t^*-compatible with T', and $\pi^b \in \Pi$ order t^*-compatible with $P \setminus T'$, we have $\widehat{C}(N \setminus T') = y^{r\pi^a}(N \setminus T')$ and $\widehat{C}(S \cup T') = y^{r\pi^b}(S \cup T')$. Under (I$a$) and (I$b$), we have $y^{r\pi^a}(T') = y^{r\pi}(T')$ and $y^{r\pi^b}(P \setminus T') = y^{r\pi}(P \setminus T')$, respectively. Hence, $\widehat{C}(N \setminus T') = y^{r\pi}(N \setminus T')$ and $\widehat{C}(S \cup T') = y^{r\pi}(S \cup T')$. Thus,

$$\begin{aligned}
\widehat{C}(S) &= \min\left\{C(S), \min_{\emptyset \neq T'' \subset P}\left\{\widehat{C}(S \cup T'') + \widehat{C}\left(N \setminus T''\right) - C(N)\right\}\right\} \\
&\leq \min_{\emptyset \neq T'' \subset P}\left\{\widehat{C}\left(S \cup T''\right) + \widehat{C}\left(N \setminus T''\right) - C(N)\right\} \\
&\leq \widehat{C}(S \cup T') + \widehat{C}(N \setminus T') - C(N) \\
&= y^{r\pi}(S \cup T') + y^{r\pi}(N \setminus T') - C(N) \\
&= y^{r\pi}(S \cup T') + y^{r\pi}(N) - y^{r\pi}(T') - C(N) \\
&= y^{r\pi}(S).
\end{aligned}$$

concluding the proof. ■

We are now ready for the main result of this section. Let $\Delta(c)$ be the convex hull of the allocations $\left(y^{r\pi}\right)_{\pi \in \Pi}$.

Theorem 3 *For any mcst game (N, C), $ExtCore(C) = \Delta(c)$.*

Proof. We proceed by contradiction. Assume $\Delta(c) \neq ExtCore(C)$. Since each $y \in \Delta(c)$ is a core allocation, we have that there exists an extreme core allocation x that does not belong to $\Delta(c)$. From this, we deduce that there exists some S such that $x(S) - y(S)$ has the same (nonzero) sign for each allocation $y \in \Delta(c)$. That is, either $x(S) - y(S) > 0$ for all $y \in \Delta(c)$, or $x(S) - y(S) < 0$ for all $y \in \Delta(c)$.

Then, S should be one of the sets that determine a saturate constraint on a face of $\Delta(c)$. Since $x \notin \Delta(c)$, then it should lay inside the opposite side of the half-space.

We can assume $x(S) - y(S) > 0$ for all $y \in \Delta$ because, in the opposite, we instead consider $T = N \setminus S$.

We now take y such that $y(S) = \widehat{C}(S)$. Existence of such a $y \in \Delta$ is guaranteed by Theorem 2. Then, $x(S) > \widehat{C}(S)$ and thus, by Lemma 1, x does not belong to $Core(\widehat{C}) = Core(C)$. Hence the contradiction. ■

As a corollary, we now obtain a complete description of the core.

Corollary 1 *For all $c \in \Gamma$, $Core(C) = \Delta(c)$.*

5. LINK WITH COST SHARING SOLUTIONS

We show that while our method is new, it can be used to reinterpret the three most famous cost-sharing solutions in the literature on mcst.

A cost sharing solution (or rule) assigns a cost allocation $y(c)$ to any admissible cost matrix c. We start by building a solution from the allocations defined in the previous sections. Let $y^r = \sum_{\pi \in \Pi(N)} \frac{1}{n!} y^{r\pi}$ be the average of the reduced marginal cost vectors. By our previous section, y^r is the barycenter of $Core(C)$.

The Bird solution

The Bird solution is defined as follows. Let $\Pi^*(c)$ be the set of permutations obtained in Prim's algorithm. For all $\pi \in \Pi^*(c)$, let $y^{b,\pi}_{\pi_i} = \min_{k=0,\ldots,i-1} c_{\pi_k \pi_i}$ where $\pi_0 = 0$. The Bird solution is

$$y^b = \frac{1}{|\Pi^*(c)|} \sum_{\pi \in \Pi^*(c)} y^{b,\pi}.$$

Theorem 4 *For all $\pi \in \Pi^*(c)$, $y^{b,\pi} = y^{r\pi}$.*

Proof. Suppose $\pi \in \Pi^*(c)$ and consider agent π_n. We have

$$y^{r\pi}_{\pi_n} = C(N) - C(N \setminus \{\pi_n\}) = \min_{k=0,\ldots,n-1} c_{\pi_k \pi_i} = y^{b,\pi}_{\pi_n}.$$

Observe that by construction, $C(S \cup \{\pi_n\}) - C(S) = \min_{k \in S} c_{\pi_k \pi_i} \leq C(N) - C(N \setminus \{\pi_n\})$ for all $S \subseteq N \setminus \pi_n$. Thus, $C^{\pi_n}(S) = \min(C(S), C(S \cup \pi_n) + C(N \setminus \pi_n) - C(N)) = C(S)$ for all $S \subseteq N \setminus \{\pi_n\}$.

Consider agent π_j for $j = 1, \ldots, n-1$.

$$\begin{aligned} y^{r\pi}_{\pi_k} &= C^{\pi_n,\ldots,\pi_{j+1}}(\{\pi_1, \ldots, \pi_j\}) - C^{\pi_n,\ldots,\pi_{j+1}}(\{\pi_1, \ldots, \pi_{j-1}\}) \\ &= C(\{\pi_1, \ldots, \pi_j\}) - C\left(\{\pi_1, \ldots, \pi_{j-1}\}\right) \\ &= \min_{k=0,\ldots,j-1} c_{\pi_k \pi_j} \\ &= y^{b,\pi}_{\pi_j} \end{aligned}$$

Hence, $y^{b,\pi} = y^{r\pi}$. ■

The Bird allocation is thus a very special case of our method, as it only uses the reduced marginal cost vector of C related to permutation π if π is a permutation corresponding to an order in which we construct the mcst using Prim's algorithm.

We thus obtain an alternative proof of the stability of the Bird allocation. In addition, we can see that the permutations used for the Bird allocation are such that there are no modifications to do on C before computing the marginal cost vector. For other permutations, we will obtain that the corresponding marginal cost vector (without modifying C) is not in the core.

The cycle-complete solution

To define the cycle-complete solution, we need to define irreducible and cycle-complete cost matrices.

From any cost matrix c, we can define the irreducible cost matrix $\bar{c}$ as follows:

$$\bar{c}_{ij} = \min_{p_{ij} \in P_{ij}(N_0)} \max_{e \in p_{ij}} c_e \text{ for all } i, j \in N_0.$$

From any cost matrix c, we can define the cycle-complete cost matrix c^* as follows:

$$\begin{aligned} c^*_{ij} &= \max_{k \in N \setminus \{i,j\}} \overline{c_{ij}^{N \setminus \{k\}}} \text{ for all } i,j \in N \\ c^*_{0i} &= \max_{k \in N \setminus \{i\}} \overline{c_{ij}^{N \setminus \{k\}}} \text{ for all } i \in N. \end{aligned}$$

where $\overline{c_{ij}^{N \setminus \{k\}}}$ indicate the cost of edge (i,j) on the matrix that we first restricted to agents in $N \setminus \{k\}$ before transforming into an irreducible matrix.

The cycle complete matrix can also be defined using cycles (Trudeau, 2012): for edge (i,j), we look at cycles that go through i and j. If there is one such cycle such that its most expensive edge is cheaper than a direct connection on edge (i,j), we assign this cost to edge (i,j).

The cycle-complete solution y^{cc} is the Shapley value of (N,C).

Let Γ^e be the set of elementary cost matrices: for any $c \in \Gamma^e$ and any $i,j \in N_0$, $c_{ij} \in \{0,1\}$. We show that for elementary cost matrices, the cycle-complete solution corresponds to y^r.

Theorem 5 *For any elementary cost matrix $c \in \Gamma^e$, $y^r = y^{cc}$.*

Proof. Trudeau (2012) showed that for elementary problems, $Core(C) = Core(C^*)$, where (N,C^*) is the mcst game associated to c^*, and that (N,C^*) is concave. By the properties of the Shapley value for concave games, y^{cc} is the average over the set of permutations of incremental cost allocations, with each of them being an extreme core allocation. It is obvious that the incremental cost vector corresponding to order π is exactly $y^{r\pi}$. We thus have that $y^r = y^{cc}$. ■

An alternative explanation of the above result is that the changes from the original to the cycle-complete cost matrix are the same as those imposed by our method. If node i has two free distinct paths to node j, say with the help of S and T, she will obtain the cost savings with both coalitions. This will result in $\widehat{C}(\{i,j\}) = 0$, the same result as if we modified directly the matrix into a cycle-complete matrix.

The folk solution

The folk solution is the Shapley value of the cost game associated with the irreducible cost matrix $\bar{c}$ defined in the previous subsection. As for the cycle-complete solution, we show a link between our method and the folk solution in elementary mcst problems. To do so, we need to define the monotonically increasing version of C, denoted as C^+:

Definition 2 *For all $c \in \Gamma$, let (N,C^+) be defined in the following way:*

Step 0: $C^+(N) = C(N)$.

Step k: For all S such that $|S| = n-k$, $C^+(S,c) = \min(C(S,c), \min_{i \in N \setminus S} C^+(S \cup i), c)$ for $0 < k < n$.

Step n: $C^+(\emptyset) = 0$.

The transformation from C to C^+ guarantees that there is never a negative cost to add an agent to a coalition. We show that applying our method to C^+ for an elementary problem yields the folk solution.

Theorem 6 *For any elementary cost matrix $c \in \Gamma^e$, $y^r(C^+) = y^f$, where $y^r(C^+)$ is y^r defined from C^+ instead of C.*

Proof. Consider a connected component $T \subseteq N_0$. If $0 \in T$, then $C(T \setminus \{0\}) = 0$. Otherwise, $C(T \setminus \{0\}) = 1$. Suppose that for $S \in T \setminus \{0\}$, $C(S \setminus \{0\}) > C(T \setminus \{0\})$. In such a case, $\bar{c}$ is such that $\bar{c}_{ij} = 0$ for all $i,j \in T$. By monotonicity, we must have that $C(S \setminus \{0\}) = C(T \setminus \{0\})$ for all $S \subseteq T$. Therefore, in both cases, the changes are identical.

We next show that there are no other changes. If i,j belong to different connected components, say T_1 and T_2, then $\bar{c}_{ij} = c_{ij}$. We also have that for all $R \subseteq T_1$ and $S \subseteq T_2$, $C(R \cup S) = C(R) = C(S)$. Therefore, once we have made the cost function monotonic within the connected components, it will be monotonic over all coalitions, meaning that we do not need to make any other changes.

After the modifications, to obtain y^f we take the Shapley value of $(N, \bar{C})$, the cost game associated with $\bar{c}$, or equivalently, $\sum_{\pi \in \Pi(N)} \frac{1}{n!} y^{\pi}(\bar{C})$. Given that $(N, \bar{C})$ is concave (Bergantiños and Vidal-Puga (2007)), $y^{r\pi}(\bar{C}) = y^{\pi}(\bar{C})$ for all $\pi \in \Pi(N)$ and thus $y^r(C^+) = y^f$. ■

We thus obtain the new result that for elementary mcst problems, the folk solution is the barycenter of $Core(C^+)$. Since in elementary mcst problems moving from $Core(C)$ to $Core(C^+)$ is the same as restricting our attention to non-negative core allocations, we can thus say that for elementary mcst problems, the folk solution is the barycenter of the non-negative core.

Therefore, we obtain, for elementary cost matrices, a clear distinction between the folk and cycle-complete solutions, based on whether or not they disqualify core allocations where some agents receive strictly negative cost shares.

6. DISCUSSION

The result of the previous section on the folk and cycle-complete solutions do not hold for non-elementary cost matrices. For those, we can compute the folk and cycle-complete solutions by decomposing the cost matrix into a series of elementary cost matrices and summing up. While that approach is computationally advantageous, one of the disadvantage is that, in general, we have that $Core(C^*)$ is a strict subset of $Core(C)$ and $Core(\bar{C})$ is a strict subset of $Core(C^+)$; the cycle-complete and folk solutions are no longer the barycenters of, respectively, the core of C and the non-negative core of C.

If we are willing to forego the Piecewise Linearity property, we therefore can use y^r and $y^r(C^+)$ as non-piecewise linear extensions of, respectively, the cycle-complete and folk solutions.

REFERENCES

Bahel, E. and Trudeau, C. (2014) Stable lexicographic rules for shortest path games. Economic Letters, 125, 266–269.

Bergantiños, G. and Vidal-Puga, J. (2007) A fair rule in minimum cost spanning tree problems. J. Econ. Theory, 137, 326–352.

Bird, C. (1976) On cost allocation for a spanning tree: a game theoretic approach. Networks, 6, 335–350.

Davis, M. and Maschler, M. (1965) The kernel of a cooperative game. Naval Research Logistics Quarterly, 12, 223–259.

Feltkamp, V., Tijs, S., and Muto, S. (1994) On the irreducible core and the equal remaining obligations rule of minimum cost spanning extension prob- lems. Technical Report 106, CentER DP 1994 nr.106, Tilburg University, The Netherlands.

Granot, D. and Huberman, G. (1981) On minimum cost spanning tree games. Mathematical Programming, 21, 1–18.

Granot, D. and Huberman, G. (1984) On the core and nucleolus of minimum cost spanning tree problems. Mathematical Programming, 29, 323–347.

Hwang, F. and Richards, D. S. (1992) Steiner tree problems. Networks, 22(1), 55–89.

Núñez, M. and Rafels, C. (1998) On extreme points of the core and reduced games. Annals of Operations Research, 84, 121–133.

Núñez, M. and Rafels, C. (2003) Characterization of the extreme core allo- cations of the assignment game. Games Econ. Behav., 44(2), 311–331.

Pérez-Castrillo, D. and Sotomayor, M. (2002) A simple selling and buying procedure. J. Econ. Theory., 103, 461–474.

Skorin-Kapov, D. (1995) On the core of the minimum cost steiner tree game in networks. Annals of Operations Research, 57, 233–249.

Tijs, S., Borm, P., Lohmann, E., and Quant, M. (2011) An average lexi- cographic value for cooperative games. European Journal of Operational Research, 213(1), 210–220.

Trudeau, C. (2012) A new stable and more responsible cost sharing solution for mcst problems. Games Econ. Behav., 75(1), 402–412.

Vidal-Puga, J. (2004) Bargaining with commitments. International Journal of Game Theory, 33(1), 129–144.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

A monotonic and merge-proof rule in minimum cost spanning tree situations

María Gómez-Rúa[1] and Juan Vidal-Puga[1]

[1] Research Group in Economic Analysis. Universidade de Vigo, Spain.

ABSTRACT

We present a new model for cost sharing in minimum cost spanning tree problems, so that the planner can identify the agents that merge. Under this new framework, and as opposed to the traditional model, there exist rules that satisfy *merge-proofness*. Besides, by strengthening this property and adding some other properties, such as *population-monotonicity* and *solidarity*, we characterize a unique rule that coincides with the weighted Shapley value of an associated cost game.

Keywords: Minimum cost spanning tree problems, cost sharing, core selection, cost-monotonicity, merge-proofness, weighted Shapley value.

1. INTRODUCTION

Problems in which supply network formation is involved have been widely studied in the literature from two different points of view. On one hand, the design of efficient algorithms and the computational complexity. On the other hand, the cost sharing of the network.

In situations where the supply can be done either directly or though other nodes, the optimal network is a tree. Examples include water supply, electricity, cable television, Internet, and so on.

In order to study these situations, two steps should be considered: in the first one, agents will construct a minimal cost tree, *mt* for short. There exist several algorithms for building a *mt*. Two of them are provided by Kruskal (1956) and Prim (1957), respectively. Once we have constructed a *mt*, the second step consists of dividing its cost between the nodes that benefit from it. In order to do that, a rule will be used.

In this paper, we focus on the cost sharing aspect. In particular, we study minimum cost spanning tree (*mcst*) situations. A group of agents, located at different geographical points, want some particular service or good which can only be provided by a common supplier, called the *source*. Agents can be served through connections which entail some cost and they do not care about whether they are connected directly or indirectly to the source.

There is a large literature on the cost-sharing related problem. Bird (1976) uses Prim's algorithm and proposes a cost allocation rule, the so-called Bird rule. He also associates a cooperative game with each *mcst* problem. Granot and Huberman (1981, 1984) study the core and the nucleolus of this cooperative game. Feltkamp et al. (1994) use Kruskal's algorithm in order to define a rule, the folk rule, which has been redefined and characterized in Bergantiños and Vidal-Puga (2007a,b). The definition of the folk rule in Bergantiños and Vidal-Puga (2007b) is as the Shapley value of a particular game associated to the problem. Other authors also use the Shapley value in order to define rules. Kar (2002) studies the Shapley value of the cooperative game defined by Bird (1976). Trudeau (2012) defines the cycle-complete rule as the Shapley value of a another particular game. Dutta and Kar (2004) propose another different rule.

A relevant approach is to study which desirable properties are satisfied by the different rules. Properties should help a planner to compare different rules and to decide which rule is preferred in a particular case.

In this paper, we focus on three important classes of properties. The first class is based on the property of *core selection*. This property states that no group of agents should subsidize the other agents, paying more than the cost of connecting themselves to the source. This is a very relevant

property in the economic literature. A typical drawback is that, in general, the core may be empty. However, it happens to be always non-empty in the game associated to a *mcst* problem. In fact, most of the rules proposed in the literature satisfy core-selection, with the remarkable exception of the one derived from the Shapley value of the associated game (Kar, 2002). A stronger version of core selection is *population monotonicity* (Thomson, 1983b,a), which requires that the cost allocated to any agent will not decrease if new agents join the society. Population monotonicity in *mcst* problems has been studied by Bergantiños and Gómez-Rúa (2010), Bergantiños and Vidal-Puga (2007a, 2009), Bogomolnaia and Moulin (2010), Lorenzo and Lorenzo-Freire (2009), and Norde et al. (2004).

The second class is based on the property of *cost monotonicity*. This property implies that the cost allocated to some node will not increase if the cost of a link involving this node goes down, nothing else changing. Hence, a violation of this appealing property could disincentive the agents to reduce the costs of constructing links (Dutta and Kar, 2004). A stronger version of cost monotonicity requires that the cost allocated to any agent will not increase if the cost of any link (involving this player or not) goes down, nothing else changing. Hence, this strong version would also prelude the agents to sabotage the construction of any link. This stronger version of cost monotonicity in *mcst* problems has been studied, among others, in Bergantiños and Gómez-Rúa (2010), Bergantiños and Vidal-Puga (2007a, 2009), Bogomolnaia and Moulin (2010), and Trudeau (2012).

There have been studied several rules that satisfy both population monotonicity and cost monotonicity. In particular, the folk rule is the only of these rules that satisfies both properties (Bergantiños and Vidal-Puga, 2008). For a characterization of rules satisfying both properties, see Bergantiños and Vidal-Puga (2015).

The third class is based on the properties of *split* and *merge-proofness*. The split-proofness property implies that one node should not have incentives to split into two or more different nodes. The merge-proofness property implies that two or more nodes should not have incentives to merge into a single node. These properties are relevant in situations where the identity of the nodes is ambiguous. For example, different departments on a University campus can be already connected by an internal network. In case they want to be connected to a wider supply network, should they be considered as a single node (the campus) or as several different nodes (the departments) connected at zero cost? Other examples are the different shops at the mall, apartments on a building, or houses in a residential area.

Merge-proofness preludes the agents to build an inefficient network, as next example shows:

Example 0.1 *Consider three agents located respectively at nodes* 1*,* 2 *and* 3*. The connection costs are depicted in Figure 1 (left). Figure 1 (right) represents the same situation after agents* 2 *and* 3 *merge, paying cost* 8*. This merging is inefficient, because the three agents end up paying no less than* $8 + 28 = 36$*, whereas the initial problem has an optimal tree with cost* 34*. Assume a rule assigns* $\Phi = (\Phi_1, \Phi_2, \Phi_3)$ *in the initial problem, and* $\Phi' = (\Phi'_1, \Phi'_2)$ *in the second one. Then, a merge-proof rule should satisfy* $\Phi_2 + \Phi_3 \leq 8 + \Phi'_2$*. Inefficiency due to a previous merge of nodes* 2 *and* 3 *is then avoided.*

The properties of split and merge-proofness have been studied in many contexts related to cost sharing problems. In a context where the agents consume arbitrary quantities of possibly different goods, Sprumont (2005) characterizes the Aumann-Shapley cost sharing method, used to distribute the total generated cost that ensures that agents never find it profitable to split or to merge their consumptions. O'Neill (1982), Chun (1988), De Frutos (1999), Moulin (2002), Ju (2003), and Ju et al. (2007) study split and merge-proofness in claim and bankruptcy problems. Moulin (2007, 2008) also studies split and merge-proofness in the context of job scheduling. Merging in exchange economies has been studied for several solution concepts by Hart (1974), Postlewaite and Rosenthal (1974), Maschler (1976), Legros (1987) and Rosenmüller and Sudhölter (2004). Merging and splitting in cooperative games have also been studied by Knudsen and Østerdal (2012) and references herein.

In the context of *mcst* problems, the folk rule satisfies split-proofness and, moreover, it is not difficult to derive a split-proof rule from a cost-monotonic one. However, this is not the case with merge-proofness. Under domain restrictions[1], the Bird rule satisfies merge-proofness (Özsoy,

[1]For example, assuming all the costs are different.

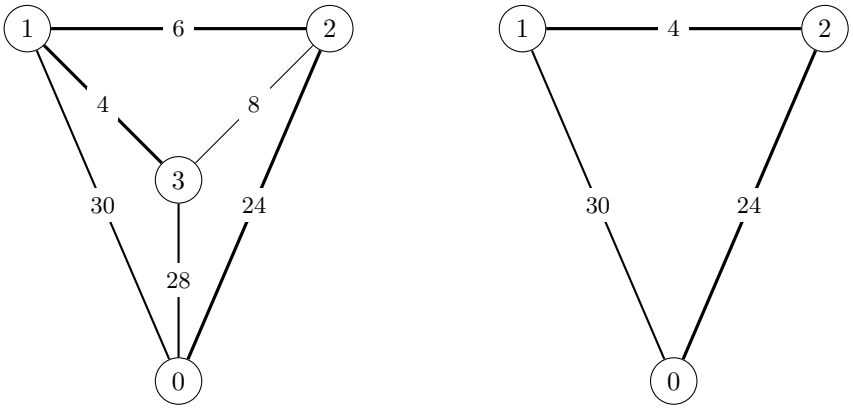


Figure 1: Merging of nodes 2 and 3.

2006; Athanassoglou and Sethuraman, 2008; Gómez-Rúa and Vidal-Puga, 2011). The Bird rule is, moreover, the only relevant rule defined in the literature that satisfies it (Gómez-Rúa and Vidal-Puga, 2011).

However, in the most general setting, no rule satisfies merge-proofness, as the next example shows:

Example 0.2 (Özsoy (2006)) *Assume we have three agents located, respectively, at nodes 1, 2 and 3. The connection cost between each agent and the source is* 1*, and the connection cost between any pair of agents is* 0*. The minimal cost is* 1 *and hence any rule* Φ *should satisfy* $\Phi_1+\Phi_2+\Phi_3=1$*. Assume w.l.o.g.* $\Phi_3 \geq \max\{\Phi_1, \Phi_2\}$*. Now, if players 1 and 3 join and appear as agent 1 alone, the planner would face the same problem as if players 2 and 3 join and appear as agent 2 alone. Then, a merge-proof rule should assign to player 1 (say* Φ'_1*) at least as much as to players 1 and 3 in the original problem, whereas it should assign to player 2 (say* Φ'_2*) at least as much as to players 2 and 3 in the original problem. This means* $\Phi_1+\Phi_3 \leq \Phi'_1$ *and* $\Phi_2+\Phi_3 \leq \Phi'_2$*, which implies* $\Phi_1+\Phi_2+2\Phi_3 \leq \Phi'_1+\Phi'_2$*. Since* $\Phi_1+\Phi_2+\Phi_3=1$ *and* $\Phi'_1+\Phi'_2=1$*, we deduce* $\Phi_3 \leq 0$*, which is impossible because* $\Phi_3 \geq \max\{\Phi_1, \Phi_2\}$ *and* $\Phi_1+\Phi_2+\Phi_3=1$*.*

The key issue in the previous example is that the planner has no way to know whether agent 3 has merged with agent 1 or with agent 2. This assumption is necessary in situations where the agents may use multiple replicas without being detected, as for example the case of users of a web page. However, this may not be a reasonable assumption in many other situations. In the *mcst* model, one may think in the case of departments in a campus or apartments in a building. In case the planner knows which mergers may have taken place, it is not difficult to derive a merge-proof rule[2]. On the other hand, it is not clear whether a merge-proof rule could also satisfy core selection and cost monotonicity[3].

In this paper, we model the *mcst* situation in such a way that the planner knows which mergers take place. If some agents merge and present themselves to the planner in this way, she should solve a situation where the nodes are formed by the union of several agents. We refer to this new problem as a (*mcst*) *situation*. A *mcst* situation generalizes the classical *mcst* problem.

Under this model, all the rules presented in the literature fail at least one of the properties. However, we propose a new rule that satisfies all of them, even the stronger versions. This rule is the weighted Shapley value of a particular cooperative game[4]. We also propose a characterization result of this rule with these and other properties. The other properties used in the characterization are *efficient merging*, *piece-wise additivity*, *symmetry* and *positivity*.

Positivity says that each node should pay at least zero. It has been used, in the context of minimum cost spanning tree problems, by Bergantiños and Vidal-Puga (2009). Piece-wise

[2]For example, charging all the cost to the merging agents.

[3]Charging all the cost to the merging agents will clearly not satisfy core selection.

[4]In the context of pricing traffic demand in a spanning network, Moulin (2014) also finds the weighted Shapley value of a cooperative game to satisfy the so-called routing-proofness. This property is related to split-proofness, since it precludes the agents to get advantage by reporting to be several different users along a path.

additivity is a weaker version of additivity. Additivity and symmetry are standard properties. They are used in the classical characterization of the Shapley value (Shapley, 1953). Different versions of piece-wise additivity property have also been used in the context of *mcst* problems by Branzei et al. (2004), Tijs et al. (2006), Bergantiños and Vidal-Puga (2009), Bergantiños et al. (2010), and Hougaard et al. (2010). Piece-wise additivity says that when the same network is optimal for two different cost matrices, then the cost-sharing is additive in the cost function.

Efficient merging, on the other hand, is related to the properties of split and merge-proofness. Even though merge-proofness precludes agents to build an inefficient network (as in Example 0.1), this concern is no longer relevant when the nodes that merge are already the closest ones (as in Example 0.2). Efficient Merging says that these nodes should not find harmful to join in advance in these cases.

REFERENCES

Athanassoglou, S. and Sethuraman, J. (2008). A characterization result in minimum cost spanning tree games. *Working paper at Columbia University*.

Bergantiños, G. and Gómez-Rúa, M. (2010). Minimum cost spanning tree problems with groups. *Economic Theory*, 43:227–262.

Bergantiños, G., Lorenzo, L., and Lorenzo-Freire, S. (2010). The family of cost monotonic and cost additive rules in minimum cost spanning tree problems. *Social Choice and Welfare*, 34:695–710.

Bergantiños, G. and Vidal-Puga, J. (2007a). A fair rule in minimum cost spanning tree problems. *Journal of Economic Theory*, 137:326–352.

Bergantiños, G. and Vidal-Puga, J. (2007b). The optimistic TU game in minimum cost spanning tree problems. *International Journal of Game Theory*, 36:223–239.

Bergantiños, G. and Vidal-Puga, J. (2008). On some properties of cost allocation rules in minimum cost spanning tree problems. *AUCO Czech Economic Review*, 2(3):251–267.

Bergantiños, G. and Vidal-Puga, J. (2009). Additivity in minimum cost spanning tree problems. *Journal of Mathematical Economics*, 1-2:38–42.

Bergantiños, G. and Vidal-Puga, J. (2015). Characterization of monotonic rules in minimum cost spanning tree problems. *International Journal of Game Theory*, Forthcoming.

Bird, C. (1976). On cost allocation for a spanning tree: a game theoretic approach. *Networks*, 6:335–350.

Bogomolnaia, A. and Moulin, H. (2010). Sharing the cost of a minimal cost spanning tree: beyond the folk solution. *Games and Economic Behavior*, 69(2):238–248.

Branzei, R., Moretti, S., Norde, H., and Tijs, S. (2004). The P-value for cost sharing in minimum cost spanning tree situations. *Theory and Decision*, 56:47–61.

Chun, Y. (1988). The proportional solution for right problems. *Mathematical Social Sciences*, 15:231–246.

De Frutos, M. (1999). Coalitional manipulation in a bankruptcy problem. *Review of Economic Design*, 4:255–272.

Dutta, B. and Kar, A. (2004). Cost monotonicity, consistency and minimum cost spanning tree games. *Games and Economic Behavior*, 48(2):223–248.

Feltkamp, V., Tijs, S., and Muto, S. (1994). On the irreducible core and the equal remaining obligations rule of minimum cost spanning extension problems. *CentER DP 1994 nr.106, Tilburg University, The Netherlands*.

Gómez-Rúa, M. and Vidal-Puga, J. (2011). Merge-proofness in minimum cost spanning tree problems. *International Journal of Game Theory*, 40(2):309–329.

Granot, D. and Huberman, G. (1981). On minimum cost spanning tree games. *Mathematical Programming*, 21:1–18.

Granot, D. and Huberman, G. (1984). On the core and nucleolus of minimum cost spanning tree problems. *Mathematical Programming*, 29:323–347.

Hart, S. (1974). Formation of cartels in large markets. *Journal of Economic Theory*, 7:453–466.

Hougaard, J. L., Moulin, H., and Østerdal, L. P. (2010). Decentralized pricing in minimum cost spanning trees. *Economic Theory*, 44:293–306.

Ju, B.-G. (2003). Manipulation via merging and splitting in claims problems. *Review of Economic Design*, 8:205–215.

Ju, B.-G., Miyagawa, E., and Sakai, T. (2007). Non-manipulable division rules in claim problems and generalizations. *Journal of Economic Theory*, 132:1–26.

Kar, A. (2002). Axiomatization of the Shapley value on minimum cost spanning tree games. *Games and Economic Behavior*, 38:265–277.

Knudsen, P. H. and Østerdal, L. P. (2012). Merging and splitting in cooperative games: some (im)possibility results. *International Journal of Game Theory*, 41(4):763–774.

Kruskal, J. (1956). On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50.

Legros, P. (1987). Disadvantageous syndicates and stable cartels: the case of the nucleolus. *Journal of Economic Theory*, 42:30–49.

Lorenzo, L. and Lorenzo-Freire, S. (2009). A characterization of Kruskal sharing rules for minimum cost spanning tree problems. *International Journal of Game Theory*, 38:107–126.

Maschler, M. (1976). An advantage of the bargaining set over the core. *Journal of Economic Theory*, 13:184–192.

Moulin, H. (2002). Axiomatic cost and surplus sharing. In Arrow, K., Sen, A., and Suzumura, K., editors, *Handbook of Social Choice and Welfare*, pages 289–357. Elsevier, Amsterdam.

Moulin, H. (2007). On scheduling fees to prevent merging, splitting and transferring of jobs. *Mathematics of Operations Research*, 32(2):266–283.

Moulin, H. (2008). Proportional scheduling, split-proofness, and merge-proofness. *Games and Economic Behavior*, 63:567–587.

Moulin, H. (2014). Pricing traffic in a spanning network. *Games and Economic Behavior*, 86:475–490.

Norde, H., Moretti, S., and Tijs, S. (2004). Minimum cost spanning tree games and population monotonic allocation schemes. *European Journal of Operational Research*, 154(1):84–97.

O'Neill, B. (1982). A problem of rights arbitration from the talmud. *Mathematical Social Sciences*, 2:345–371.

Özsoy, H. (2006). A characterization of Bird's rule. *Working paper at Rice University*.

Postlewaite, A. and Rosenthal, R. (1974). Disadvantageous syndicates. *Journal of Economic Theory*, 9:324–326.

Prim, R. (1957). Shortest connection network and some generalization. *Bell Systems Technology Journal*, 36:1389–1401.

Rosenmüller, J. and Sudhölter, P. (2004). Cartels via the modiclus. *Discrete Applied Mathematics*, 134:263–302.

Shapley, L. (1953). A value for n-person games. In Kuhn and Tucker, editors, *Contributions to the theory of games II*, pages 307–317. Princeton University Press, Princeton NJ.

Sprumont, Y. (2005). On the discrete version of the Aumann-Shapley cost-sharing method. *Econometrica*, 73(5):1693–1712.

Thomson, W. (1983a). The fair division of a fixed supply among a growing population. *Mathematics of Operations Research*, 8(3):319–326.

Thomson, W. (1983b). Problems of fair division and the egalitarian solution. *Journal of Economic Theory*, 31:211–226.

Tijs, S., Moretti, S., Branzei, R., and Norde, H. (2006). The Bird core for minimum cost spanning tree problems revisited: Monotonicity and additivity aspects. In Seeger, A., editor, *Recent Advances in Optimization*, pages 305–322. Springer Berlin Heidelberg.

Trudeau, C. (2012). A new stable and more responsible cost sharing solution for mcst problems. *Games and Economic Behavior*, 75(1):402–412.

Comunicacións orais

Sesión 1 de Estatística

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

ASPETOS GEOMÉTRICOS NA ESCOLHA DA JANELA PARA O ESTIMADOR DO NÚCLEO

Paulo Eduardo Oliveira[1]

[1]CMUC, Departamento de Matemática, Universidade de Coimbra, Portugal

RESUMO

A escolha da janela para o método do núcleo, feita a partir da caracterização assintótica do erro quadrático médio, está bem estudada para variáveis com valores num espaço euclideano de dimensão finita. Quando a variável independente toma valores em espaços mais abstratos a descrição do comportamento do erro quadrático torna-se mais difícil, dependendo bastante da geometria do espaço e da regularidade da distribuição. Estes aspetos geométricos ficam escondidos nos espaços de dimensão finita em que tudo se exprime em relação à medida de Lebesgue. Em espaços abstratos não existe, em geral, um análogo da medida de Lebesgue pelo que as propriedades geométricas das distribuições se tornam bastante mais relevantes.

Nesta comunicação apresentamos uma caracterização do erro quadrático médio do estimador do núcleo para a regressão válida em espaços abstratos. Esta representação inclui explicitamente termos que dependem da geometria e regularidade da distribuição da variável independente. A caracterização encontrada reduz-se às habituais representações no caso de variáveis com valores em espaços euclideanos de dimensão finita. As hipóteses sobre a função de regressão incluem apenas a continuidade do primeiro e segundo momentos, não sendo necessárias quaisquer propriedades de diferenciabilidade.

Palabras e frases chave: estimador do núcleo, probabilidades locais, erro quadrático, comportamento assintótico.

REFERENCIAS

Ferraty, F. and Vieu, P. (2006) Nonparametric Functional Data Analysis. Theory and Practice. New York, Springer.

Ferraty, F., Mas, A. and Vieu, P. (2007) Nonparametric Regression on Functional Data: Inference and practical aspects. Aus. N. Z. J. Stat., 49, 267–286.

Jacob, P. and Oliveira, P.E. (1995) A general approach to nonparametric histogram estimation. Statistics, 27, 73–92.

Jacob, P. and Oliveira, P.E. (1997) Kernel estimators of general Radon-Nikodym derivatives. Statistics, 30, 25–46.

Karol, A. and Nazarov, A. (2014) Small ball probabilities for smooth Gaussian fields and tensor products of compact operators. Math. Nachr., 287, 595-609.

Li, W.V. and Shao, Q.M. (2001) Gaussian processes: inequalities, small-ball probabilities and applications, In Handbook of Statistics, 19, 533–597, Elsevier.

Masry, E. (2005) Nonparametric regression estimation for dependent functional data: asymptotic normality. Stochastic Process. Appl., 115, 155–177.

Oliveira, P.E. (2000) Mean square error for histograms when estimating Radon-Nikodym derivatives. Portugal. Math., 57, 1-16.

Oliveira, P.E. (2005) Nonparametric density and regression estimation for functional data. Preprint, Publ. Dep. Matemática Univ. Coimbra 05-09.

Ramsay, J. and Silverman, B. (2002) Applied Functional Data Analysis: Methods and Case Studies. New York, Springer.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

A survey on statistical choice of extreme value models

M. Manuela Neves[1]

[1]CEAUL and ISA, Universidade de Lisboa, Portugal

Abstract

Statistical analysis of extreme large (or small) values is of major importance in several fields such as environmental science, finance, insurance, hydrology, climatology, telecommunications to refer only a few examples. Extreme value theory (EVT) is the branch of probability and statistics dedicated to characterising the behaviour of the extreme observations. The tails of the underlying model are more informative than the central part. Interpreting data for extreme values with the adequate choice of the tail represents an important challenge when dealing with real practical applications. However a goodness-of-fit test would not give the adequate information regarding the shape of the tails. The objective of this work is to present a brief overview about several tests for the statistical choice of extreme value models. An application to a data set of flood discharge rate in a hydrometric station in the North of Portugal will illustrate some statistical testing procedures.

Palavras e frases chave: Extreme value theory; extreme value index; parametric and semi-parametric estimation; statistical testing; R software.

1. Introduction

Statistical inference about extreme events is dedicated to the estimation of the probability of occurrence of extreme events and can clearly be deduced only from those observations that are extreme in some sense. If one is interested in *extreme quantities*, a challenge faced by hydrologists for the prediction of large floods for example, see Hosking and Wallis (1987). Model assumptions on the upper tail of the underlying distribution function (df) are necessary. In recent years there has been a significant development of several procedures to infer, from a data set, about the extremal model that most conveniently describes the df of the underlying population.

In EVT there are three domains of attraction, the Gumbel, Fréchet, and Weibull domains. All distributions in the Gumbel domain have the exponential as the limiting distribution of their tail. This domain contains the majority of well-known distributions such as the normal, the exponential, and the gamma. The Fréchet domain contains distributions with an infinite and heavier tail than the exponential, e.g., the Cauchy distribution. The Weibull domain contains distributions with lighter tails than exponential, which possess a finite upper bound, e.g., the uniform distribution.

The Gumbel domain of attraction is the prominent component of EVT and any attempt to test the Gumbel hypothesis for maxima, reduces to the determination of whether or not data from the extreme right tail of a distribution appear to be exponential. Being Gumbel distributions the heart of EVT it is adopted in many applied situations. However deciding by an exponential or a light/heavy tail is the first question to answer and to influence the study of other parameters of interest.

The problem is not recent, as seen in Segers and Teugels (2000). Various test procedures have appeared in the literature; we can refer to Van Montfort (1970), Stephens (1977), Tiago de Oliveira (1981), Gomes (1982), Tiago de Oliveira and Gomes (1984), Wang *et al.* (1996), Marohn (1998a,b), Marohn (2000), Jureckova and Picek (2001), Cabana and Quiroz (2005), and more recently Neves and Fraga Alves (2007, 2008), and the references in all these papers.

For the statistical analysis and inference on extreme values two approaches have been developed: the parametric and the semi-parametric approaches. In both situations the statistical inference results will be improved if the choice about the most adequate tail decay of the underlying df is made: an exponential, a polynomial or a light one.

The extreme value index, ξ, is the parameter directly related with the heaviness of the right tail of the underlying distribution of the data but other parameters of extreme values are also of great interest, such as, for example, the high quantile of probability $1-p$, with p small, i.e., *high level* usually designed by the *return level* associated to the *return period* $1/p$, i.e., the expected waiting time between independent exceedances of specific high level or the right endpoint of an underlying model.

In all areas of application it is of major importance to use adequate, accurate and if possible free statistical software. The R software (R Development Core Team, 2012) is an open source environment that incorporates in its base a huge amount of statistical techniques made available by the scientific community. The application of several tests (from which we have here plotted only two of them) and a simulation study have been performed in R.

The aim of this work is to prepare a survey on tests available for the statistical choice of extreme value conditions.

2. Background and notation in EVT

Let us consider an independent and identically distributed (iid) sample of random variables $(X_1, \ldots, X_n)$, with common df F; if we want to derive the distribution of the maximum $M_n = \max(X_1, \ldots, X_n)$, we know that,

$$P[M_n \leq x] = P[X_1 \leq x, \ldots, X_n \leq x] = P[X_1 \leq x] \ldots P[X_n \leq x] = F^n(x).$$

However the df F^n is of little help in practice since F is itself unknown and even the misspecification of F can lead to large errors in the distribution of the maximum.

The asymptotic theory of sample extremes distribution has been developed under analogous arguments to those considered in the *central limit theorem*. First results date back to Fréchet (1927), who obtained the asymptotic distribution of the maximum. Fisher and Tippet (1928), Gumbel (1958) and von Mises (1936) presented the first studies on the *extremal limit problem*. Finally Gnedenko (1943) and de Haan (1970) solved the most important problems related with the asymptotic behaviour of statistical extremes, giving conditions for the existence of sequences $\{a_n\} \in R^+$ and $\{b_n\} \in R$ such that,

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = EV_\xi(x) \quad \forall x \in R, \tag{1}$$

where EV_ξ is a nondegenerate distribution function. This function, called *Extreme Value* df is given by

$$EV_\xi(x) = \begin{cases} \exp\left\{-\left[1+\xi x\right]^{-1/\xi}\right\}, & 1+\xi x > 0,\ if\ \xi \neq 0\ ; \\ \exp\left\{-\exp\left[-x\right]\right\}, & x \in R,\ if\ \xi = 0, \end{cases} \tag{2}$$

where $\xi \in R$ is the shape parameter, usually designated as extreme value index (EVI). When (1) holds we say that F is in the domain of attraction (for maxima) of EV_ξ and write $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$.

The shape parameter, ξ, measures the heaviness of the right-tail, $\overline{F} := 1-F$. If $\xi = 0$, the right tail is of an exponential type. The *right endpoint* of F, $x^* \equiv x^F := \sup\{x : F(x) < 1\}$, can then be either finite or infinite. Many common distributions such as normal, lognormal, exponential and gamma distributions are in the domain of attraction for maxima of the **Gumbel df**. If $\xi > 0$, the right tail is heavy, it is of a negative polynomial type and F has an infinite *right endpoint*. Moments of order r are infinite if $r \geq 1/\xi$. Pareto and Cauchy are examples of distributions in the domain of attraction for maxima of the **Fréchet df**. If $\xi < 0$, the right tail is short and F has a finite *right endpoint*, $x^* = -1/\xi < +\infty$. As an example of a distribution belonging to the domain of attraction for maxima of the **Weibull df** we can refer to the uniform distribution.

Figure 1 shows plots of density functions for the Gumbel, Weibull and Fréchet (left) and a zoom of the plot for comparing with the standard normal density (right).

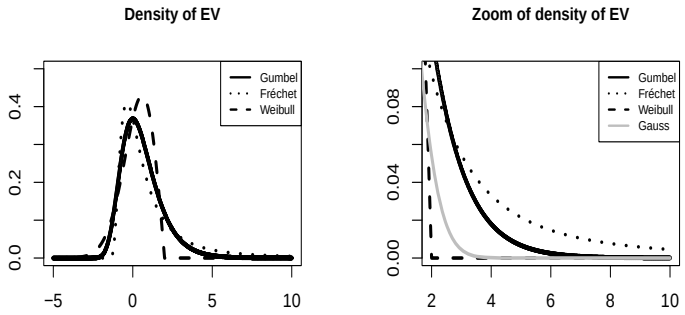


Figura 1: Density functions for the Gumbel, Weibull and Fréchet (left) and a zoom of those functions compared with the standard normal density (right).

Rather than just considering the maxima or the k largest values in the sample, we can consider all values larger than a given high threshold u_n. This was formalized by Balkema and de Haan (1974) and Pickands III (1975), who proved that $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$ if and only if the upper tail of F is, in a certain sense, close of the upper tail of GP_ξ, the *Generalized Pareto* df, where,

$$GP_\xi(y) = \begin{cases} 1-(1+\xi y/\widetilde{\delta})^{-1/\xi}, & (1+\xi y/\widetilde{\delta})>0, \text{ and } y>0, \text{ if } \xi \neq 0; \\ 1-\exp(-y), & y>0, \text{ if } \xi = 0. \end{cases} \tag{3}$$

where ξ is the shape parameter, equal to that of the corresponding EV distribution, and the scale parameter $\widetilde{\delta} = \delta + \xi(u_n - \lambda)$, where λ is the location parameter in the EV df.

3. Statistical choice in extreme value models

As we mentioned before, statistical inference in EVT has developed under two approaches based on the different ways of defining observations as extremes, leading to the application of different procedures. In any of these procedures it is assumed that the underlying distribution function F belongs to $\mathcal{D}_{\mathcal{M}}(EV_\xi)$, for a appropriate value of ξ, or is in specific sub-domains of $\mathcal{D}_{\mathcal{M}}(EV_\xi)$. These condition is called the *extreme value condition* and is not always fulfilled. So, before dealing with an application, it is important to check whether the extreme value condition is reasonable for a data set. We must check the assumption:

$$H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_\xi) \text{ for some } \xi \in R.$$

We can refer to some tests for the hypothesis $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$ such as Dietrich et al. (2002) and Drees et al. (2006) that considered the case of testing the general extreme value conditions. Hüsler and Li (2006) provided quite accuracy tables of critical points for those tests.

Beirlant et. al (2006) derived a goodness-of-fit test for testing the Pareto-type behaviour on the tail of the underlying df.

Here we will give more emphasis to the problem of choosing one of the three extremal types, giving preference to the Gumbel model for the null hypothesis; this problem has frequently received the general designation of *statistical choice of extremal models* and has been handled by several authors.

Classical parametric frameworks for modeling and estimation were the first to be defined, working with models in (2). Nevertheless, in the late seventies, estimation procedures in EVT began to be made on a semi-parametric framework based on probabilistic asymptotic results in the tail of the unknown distribution.

The parametric framework

In parametric framework there are only three classes of models, given by the EVT with shape parameter ξ and possibly location and scale parameters. The Gumbel type df, EV_0, and the exponential type df, GP_0, received a central place given the simplicity of the inference associated with them.

The important testing problem, named by Tiago de Oliveira as *'the trilema'* consist in performing the test

$$H_0 : \xi = 0 \quad vs \quad H_1 : \xi \neq 0 \quad \text{or} \quad H_1 : \xi > 0 \quad \text{or} \quad H_1 : \xi < 0.$$

Several papers regarding these tests can be mentioned such as Van Montfort (1970), Bardsley (1977), Otten and Van Montfort (1978), Tiago de Oliveira (1981), Gomes (1982), Tiago de Oliveira (1984), Tiago de Oliveira and Gomes (1984), Hosking (1984), among others.

The semi-parametric framework

As said before, in a semi-parametric framework we need only to assume the assumption that $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)$. Here ξ is the primordial parameter since determines the shape of the tail of the underlying distribution function F. A negative value for ξ is associated to the Weibull domain of attraction in which all the df's are short tailed with finite right endpoint. If $\xi > 0$, we have the Fréchet domain of attraction to which it belongs the heavy tailed df's with polynomially decaying tail. The case of $\xi = 0$ is particularly important due to the simplicity of inference within the Gumbel domain wish contains a great variety of df's with an exponential tail having finite right end point or not.

The natural hypothesis testing are now:

$$H_0 : F \in \mathcal{D}_{\mathcal{M}}(EV_0) \quad vs \quad H_1 : F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi \neq 0}. \tag{4}$$

or against the one-sided alternatives $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi<0}$ or $F \in \mathcal{D}_{\mathcal{M}}(EV_\xi)_{\xi>0}$.

This is the already mentioned *statistical choice of extreme domains of attraction.*

Under this framework several tests have been proposed in literature, among which we can mention Galambos (1982), Castillho *et al* (1989), Hasofer and Wang (1992), Falk (1995), Fraga Alves and Gomes (1996), Marohn (1998a,b), Fraga Alves (1999) and Segers and Teugels (2000).

Correia and Neves (1996) considered the Hasofer and Wang statistic and presented a slight modification. They have performed an extensive simulation study.

Brilhante (2004) derived a resistant and robust test for testing exponentially versus generalized Pareto, showing that her proposal as a good compromise among several scenarios.

Castillo (2014) provided tests to distinguish between polynomial and exponential tails, based on properties of the *coeficient of variation* (CV). It is especially important for applications to determine the threshold where the tail begins.

Neves et al. (2006) and Neves and Fraga Alves (2007) introduced new procedures that were completely studied in Neves and Fraga Alves (2008). In these procedures the statistics for testing (4) are based on the k excesses over the $(n-k)th$ ascending intermediate order statistic $X_{n-k:n}$, where $k = k_n \to \infty$, $k/n \to 0$ as $n \to \infty$. Thus, under the null hypothesis of Gumbel domain of attraction and further assuming second order conditions on the upper tail of F and that the intermediate sequence k_n is such that $k^{1/2}A(n/k) \to 0$ as $n \to \infty$ where A is related to the second order condition, the authors presented:

$$Ratio - test: \qquad R_n^*(k) := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}(X_{n-i+1:n} - X_{n-k:n})} - \log k \longrightarrow EV_0 \qquad \text{as} \quad n \to \infty \tag{5}$$

$$Gt - test: \qquad G_n(k) := \frac{\frac{1}{k}\sum_{i=1}^{k}(X_{n-i+1:n} - X_{n-k:n})^2}{\left(\frac{1}{k}\sum_{i=1}^{k} X_{n-i+1:n} - X_{n-k:n}\right)^2} \tag{6}$$

$$G_n^*(k) = \sqrt{k/4}\,(G_n(k) - 2) \longrightarrow N(0,1) \qquad \text{as} \quad n \to \infty$$

$$HW - test: \qquad W_n(k) := \frac{1}{k}\left[1 - \frac{G_n(k) - 2}{1 + (G_n(k) - 2)}\right], \tag{7}$$

$$W_n^*(k) = \sqrt{k/4}\,(kW_n(k) - 1) \longrightarrow N(0,1) \qquad \text{as} \quad n \to \infty$$

The null hypothesis is rejected in (4) if $T_n^* < \chi_{\alpha/2}$ or $T_n^* > \chi_{1-\alpha/2}$, where T^* has to be replaced by R^*, G^* or W^* and χ_p is the p probability quantile of the corresponding limiting distribution. For the one-sided test, the null hypothesis is rejected in favour of either unilateral alternatives $H_1^l : F \in D_M(EV_\xi)_{\xi<0}$ or $H_1^r : F \in D_M(EV_\xi)_{\xi>0}$ if $R_n^*(k) < -\chi_\alpha (G_n^*(k) < -\chi_\alpha)$ or $R_n^*(k) > \chi_{1-\alpha} (G_n^*(k) > \chi_{1-\alpha})$.

For the HW-test, the H_0 is rejected in favour $H_1^l : F \in D_M(EV_\xi)_{\xi<0}$ or $H_1^r : F \in D_M(EV_\xi)_{\xi>0}$ if $W_n^*(k) > \chi_{1-\alpha}$ or $W_n^*(k) < -\chi_\alpha$.

Neves and Fraga Alves (2007) performed an extensive simulation study concerning the proposed procedures.

5. A case study

The data set available in this study refers to the daily mean flow discharge rate in the hydrometric station of Fragas da Torre in Paiva river, (m^3/s) from 1 October, 1946 to 30 September, 2006, collected from the "SNIRH: Sistema Nacional de Informação dos Recursos Hídricos".

The source of Paiva river is in the Serra de Leomil, in the North of Portugal, and its hydrographic basin has an area of approximately 700 Km^2. The flow study of this river is a matter of major importance since it is one of the main alternatives to the Douro river as source of water supply in the south of Oporto region.

As we are interested in maximum values, after a graphical analysis of the tail behaviour for the different months and also in order to have stationarity of the data only the months from November until April are considered. Several analysis have been presented in Neves *et al* (2015).

The application of tests described in (5) and (6) are plotted in Figure 2,

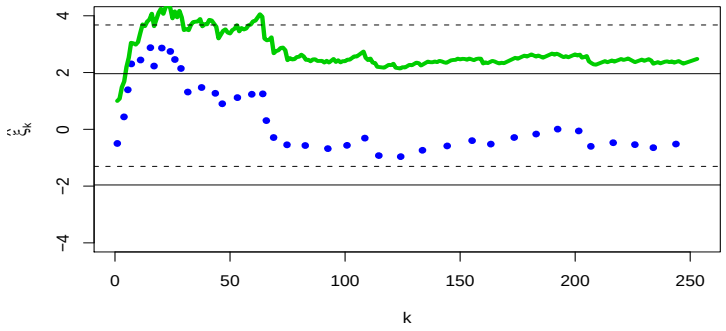


Figura 2: Sample paths of the statistics R_n^*, in green solid line and the quantiles $\chi_{0.025}$ and $\chi_{0.975}$ for the standard Gumbel distribution in dashed lines. The dotted blue lines is the sample path for the G_n^* statistics, and the associated quantiles of the standard normal, solid black lines.

These tests suggest the non rejection of the null hypothesis, leading us to consider that the data are in the domain of attraction of the Gumbel distribution.

Acknowledgement

This research was partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2014 (CEAUL)

REFERENCES

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. Annals of Probability, 2, 792–804.

Bardsley, W. E. (1977) A test for distinguishing between extreme value distributions. Journal of Hydrology, 34, 377–381.

Beirlant, J., de Wet, T., and Goegebeur, Y. (2006) A goodness-of-fit statistic for Pareto-type behaviour. Journal of Computational and Applied Mathematics, 186, 99–116.

Brilhante, M. F. (2004) Exponentiality versus generalized Pareto – a resistant and robust test. RevStat, 2(1), 1–13.

Cabana, A. and Quiroz,A. J. (2005) Using the empirical moment generating function in testing for the Weibull and the Type I extreme value distribution, Test 13, 417-431

Castillo, E., Galambos, J., and Sarabia, J. M. (1989) The selection of the domain of attraction of an extreme value distribution from a set of data. In J. HÄusler and R.-D. Reiss, eds., Extreme value theory (Ober- wolfach, 1987)– Lecture Notes in Statistics 51 , 181–190. Springer, Berlin-Heidelberg.

Castillo, J. del, Daoudi,J. and Lockhart, R. (2014) Methods to distinguish between polynomial and exponential tails. Scandinavian Journal of Statistics, 41, 382–393

Correia, A. L. e Neves, M. (1996) Escolha Estatística em Modelos Extremais - Testes de Ajustamento . in Bom Senso e Sensibilidade, (Branco, J., Gomes, P. e Prata, J. eds.), Actas do III Congresso Anual da Sociedade Portuguesa de Estatística, Edições Salamandra, 223-236.

Dietrich, D., de Haan, L., and Husler, J. (2002) Testing extreme value conditions. Extremes, 5, 71–85.

Drees, H., de Haan, L., and Li, D. (2006) Approximations to the tail empirical distribution function with application to testing extreme value conditions. Journal of Statistical Planning and Inference, 136, 3498–3538.

Falk, M. (1995). On testing the extreme value index via the POT-method. Annals of Statistics, 23, 2013–2035.

Fisher, R. A. and Tippett, L. H. C. (1928) On the estimation of the frequency distributions of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society, 24, 180-190.

Fraga Alves, M. I. (1999) Asymptotic distribution of Gumbel statistic in a semi-parametric approach. Portugalia Mathematica, 56(3),282–298.

Fraga Alves, M. I. and Gomes, M. I. (1996) Statistical choice of extreme value domains of attraction – a comparative analysis. Communications in Statistics – Theory and Methods, 25(4), 789–811.

Fréchet, M. (1927) Sur la loi de probabilité de l'écart maximum. Ann. Soc. Polon. Math. (Cracovie), 6, 93-116.

Galambos, J. (1982) A statistical test for extreme value distributions. In e. B.W. Gnedenko et al., ed., Nonparametric Statistical Inference, 221–230. North Holland, Amsterdam.

Gnedenko, B. V. (1943) Sur la distribution limite du terme maximum d'une série aléaatoire. Annals of Mathematics, 44, 423–453.

Gomes, M. (1982) A note on statistical choice of extremal models. In Proc. IX Jornadas Mat. Hispano-Lusas, Salamanca. 653–655.

Gomes, M. I. (1987) Extreme value theory – statistical choice. In Colloq. Math. Soc. Janos Bolyai 45 , 195–210.

Gomes, M. I. (1989) Comparison of extremal models through statistical choice in multidimensional backgrounds. In J. HÄusler and R.-D. Reiss, eds., Extreme value theory (Oberwolfach, 1987) – Lecture Notes in Statistics 51 , 191–203. Springer, Berlin-Heidelberg.

Gomes, M. I. and Van Montfort, M. A. J. (1986) Exponentiality versus generalized Pareto-quick tests. In Proc. III Internat. Conf. Statistical Climatology, 185–195.

Gumbel, E. J. (1935) Les valeurs extrêmes des distributions statistiques. Ann. Inst. Henri Poincaré, 5(2), 115-158.

de Haan, L. (1970). On Regular Variation and its Applications to the Weak Convergence of Sample Extremes, Mathematical Centre Tract 32, Amesterdam, Dordrecht: D.Reidel.

Hasofer, A. and Wang, J. Z. (1992) A test for extreme value domain of attraction. Journal of the American Statistical Association, 87, 171–177.

Hosking, J. (1984) Testing whether the shape parameter is zero in the generalized extreme value distribution. Biometrika, 71, 367–374.

Hosking, J. and Wallis J.R.(1987) Parameter and quantile estimation for the generalized Pareto distribution. Technometrics, 29, 339-349.

Husler, J. and Li, D. (2006) On testing extreme value conditions. Extremes, 9, 69–86.

Jureckova, J. and Picek, J. (2001) A class of tests on the tail index, Extremes 4, 165–183.

Marohn, F. (1998a) An adaptive eficient test for Gumbel domain of attraction. Scandinavian Journal of Statistics, 25, 311–324.

Marohn, F. (1998b) Testing the Gumbel hypothesis via the POT- method. Extremes, 1:2, 191–213.

Marohn, F. (2000). Testing extreme value models. Extremes, 3, 363–384.

Neves, C. and Fraga Alves, M. I. (2007) Semi-parametric approach to Hasofer-Wang and Greenwood statistics in extremes. Test, 16, 297-313.

Neves, C. and Fraga Alves, M. I. (2008) Testing extreme value conditions– an overview and recent approaches. Revstat, 6(1), 83–100.

Neves, C., Picek, J., and Fraga Alves, M. I. (2006) The contribution of the maximum to the sum of excesses for testing max-domains of attraction. Journal of Statistical Planning and Inference, 136(4), 1281–1301.

Neves, M.M., Penalva, H. and Nunes, S. (2015) Extreme value analysis of river levels in a hydrometric station in the North of Portugal. Current Topics on Risk Analysis: ICRA6 and RISK 2015 Conference proceedings. M. Guillén, A. Juan, H. Ramalhinho, I. Serra and C. Serrat (edts), 533–538.

Otten, A. and Van Montfort, M. (1978). The power of two tests on the type of distribution of extremes. Journal of Hydrology, 37, 195–199.

Pickands III, J. (1975) Statistical inference using extreme order statistics. Annals of Statistics, 3, 119–131.

R Development Core Team (2012), R: A Language and Environment for Statistical Computing, R. Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. ISBN 3-900051-07-0.

Segers, J. and Teugels, J. (2000) Testing the Gumbel hypothesis by Galton's ratio. Extremes, 3(3)291–303.

Stephens, M. A. (1977) Goodness-of-fit for the extreme value distribution. Biometrika, 64, 583–588.

Tiago de Oliveira, J. (1981). The selection of the domain of attraction of an extreme value distribution from a set of data. In C. Taillie, ed., Statistical Distributions in Scientific Work, Vol. 6 , 367–387. D. Reidel, Dordrecht.

Tiago de Oliveira, J. (1984). Univariate extremes: Statistical choice. In J. Tiago de Oliveira, ed., Statistical Extremes and Applications, 91–107. D. Reidel, Dordrecht.

Tiago de Oliveira, J. and Gomes, M. I. (1984). Two statistics for choice of univariate extreme value models. In J. Tiago de Oliveira, ed., Statistical Extremes and Applications, 651–668. D. Reidel, Dordrecht.

Van Montfort, M. A. J. (1970). On testing that the distribution of extremes is of type i when type ii is the alternative. Journal of Hydrology, 11, 421–427.

Wang, J. Z., Cooke, P., and Li, S. (1996). Determination of domains of attraction based on a sequence of maxima. Australian Journal of Statistics, 38, 173–181.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

DIAGNÓSTICO DE MODELOS DE CLASES LATENTES EN TABLAS CON FRECUENCIAS PEQUEÑAS O NULAS: EFECTO DE SUMAR A LAS FRECUENCIAS DE LOS PATRONES DE RESPUESTAS UNA CONSTANTE

Carlomagno Araya Alpízar 1[1]

[1] Universidad de Costa Rica, Sede de Occidente.

RESUMEN

El artículo trata del problema que puede surgir en la aplicación de los Modelos de Clases Latentes, cuando se tienen conjuntos de datos binarios con frecuencias de patrones de repuestas nulos o pequeños. En este trabajo analiza el efecto de sumar una pequeña constante a las frecuencias observadas de cada patrón de respuesta celdas de la tabla múltiple sobre los errores relativos de los estadísticos de bondad de ajuste y las probabilidades totales de clases. Se determina por medio de la simulación de datos que, entre más pequeña sea la constante sumada a las celdas mayor exactitud tienen los resultados. Considerando los resultados, se recomienda utilizar la constante $1/2^p = 1/R$, donde p es número de variables con respuesta binaria y R representa el total de patrones de respuestas.
Palabras claves: clases latentes, tablas escasas, bondad de ajuste, datos binarios.

1. INTRODUCCIÓN

El modelo de clases latentes (MCL) es una técnica estadística que permite estudiar la existencia de una (o varias) variable(s) latente(s) a partir de un conjunto de variables explicativas observadas, y definir a partir de sus clases una clasificación o tipología de los individuos analizados. La técnica utiliza los estadísticos de bondad de ajuste (EBA) son un criterio utilizado para la selección de un MCL. Estos estadísticos tienen, bajo ciertas condiciones una distribución de probabilidad teórica χ^2.

Las celdas con frecuencias muy pequeñas o cero, tienen una contribución muy importante en los EBA, al provocar un aumento del error Tipo I. Por esa razón, los EBA están muy distorsionados a medida que el número de celdas de la tabla múltiple tienen frecuencias observadas pequeñas. En este sentido, el objetivo de estudio es determinar el efecto de sumar una pequeña constante a la frecuencia de los patrones de respuestas sobre los errores relativos de los estadísticos de bondad de ajuste y las probabilidades totales de clases en MCL cuando se tienen frecuencias pequeñas o nulas en los patrones de respuesta.

2. MÉTODO Y RESULTADOS

Para medir el efecto de las constantes sobre los errores relativos de los EBA y las probabilidades totales de clases, fueron simulados conjuntos de datos de 1000 patrones, para cada nivel de densidad de patrones (1,2,…,20) y 5 constantes. En total, para cada MCL fueron generados 100,000 patrones. También, se consideró sumar la constante a todas las frecuencias de los patrones de respuesta o únicamente a aquellos con frecuencias nulas (celdas vacías). Esto para variables manifiestas de 4 hasta 10 y tamaños de muestra $n = 2^p$ hasta $n = 20 \times 2^p$.

En cuanto a los resultados, los EBA presentan diferencias promedios significativas en los errores relativos según las constantes sumadas a la frecuencia de las celdas (las celdas representan los patrones posibles) y los niveles de densidad de los patrones, tanto para la alternativa de sumar la constante a todas las celdas o solamente a las vacías. En general, para todas las simulaciones, existe una interacción altamente significativa entre la densidad de los patrones y las constantes, para las medias de los errores relativos de los EBA y probabilidades totales de clase.

Los errores relativos de los EBA son grandes cuando el nivel de densidad de los patrones es pequeño, tendiendo a disminuir considerablemente al incrementarse el tamaño de la muestra y por ende, al aumentar el porcentaje de patrones de respuestas observados, estos son aquellos que tienen frecuencias no nulas. Lo anterior, queda mostrado en el **Figura 1**, por ejemplo para 5 variables manifiestas binarias y 2 clases latentes, sumando las constante a todas las celdas y para el estadístico Cressie-Read.

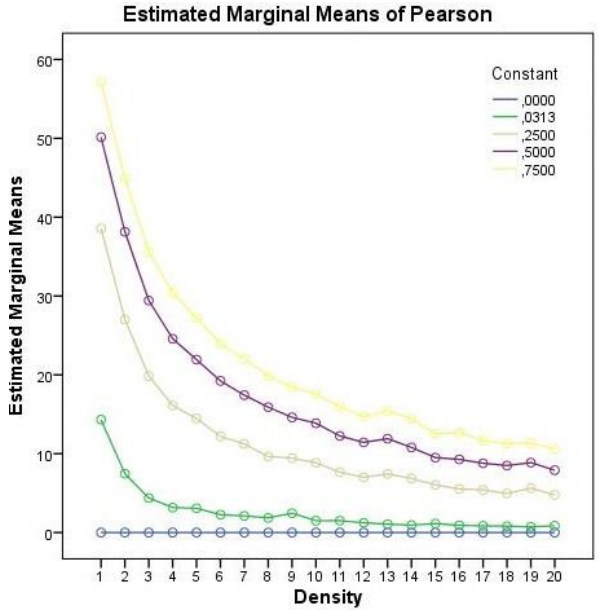


Figura 1: Errores relativos para el estadístico Cressie-Read: 5 variables manifiestas y 2 clases latentes (sumando constantes a todas las celdas).

Como se ha mencionó, los errores relativos decrecen al aumentar el nivel de densidad de los patrones, pero el efecto sobre los resultados de los errores relativos usando R es más significativo, ya que tienden a hacer muy pequeños y no presentar diferencias significativas respecto al grupo control, que representa los EBA calculados sin sumar ninguna constante a las frecuencias de los patrones de respuesta.

Por otra parte, respecto a los errores relativos sobre las probabilidades totales de clases del MCL, en general, los errores tienden a ser pequeños cuando los tamaños de las clases son grandes, y aumentan conforme las probabilidades totales son pequeñas. Respecto a lo anterior, observarse el **Figura 2** para clase latente 1, para 6 variables manifiestas y 2 clases latentes.

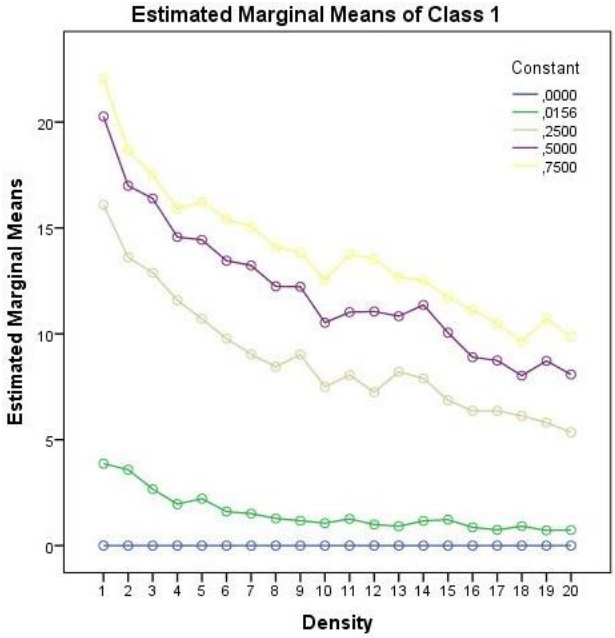


Figura 2. Errores relativos para la clase latente 1: 6 variables manifiestas y 2 clases latentes (sumando constante celdas vacías)

Utilizando la constante 0.75 se obtienen los mayores valores medios de los errores relativos para las dos clases latentes. Lo anterior, se cumple para la opción de sumar la constante únicamente a las frecuencias

de patrones no observados en la muestra (es decir, celdas vacías), como también para opción de sumar la constante a todas las celdas.

3. CONCLUSIONES

Los resultados obtenidos nos permiten llegar a las siguientes conclusiones:

- El sumar una constante a las frecuencias de los patrones de respuesta (celdas) modifica significativamente los errores relativos de los estadísticos de bondad de ajuste y las probabilidades totales de clase.
- Se obtienen resultados similares en cuanto a los errores relativos, al sumar la constante a todas las celdas o únicamente a las vacías (es decir, aquellos patrones de respuesta con frecuencias nulas).
- El efecto los errores relativos al sumar la constante R, es más pequeño en comparación al obtenidos al sumar 0.25, 0.50 y 0.75.
- Entre los estadísticos de bondad de ajuste, el estadísticos Freeman-Tukey proporciona los errores relativos más grandes, en comparación a la razón de verosimilitud, Pearson y Cressie-Read.
- Y finalmente, los errores relativos son las probabilidades a priori (tamaños de las clases latentes), aumenta conforme el tamaño de la muestra decrece. Además, se obtienen los errores relativos más pequeños usando la constante R.

REFERENCIAS

Agresti, A. (1990) Categorical Data Analysis. Wiley, New York.

Araya, A.C. (2010). Modelos de clases latentes en tablas poco ocupadas: una contribución basada en bootstrap (Tesis de Doctorado). Universidad de Salamanca, España.

Clogg, C.C. and Goodman, L. A (1984) Latent Structure Analysis of a set of Multidimensional Contingency Tables. Journal of the American Statistical Association, 79 (388), 762-771.

Dempster, A.P.; Laird, N.M.; Rubin, D.B (1977) Maximum likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B (Methodological), 39(1), 1-38.

Goodman, L.A. (1974) Exploratory Latent Structure Analysis using both Identifiable and Unidentifiable Models. Biometrika, 61(2), 215-231.

Henry, N.W. and Lazarsfeld, P. F. (1968) Latent Structure Analysis. Houghton Mifflin, Boston.

Read, D, T. & Cressie, N. (1988) Goodness-of-Fit Statistics for Discrete Multivariate Data. Springer, New York.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

NONPARAMETRIC INCIDENCE AND LATENCY ESTIMATION IN MIXTURE CURE MODELS

A. López-Cheda[1], R. Cao[1], M.A. Jácome[1] and I. Van Keilegom[2]

[1]MODES group, University of A Coruña; ana.lopez.cheda@udc.es, rcao@udc.es, majacome@udc.es
[2]Université catholique de Louvain; ingrid.vankeilegom@uclouvain.be

ABSTRACT

A completely nonparametric method for the estimation of mixture cure models is proposed in this paper. The nonparametric estimator of the incidence introduced by Xu and Peng (2014) is extensively studied and a nonparametric estimator of the latency is presented. These estimators, which are based on the Beran estimator of the conditional survival function, are proved to be the local maximum likelihood estimators. An iid representation is obtained for the nonparametric incidence estimator. As a consequence, an asymptotically optimal bandwidth is found. Moreover, a bootstrap bandwidth selection method for the nonparametric incidence estimator is proposed. The introduced nonparametric estimators are compared with existing semiparametric approaches in a simulation study, in which the performance of the bootstrap bandwidth selector is also assessed. Finally, the presented method is applied to a database of colorectal cancer from the University Hospital of A Coruña (CHUAC).

Keywords: Survival analysis, censored data, maximum likelihood, kernel estimation, bootstrap bandwidth selector.

1. INTRODUCTION

Thanks to the effectiveness of current cancer treatments, the proportion of patients who get cured (or who at least survive for a long time) is increasing over time. Therefore, for cancer studies a cure fraction should be incorporated into survival models, and a standard survival model would be inappropriate.

There are two main classes of cure models: mixture and non-mixture models. In this paper we consider a model which belongs to the so called two-component mixture cure models. The mixture cure model was proposed by Boag (1949) and it explicitly expresses the survival function as a mixture of two types of patients: those who are cured and those who are not. An advantage of this model is that it allows the covariates to have different influence on cured and uncured patients. The target of these models is to estimate the cured proportion (incidence) and also the probability of survival of the uncured patients up to a given time point (latency). Maller and Zhou (1996) provided a detailed review of this model.

In mixture cure models, the incidence is usually assumed to have a logistic form and the latency is usually estimated parametrically. Due to the fact that the effects of the covariate cannot always be well approximated using parametric or semiparametric methods, a nonparametric approach is needed. In the literature, some nonparametric methods for the estimation of the cure rate have been studied: Maller and Zhou (1992) proposed a consistent nonparametric estimator of the incidence, but it cannot handle covariates. In order to overcome this drawback, Laska and Meisner (1992) proposed another nonparametric estimator of the cure rate, but it only works for discrete covariates. Furthermore, Wang et al. (2012) proposed a cure model with a nonparametric form in the cure probability. To ensure model identifiability, they assumed a nonparametric proportional hazards model for the hazard function. More recently, Xu and Peng (2014) extended the existing work by proposing a nonparametric incidence estimator which allows for a continuous covariate.

The present paper will study that nonparametric incidence estimator deeply and fill this important gap for the latency function.

In this paper, we propose a two-component mixture model with nonparametric forms for both the cure probability and the survival function of the uncured individuals. Identifiability of the model is guaranteed by considering the time beyond which we have no interest as infinity and by identifying censored subjects beyond the largest observable failure time as cured, so that our estimator does not overestimate the true cure rate.

The rest of the article is organized as follows. In Section 2 we give a detailed description of our nonparametric mixture cure model, we study the estimator of the incidence proposed in Xu and Peng (2014) and introduce a nonparametric estimator of the latency. Moreover, we address the identifiability problem. We also present a local maximum likelihood result as well as an iid representation and the asymptotic mean squared error for the nonparametric incidence estimator. A bootstrap bandwidth selection method is introduced in Section 3. Section 4 includes a comparison between these nonparametric estimators and the semiparametric ones proposed in Peng and Dear (2000) in a simulation study and assesses the practical performance of the bootstrap bandwidth selector. Finally, in Section 5 we apply the proposed nonparametric method to real data related to colorectal cancer patients in CHUAC.

2. NONPARAMETRIC MIXTURE CURE MODEL

2.1 Notation

Let ξ be a binary variable where $\xi = 0$ indicates if the individual belongs to the susceptible group (the individual will eventually experience the event of interest if followed for long enough) and $\xi = 1$ indicates if the subject is cured. The proportion of cured patients and the survival function in the group of uncured patients can depend on certain characteristics of the subject, represented by a set of covariates X. Let $p(x) = P(\xi = 0|X = x)$ be the conditional probability of not being cured, and let Y be the time to occurrence of the event. When $\xi = 1$ it is assumed that $Y = \infty$.

The conditional distribution function of Y is $F(t|x) = P(Y \leq t|X = x)$. Note that the corresponding survival function, $S(t|x)$, is improper when cured patients exist, since $\lim_{t\to\infty} S(t|x) = 1 - p(x) > 0$. The conditional survival function of Y given that the subject is not cured is denoted by $S_0(t|x) = P(Y > t|X = x, \xi = 0)$. Then, the mixture cure model can be written as:

$$S(t|x) = 1 - p(x) + p(x)S_0(t|x), \tag{1}$$

where $1 - p(x)$ is the incidence and $S_0(t|x)$ is the latency. We assume that each individual is subject to random right censoring and that the censoring time, C, with distribution function G, is independent of Y given the covariates X. Let $T = \min(Y, C)$ be the observed time with distribution function H and $\delta = I(Y \leq C)$ the uncensoring indicator. Observe that $\delta = 0$ for all the cured patients, and it also happens for uncured patients with censored lifetime. Let X be a univariate continuous covariate with density function $m(x)$. Therefore, the observations will be $(X_i, T_i, \delta_i), i = 1, \ldots, n$ independent and identically distributed copies of the random vector (X, T, δ).

Xu and Peng (2014) introduced the following kernel type estimator of the incidence:

$$1 - \hat{p}_h(x) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right) = \hat{S}_h(T^1_{\max}|x), \tag{2}$$

where $T^1_{\max} = \max_{i:\delta_i=1}(T_i)$ is the largest uncensored failure time,

$$\hat{S}_h(t|x) = \prod_{T_i \leq t} \left(1 - \frac{\delta_{(i)} B_{h(i)}(x)}{\sum_{r=i}^{n} B_{h(r)}(x)}\right), \tag{3}$$

is the generalized Kaplan-Meier estimator by Beran (1981) to estimate the conditional survival function with covariates, and

$$B_{h(i)}(x) = \frac{K_h(x - X_{(i)})}{\sum_{j=1}^{n} K_h(x - X_{(j)})} \tag{4}$$

are the Nadaraya-Watson (NW) weights with $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ the rescaled kernel with bandwidth $h \to 0$. Here $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(n)}$ are the ordered T_i's, and $\delta_{(i)}$ and $X_{(i)}$ are the corresponding uncensoring indicator and covariate concomitants. We will also denote $\hat{F}_h(t|x) = 1 - \hat{S}_h(t|x)$ the Beran estimator of $F(t|x)$.

Using (1), we propose the following nonparametric estimator of the latency:

$$\hat{S}_{0,h}(t|x) = \frac{\hat{S}_h(t|x) - (1 - \hat{p}_h(x))}{\hat{p}_h(x)}. \tag{5}$$

The identifiability of a cure model allows to obtain unique estimates of the functions of interest. In a cure model, all observed uncensored subjects ($\delta_i = 1$) are necessarily uncured ($\xi_i = 0$); but it is impossible to say if an observed censored individual ($\delta_i = 0$) belongs to the susceptible group ($\xi_i = 0$) or to the cured group ($\xi_i = 1$), because some censored subjects may experience failures beyond the study period. This leads to difficulties in making a distinction between models with high incidence of being susceptible and long tails of the latency distribution and low incidence of being susceptible and short tails of the latency distribution. Lemma 1 addresses this problem.

Lemma 1. *Let T^+ be an arbitrary large time (possibly $T^+ = \infty$) and let D be the support of X. Model (1), with $p(x)$ and $S_0(t|x)$ unspecified, is identifiable if T^+ is large enough such that $S_0(T^+|x) = 0$ for all $x \in D$ and $t < T^+$.*

2.2 Asymptotic properties

The Beran estimator (3) has been deeply studied in the literature (Dabrowska, 1989; González-Manteiga and Cadarso-Suárez, 1994; van Keilegom and Veraverbeke, 1997a,b among others). If $\bar{G}(t|x) = 1 - G(t|x)$ denotes the proper conditional survival function of the censoring time C, $H(t|x) = P(T \leq t|X = x)$ the conditional distribution of the observations, and assuming that $S_0(t|x)$ is a proper survival function, we define:

$$\begin{aligned} \tau_H(x) &= \sup\{t : H(t|x) < 1\}, \\ \tau_{S_0}(x) &= \sup\{t : S_0(t|x) > 0\}, \\ \tau_G(x) &= \sup\{t : G(t|x) < 1\}. \end{aligned}$$

Since $S(t|x)$ is an improper survival function and $1 - H(t|x) = S(t|x) \times \bar{G}(t|x)$, we have $\tau_H(x) = \tau_G(x)$. Let $\tau_0 = \sup_{x \in D} \tau_{S_0}(x)$. As in Xu and Peng (2014), we assume

$$\tau_0 < \tau_G(x), \forall x \in D. \tag{6}$$

This condition states that the support of the censoring variable is not contained in the support of Y, which guarantees that censored subjects beyond the largest observable failure time are cured. Hence, our estimator does not overestimate the true cure rate. A similar assumption was used by Maller and Zhou (1992, 1994) in homogeneous cases. As pointed out in Laska and Meisner (1992), if the censoring variable takes values always below a time $\tau_G < \tau_0$, for example in a clinical trial with a fixed maximum follow-up period, the largest uncensored observation $T^1_{\max}$ may occur at a time not larger than τ_G and therefore always before τ_0. In such a case, for a large sample size, the estimator in (2) is an estimator of $1 - p(x) + p(x)S_0(\tau_G)$ which is strictly larger than $1 - p(x)$. This comment shows the need of considering the length of follow-up in the design of a clinical trial carefully, so that $S_0(\tau_G)$ is sufficiently small to take the estimator (2) of $1 - p(x) + p(x)S_0(\tau_G)$ as a good estimator of $1 - p(x)$ for practical purposes. The simulations in Xu and Peng (2014) show that if the censoring distribution $G(t|x)$ has a heavier tail than $S_0(t|x)$, the estimates from the proposed method will tend to have smaller biases regardless of the value of $\tau_{S_0}(x)$.

Define $H^1(t|x) = P(T \leq t, \delta = 1|X = x)$ the conditional distribution of the uncensored observations. We further assume that:

(A1) X, Y and C are absolutely continuous random variables.

(A2) (a) Let $I = [x_1, x_2]$ be an interval contained in the support of m, and $I_\delta = [x_1 - \delta, x_2 + \delta]$ for some $\delta > 0$ such that

$$0 < \gamma = \inf[m(x) : x \in I_\delta] < \sup[m(x) : x \in I_\delta] = \Gamma < \infty$$

and $0 < \delta\Gamma < 1$. And for all $x \in I_\delta$ the random variables Y and C are conditionally independent given $X = x$.

(b) There exist $a, b \in \mathbb{R}$, with $a < b$ satisfying $1 - H(t|x) \geq \theta > 0$ for $(t, x) \in [a, b] \times I_\delta$.

(A3) The first derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the first derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A4) The second derivative of the function $m(x)$ exists and is continuous in $x \in I_\delta$ and the second derivatives with respect to x of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous and bounded in $(t, x) \in [0, \infty) \times I_\delta$.

(A5) The first derivatives with respect to t of the functions $G(t|x)$, $H(t|x)$, $H^1(t|x)$ and $S_0(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A6) The second derivatives with respect to t of the functions $H(t|x)$ and $H^1(t|x)$ exist and are continuous in $(t, x) \in [a, b] \times D$.

(A7) Let us define $H_{c,1}(t) = P(T < t|\delta = 1)$. The first and second derivatives of the distribution and subdistribution functions $H(t)$ and $H_{c,1}(t)$ are bounded away from zero in $[a, b]$. Moreover, $H'_{c,1}(\tau_0) > 0$.

(A8) The kernel function K is a symmetric density vanishing outside $(-1, 1)$ and the total variation of K is less than some $\lambda < \infty$.

(A9) The kernel K is a twice differentiable function with K'' continuous.

(A10) The functions $H(t|x)$, $S_0(t|x)$ and $G(t|x)$ have bounded second-order derivatives with respect to x for any given value of t.

(A11) $\displaystyle\int_0^\infty \frac{dH^1(t|x)}{(1 - H(t|x))^2} < \infty \ \forall x \in I.$

The consistency and asymptotic normality of the incidence estimator are proved in Xu and Peng (2014). In the next theorem we show that both the proposed nonparametric incidence and latency estimators are the local maximum likelihood estimators of $1 - p(x)$ and $S_0(t|x)$.

Theorem 1. *The kernel type estimators $1 - \hat{p}_h(x)$ and $\hat{S}_{0,h}(t|x)$, given in (2) and (5) respectively, are the local maximum likelihood estimators of $1 - p(x)$ and $S_0(t|x)$ for the mixture cure model (1), for any $x \in D$ and $t \geq 0$.*

We also obtain an iid representation of the incidence estimator.

Theorem 2. *Under assumptions $(A1) - (A11)$ and assuming (6), for a sequence of bandwidths satisfying $nh^5(\ln n)^{-1} = O(1)$ and $\ln n/(nh) \to 0$, then*

$$(1 - \hat{p}_h(x)) - (1 - p(x)) = (1 - p(x)) \sum_{i=1}^{n} \tilde{B}_{hi}(x) \xi(T_i, \delta_i, x) + R_n(x)$$

where

$$\tilde{B}_{hi}(x) = \frac{\frac{1}{nh} K\left(\frac{x - X_i}{h}\right)}{m(x)},$$

$$\xi(T_i, \delta_i, x) = \frac{I(\delta_i = 1)}{1 - H(T_i|x)} - \int_0^{T_i} \frac{dH^1(t|x)}{(1 - H(t|x))^2}$$

and

$$\sup_{x\in I}|R_n(x)| = O\left(\left(\frac{\ln n}{nh}\right)^{3/4}\right) \text{ a.s.}$$

Finally, from the representation in Theorem 2 with straightforward calculations, the asymptotic expression of the mean squared error of the incidence estimator is given by:

$$AMSE_x(h) = \frac{1}{nh}(1-p(x))^2 c_K \sigma^2(x) + \left[h^2 \frac{1}{2} d_K (1-p(x))\mu(x)\right]^2, \tag{7}$$

where the first term corresponds to the asymptotic variance and the second one to the asymptotic squared bias, with $d_K = \int v^2 K(v)dv$, $c_K = \int K^2(v)dv$ and, following a notation similar to that in Dabrowska (1992):

$$\mu(x) = \frac{2\Phi'(x,x)m'(x) + \Phi''(x,x)m(x)}{m(x)},$$

$$\sigma^2(x) = \frac{1}{m(x)}\int_0^\infty \frac{dH^1(t|x)}{(1-H(t|x))^2},$$

where

$$\Phi(u,x) = \int_0^\infty \frac{dH^1(t|u)}{1-H(t|x)} - \int_0^\infty \frac{1-H(t|u)}{(1-H(t|x))^2} dH^1(t|x),$$

with $\Phi'(u,x) = \partial/(\partial u)\Phi(u,x)$ and $\Phi''(u,x) = \partial^2/(\partial u^2)\Phi(u,x)$.

3. BANDWIDTH SELECTION

The choice of the bandwidth is a crucial issue for kernel estimation, since it controls the trade-off between bias and variance. When the bandwidth is too large, then $B_{hi}(x) \simeq n^{-1}$ and the nonparametric estimator reduces to the homogeneous cure rate estimator proposed by Maller and Zhou (1992), in which the effect of the covariate is left out, increasing the bias. On the other hand, if the bandwidth is too small, then the error of estimation shoots up at the expense of the variance. Several methods for the selection of the smoothing parameter have been proposed in the literature. The asymptotically optimal local bandwidth to estimate the cure rate in the sense of minimizing the asymptotic expression of the MSE_x in (7) is given by:

$$h_{x,AMSE} = \left(\frac{c_K \sigma^2(x)}{d_K^2 \mu^2(x)}\right)^{1/5} n^{-1/5}.$$

Considering Dabrowska (1989), a plug-in bandwidth selector can be obtained by replacing the unknown functions in $\mu(x)$ and $\sigma^2(x)$ by consistent nonparametric estimates, giving rise to a never-ending process which seems even harder than the original problem of incidence estimation. On the other hand, unfortunately, the finite-sample behavior of the cross validation (CV) bandwidth selector in this context turned out to be disappointing; the CV bandwidth was highly variable and tended to undersmooth its kernel estimate.

3.1. Bootstrap bandwidth selector

The bootstrap method consists of minimizing a bootstrap estimate of $MSE_x(h_x)$, rather than minimizing the estimation of its asymptotic expression in (7). The bootstrap bandwidth is the minimizer of the bootstrap version of $MSE_x(h_x)$, approximated using Monte Carlo by:

$$MSE^*_{x,g_x}(h_x) \simeq \frac{1}{B}\sum_{b=1}^{B}(\hat{p}^{*b}_{h_x,g_x}(x) - \hat{p}_{g_x}(x))^2, \tag{8}$$

where $\hat{p}^{*b}_{h_x,g_x}(x)$ is the kernel estimator of p using bandwidth h_x and based on the b-th bootstrap resample, and $\hat{p}_{g_x}(x)$ is the kernel estimator of p computed with the original sample and pilot bandwidth g_x.

In this paper, we consider the simple weighted bootstrap by Li and Datta (2001), without resampling the covariate X but just fixing the covariate sample in the resamples, which is equivalent to the original simple weighted bootstrap. For fixed x and $i = 1, \ldots, n$, we set $X_i^* = X_i$ and generate a pair (T_i^*, δ_i^*) from the weighted empirical distribution $\hat{F}_{g_x}(\cdot, \cdot | X_i^*)$, where

$$\hat{F}_{g_x}(u, v|x) = \sum_{i=1}^{n} B_{g_x i}(x) I(T_i \leq u, \delta_i \leq v)$$

and $B_{g_x i}(x)$ is the NW weight in (4) with pilot bandwidth g_x. The resulting bootstrap resample is $\{(X_1, T_1^*, \delta_1^*), \ldots, (X_n, T_n^*, \delta_n^*)\}$.

The procedure for obtaining the bootstrap bandwidth selector for a fixed x is as follows:

1. Select the pilot bandwidth g_x and generate B bootstrap resamples $\left\{ \left(X_i^{(b)}, T_i^{*(b)}, \delta_i^{*(b)} \right), i = 1 \ldots, n \right\}$, $b = 1, \ldots, B$.
2. With the b-th bootstrap resample ($b = 1, ..., B$), compute $\hat{p}^{*b}_{h_l, g_x}$ with bandwidth h_l $l = 1, 2, \ldots, L$.
3. With the original sample and pilot bandwidth g_x, compute $\hat{p}_{g_x}(x)$.
4. For each bandwidth h_l in the grid, compute the Monte Carlo approximation of $MSE^*_{x,g_x}(h_l)$ given by (8).
5. The bootstrap bandwidth is $h_x^* = \arg \min_{l=1,\ldots,L} \frac{1}{B} \sum_{b=1}^{B} (\hat{p}^{*b}_{h_l, g_x}(x) - \hat{p}_{g_x}(x))^2$.

The optimal pilot bandwidth g_x should be chosen so that it minimizes the dominant term of the asymptotic representation of $MSE^*_{x,g_x}(h_x)$. Similarly as in Theorem 3 in Van Keilegom and Veraverbeke (1997b) for the bootstrapped Beran estimator $\hat{F}^*_{h_x, g_x}(\cdot|x)$, and keeping in mind that $\hat{p}^*_{h_x, g_x}(x) = \hat{F}^*_{h_x, g_x}(T^1_{\max}|x)$, with $h_x = Cn^{-1/5}$ for some $C > 0$, the pilot bandwidth g_x must be chosen to minimize the MISE of $\hat{\ddot{H}}_{g_x}(t|x) = \sum_{i=1}^{n} B^{(2)}_{g_x(i)}(x) I(T_{(i)} \leq t)$, the kernel estimator of $\ddot{H}(t|x) = (\partial^2/\partial x^2) H(t|x)$ with NW weights, bandwidth g_x and $B^{(2)}_{g_x(i)}(x) = \frac{\partial^2}{\partial x^2} B_{g_x(i)}(x)$.

Lemma 2. *Assume (A1)-(A4), (A6), (A8), (A9), $h_x = Cn^{-1/5}$ for some $C > 0$, $g_x \to 0$ and $ng_x^5 (\ln n)^{-1} \to \infty$. Then, the asymptotically optimal pilot bandwidth g_x for the bootstrap resampling, which minimizes the MISE of $\hat{\ddot{H}}_{g_x}(t|x)$, is:*

$$g_x = \left(\frac{5 c_{K''} m(x) \int_0^\infty H(t|x)(1 - H(t|x))\, dt}{\int_0^\infty \left(\frac{\partial^4}{\partial x^4} (H(t|x) m(x)) - \frac{\partial^2}{\partial x^2} (H(t|x) m''(x)) \right)^2 dt} \right)^{1/9} n^{-1/9}. \quad (9)$$

The proof of Lemma 2 derives directly from the proof of the main result in Cao (1991).

Remark: The bandwidth sequence $g_x = g_n$ is typically asymptotically larger than $h_x = h_n$. This oversmoothing pilot bandwidth is required for the bootstrap bias and variance to be asymptotically efficient estimators for the bias and variance terms. The order $n^{-1/9}$ of this asymptotically optimal pilot bandwidth satisfies the conditions in Theorem 1 of Li and Datta (2001), and is also equal to the order obtained by Cao and González-Manteiga (1993) for the uncensored case.

4. SIMULATION STUDY

In this section the proposed nonparametric estimators are compared with the semiparametric estimators in Peng and Dear (2000), which are implemented in the *smcure* package in R (Cai et al, 2012). These estimators assume a logistic expression for the incidence and a proportional hazards (PH) model for the latency.

The simulation study is carried out with two purposes: (a) to evaluate the finite sample performance of the nonparametric estimators with respect to the semiparametric estimators, and (b) to assess the practical behavior of the bootstrap bandwidth selector. Two different models are considered, and for both the censoring times are generated according to the exponential distribution with mean 1/0.3 and the covariate X is $U(-20, 20)$.

Model 1: For comparison reasons, this simulated configuration is the same as the so-called mixture cure (MC) model considered in Xu and Peng (2014). The data are generated from a logistic-exponential MC model, where the probability of not being cured is

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)},$$

with $\beta_0 = 0.476$ and $\beta_1 = 0.358$, and the survival function of the uncured subjects is:

$$S_0(t|x) = \begin{cases} \dfrac{\exp(-\lambda(x)t) - \exp(-\lambda(x)\tau_0)}{1 - \exp(-\lambda(x)\tau_0)} & \text{if } t \leq \tau_0 \\ 0 & \text{if } t > \tau_0 \end{cases},$$

where $\tau_0 = 4.605$ and $\lambda(x) = \exp((x + 20)/40)$. The percentage of censored data is 62% and of cured data is 53%. In this model the incidence is a logistic function and the latency fulfills the PH assumption, so the assumptions of the semiparametric estimator by Peng and Dear (2000) are satisfied, and therefore it is expected to give very good results.

Model 2: The data are generated from a cubic logistic-exponential mixture model, where the probability of not being cured is:

$$p(x) = \frac{\exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right)}{1 + \exp\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\right)},$$

with $\beta_0 = 0.0476$, $\beta_1 = -0.2558$, $\beta_2 = -0.0027$ and $\beta_3 = 0.0020$, and the survival function of the uncured subjects is:

$$S_0(t|x) = \frac{1}{2}\left(\exp(-\alpha(x)t^5) + \exp(-100t^5)\right), \quad \text{with} \quad \alpha(x) = \frac{1}{5}\exp((x + 20)/40).$$

The percentages of censored and cured data are 58% and 47%, respectively. The incidence is not a logistic function and the effect of the covariate on the failure time of the uncured patients does not fit a PH model. So, the results will show the gain of using the proposed nonparametric estimators, that do not require any parametric or semiparametric assumptions, with respect to the semiparametric ones.

4.1. Efficiency of the nonparametric estimators

The mean squared error (MSE) of the incidence estimators and the mean integrated squared error (MISE) of the latency estimators were approximated from a number of $m = 1000$ samples of size $n = 100$ and for a grid of bandwidths, from $h_0 = 1.2$ to $h_{99} = 20$ for the incidence function and from $h_0 = 10$ to $h_{99} = 40$ for the latency.

Regarding the MSE of the incidence estimators, Figure 1 shows that in Model 1 there is a range of bandwidths, from $h = 4.8$ to $h = 8.5$, for which the nonparametric estimator is quite competitive with respect to the semiparametric estimator in some values x of the covariate, and it works much better when the value of the covariate is around $x = 0$. In Model 2, as expected, the nonparametric estimator outperforms the semiparametric one, regardless how small or large the bandwidth is, except for the extreme values of the interval $[-20, 20]$.

Considering the latency estimators, it is noteworthy that in Model 1, in which the assumptions of the semiparametric estimator are fulfilled, there is a wide range of bandwidths for which the MISE of the nonparametric estimator is smaller than the MISE of the semiparametric estimator for values of the covariate greater than 0 (see Figure 1). In Model 2, the nonparametric estimator of the latency outperforms the semiparametric estimator for all the values of the covariate with almost any value of the bandwidth, regardless how large it is, as long as the bandwidth is greater than 17.

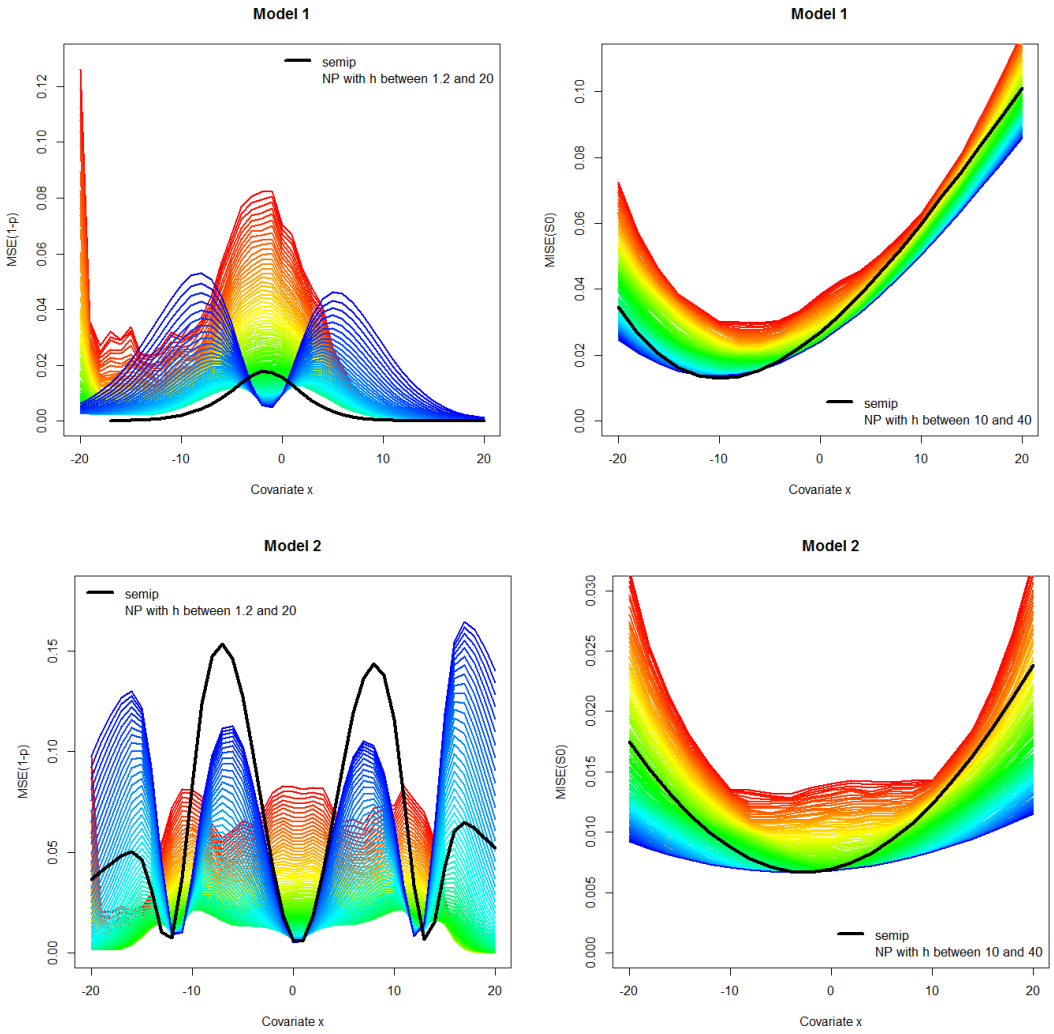


Figura 1: On the left, MSE for the semiparametric (black line) and nonparametric estimators of $1-p(x)$ computed with bandwidths from $h_0 = 1.2$ (red line) to $h_{99} = 20$ (blue line). On the right, MISE for the semiparametric (black line) and nonparametric estimators of $S_0(t|x)$ computed with bandwidths from $h_0 = 10$ (red line) to $h_{99} = 40$ (blue line). The data were generated from Model 1 (top) and Model 2 (bottom).

In short, the nonparametric estimators are quite competitive in situations where the assumptions of the semiparametric estimator are fulfilled, and clearly better when whose assumptions are not true. The efficiency of the nonparametric estimators depends on the choice of the bandwidth, but although in practice the optimal value of the bandwidth remains unknown, the simulations show that, for quite wide ranges of bandwidth values, the proposed nonparametric methods outperform the existing semiparametric estimator of Peng and Dear (2000).

4.2. Efficiency of the bootstrap bandwidth selector

To assess the practical behavior of the bootstrap bandwidth selector for the proposed nonparametric estimator of the incidence, the bootstrap bandwidth h_x^* is compared to the optimal $h_{x,MSE}$ approximated by Monte Carlo for a number of $m = 1000$ trials of sizes $n = 50, 100$ and 200.

Note that the computation of h_x^* requires to minimize MSE^*_{x,g_x} in $h_l, l = 1, \ldots, L$ for each value x of the covariate, which is a highly time consuming algorithm. For that reason a two-step method, with a double search in each step, was carried out. In the first step, $B = 80$ bootstrap resamples

were drawn and the minimization was performed in a number of 21 bandwidths equispaced on a logarithmic scale, in the first search from $h_0 = 0.2$ to $h_{20} = 25$, whereas in the second search the grid was centered around the optimal bandwidth obtained in the first search. In the second step, a double search was carried out in a similar way as in the first step, but now with two differences: using $B = 1000$ bootstrap resamples and considering a finer smaller grid of 5 bandwidths in both the first and second search.

The asymptotically optimal pilot bandwidth g_x has a quite involved expression, see (9), which depends on unknown functions in a rather awkward way. In view of the fact that the choice of g_x had a low effect on the final bootstrap bandwidth, a naive selector keeping the $n^{-1/9}$ optimal order was used. Since the distribution of the covariate is uniform, the following global pilot bandwidth was considered:

$$g = \frac{X_{(n)} - X_{(1)}}{10^{7/9}} n^{-1/9}. \tag{10}$$

For a naive pilot bandwidth selector if the distribution for X can not be assumed uniform, see Section 5.

The performance of the bootstrap bandwidth for Models 1 and 2 is shown in Figure 2. The optimal $h_{x,MSE}$, approximated by Monte Carlo, is displayed together with the mean and the 25th and 75th percentiles of the 1000 bootstrap bandwidths h_x^*. The bootstrap bandwidth h_x^* approaches $h_{x,MSE}$, adapting properly to the shape of $h_{x,MSE}$ for any sample size. Note that the optimal bandwidth $h_{x,MSE}$ has got peaks at the values x of the covariate for which $p''(x) = 0$. In other terms, those peaks only occur at points x for which the optimal bandwidth is infinitely large because the best choice is to smooth as much as possible, and the best local fit is a global fit.

5. APPLICATION TO REAL DATA

Both the semiparametric and nonparametric estimators are applied to a real dataset of 414 colorectal cancer patients from CHUAC (Complejo Hospitalario Universitario de A Coruña), Spain. Besides the lifetimes, two important variables are the stage, from 1 to 4, and the age, from 23 to 103. About 50% of the observations are censored, with the percentage of censoring varying from 30% (stage 4) to almost 71% (stage 1). For the computation of the estimators, the age of the patients has been considered as the covariate.

For the bootstrap bandwidth selector, a naive pilot bandwidth selector has been proposed in (10) if the distribution of X is uniform. Taking into account that in this case the distribution of the covariate is not uniform (see Figure 3), the following local pilot bandwidth was used instead:

$$g_x = \frac{d_k^+(x) + d_k^-(x)}{2} 100^{1/9} n^{-1/9},$$

where $d_k^+(x)$ (similarly $d_k^-(x)$) is the distance from x to the k-th nearest neighbor on the right (on the left), and k a suitable integer depending on the sample size. Numerical results showed that a good choice is to consider $k = n/4$. If there are not at least k neighbors on the right (or left), use $d_k^+(x) = d_k^-(x)$ (or $d_k^-(x) = d_k^+(x)$) respectively.

Figure 3 shows the estimations of the probability of being cured for the different stages with respect to the age of the patients. For the nonparametric estimator, and following Cao, Janssen and Veraverbeke (2001), alongside the bootstrap bandwidth a smoothed bandwidth was also used considering 5 neighbors on each side. Note that the effect of the age on the probability of being cured changes with the stage. For example, in Stage 1, patients have a probability of survival between 25% and 65%, depending on the age; whereas in Stage 3, for patients above 60, in a 10 years gap that probability decreases considerably from 40% to almost 0%. It is important to highlight the difference between the nonparametric and the semiparametric curves, that seems to indicate that the logistic model is not valid for the data. The results in Stage 4 deserve some comments. A total of 11 in the 12 greatest lifetimes in Stage 4, including the largest lifetime, are uncensored and, consequently, uncured. This causes that the nonparametric estimation of the probability of being cured is equal to 0. Although it should not be stated that it is impossible for a patient with Stage 4 colorectal cancer to survive, this estimation reinforce the assertion that long-term survival in patients with Stage 4 colorectal cancer is uncommon (Miyamoto et al, 2015). This fact, far from being a weakness of the nonparametric method, is an important advantage,

since it allows to detect situations in which introducing the possibility of cure does not contribute to improve the model.

In Figure 4 we show the latency estimation for Stages 1, 2, 3 and 4 for age 76. The nonparametric estimator $\hat{S}_{0,h_x}$ is computed with five different constant bandwidths: $h = 10, 15, 20, 25$ and 30. It is noteworthy that the bandwidth selection does not influence too much the latency estimation. This is due to the high density of the covariate around this age (see Figure 3).

REFERENCES

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society, Series B, 11, 15–53.

Cai, C.; Zou Y.; Peng, Y. and Zhang, J. (2012). smcure: Fit Semiparametric Mixture Cure Models. R package version 2.0. http://CRAN.R-project.org/package=smcure

Cao, R. (1991). Rate of convergence for the wild bootstrap in nonparametric regression. The Annals of Statistics, 19, 2226–2231.

Cao, R. and González-Manteiga, W. (1993). Bootstrap methods in regression smoothing. Journal of Nonparametric Statistics, 2, 379–388.

Cao, R.; Janssen, P. and Veraverbeke, N. (2001). Relative density estimation and local bandwidth selection for censored data. Computational Statistics and Data Analysis, 36, 497–510.

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. The Annals of Statistics, 17, 1157–1167.

Dabrowska, D. (1992). Variable bandwidth conditional Kaplan-Meier estimate. The Scandinavian Journal of Statistics, 19, 351–361.

González-Manteiga, W. and Cadarso-Suárez, C. (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. Journal of Nonparametric Statistics, 4, 65–78.

Laska, E. M. and Meisner, M. J. (1992). Nonparametric estimation and testing in a cure model. Biometrics, 48, 1223–1234.

Li, G. and Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. Annals of the Institute of Statistical Mathematics, 53, 708–729.

Maller, R. A. and Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. Biometrika, 79, 731–739.

Maller, R. A. and Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. Journal of the American Statistical Association, 89, 1499–1506.

Maller, R. A. and Zhou, S. (1996). Survival analysis with long-term survivors. Chichester, U.K.:Wiley.

Miyamoto, Y.; Hayashi, N.; Sakamoto, Y.; Ohuchi, M.; Tokunagam, R.; Kurashige, J.; Hiyoshi, Y.; Baba, Y.; Iwagami, S.; Yoshida, N.; Yoshida, M. and Baba, H. (2015). Predictors of long-term survival in patients with stage IV colorectal cancer with multi-organ metastases: a single-center retrospective analysis. International Journal of Clinical Oncology, DOI 10.1007/s10147-015-0835-2.

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. Biometrics, 56, 237–243.

Van Keilegom, I. and Veraverbeke, N. (1997a). Estimation and bootstrap with censored data in fixed design nonparametric regression. Annals of the Institute of Statistical Mathematics, 49, 467–491.

Van Keilegom, I. and Veraverbeke, N. (1997b). Weak convergence of the bootstrapped conditional Kaplan-Meier process and its quantile process. Communications in Statistics, Theory and Methods, 26, 853–869.

Wang, L.; Du, P. and Lian, H. (2012). Two-component mixture cure rate model with spline estimated nonparametric components. Biometrics, 68, 726–735.

Xu, J. and Peng, Y. (2014). Nonparametric cure rate estimation with covariates. The Canadian Journal of Statistics, 42, 1–17.

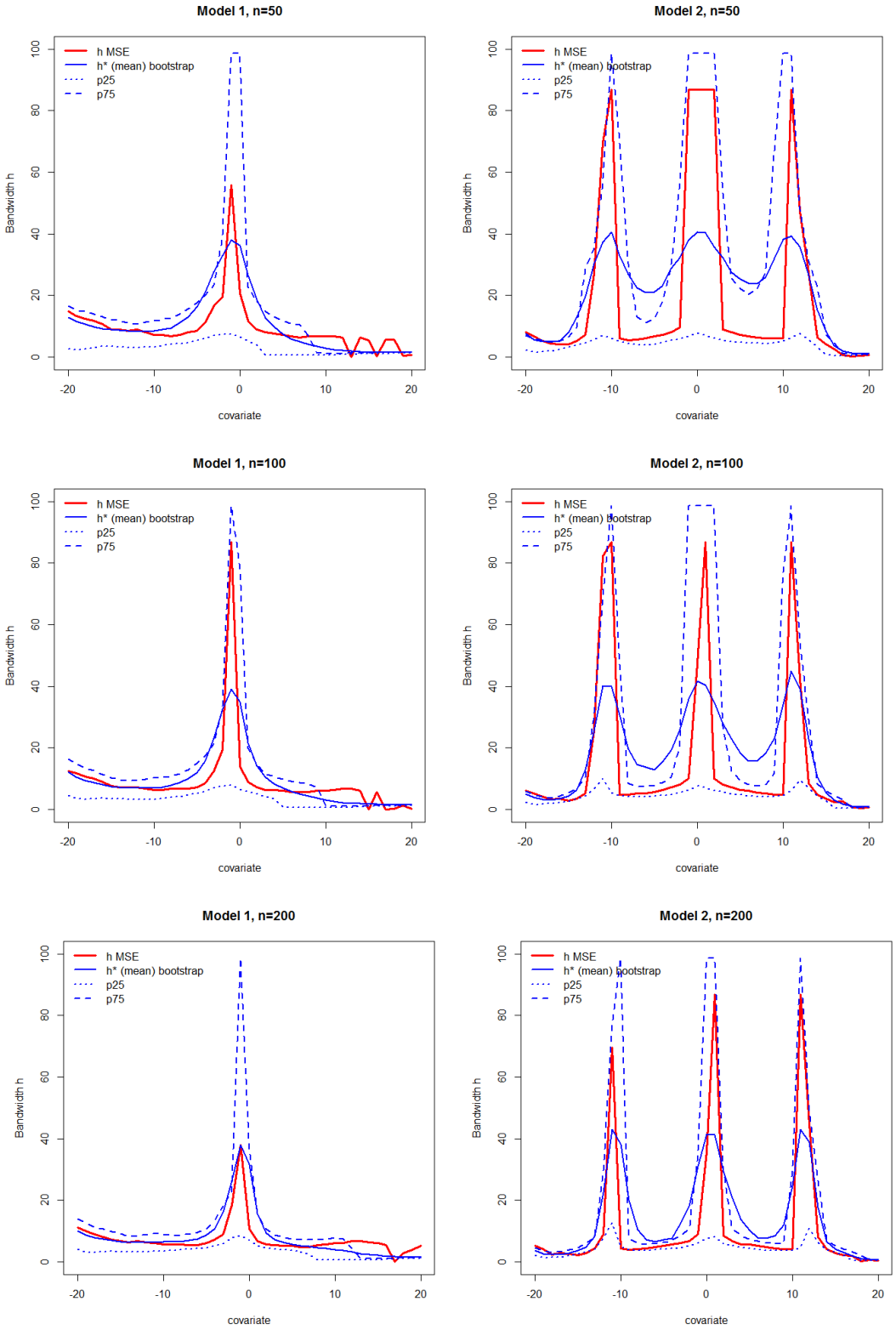


Figura 2: Optimal $h_{x,MSE}$ (red line), mean (blue line) and the 25th (dotted blue line) and 75th (dashed blue line) percentiles of the bootstrap bandwidth h_x^* along $m = 1000$ samples, with sample sizes $n = 50$ (top), $n = 100$ (center) and $n = 200$ (bottom), for Model 1 (left) and Model 2 (right).

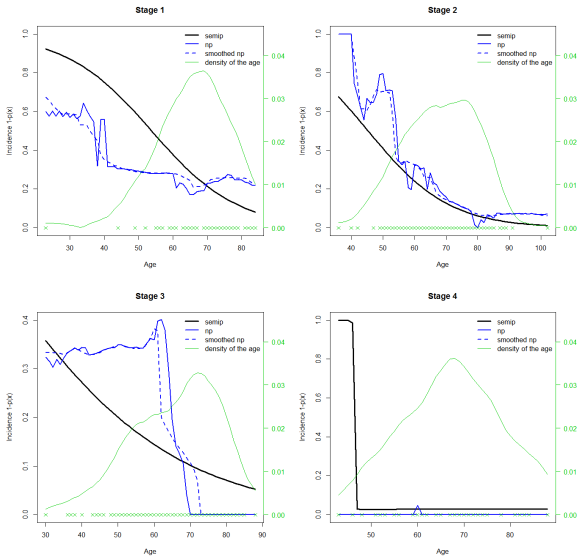


Figura 3: Semiparametric (black line) and nonparametric estimators of the probability of cure (incidence)

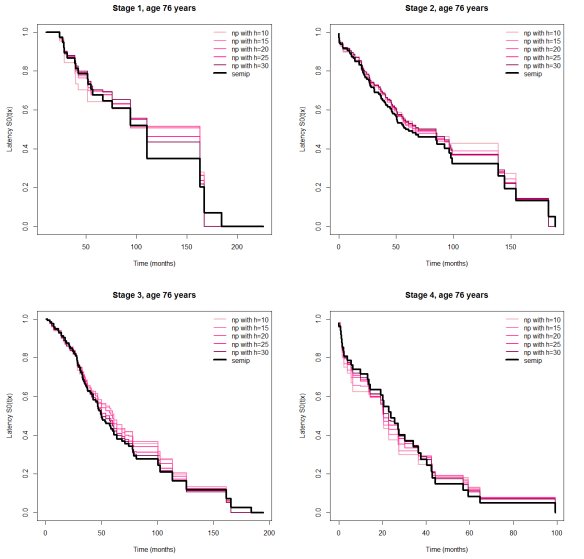


Figura 4: Estimated latency for patients of age 76 in Stages 1-4, using the semiparametric (black line)

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

BANDWIDTH SELECTION FOR THE PRESMOOTHED LOGRANK TEST

M.A. Jácome[1] and I. López-de-Ullibarri[2]

[1]Facultade de Ciencias, Universidade da Coruña, 15071 A Coruña, Spain
[2]Escuela Universitaria Politécnica, Universidade da Coruña, 15405 Ferrol, Spain

ABSTRACT

The most popular test in the two-sample random censorship model is the logrank test. It is based on the comparison of the Nelson-Aalen estimates of the corresponding cumulative hazard functions. A new logrank-type test, based on the presmoothed counterpart of the Nelson-Aalen estimators, has been shown to have the proper size under the null hypothesis, while improving the power over a wide range of alternatives. The success of the presmoothed methods depends on the choice of a smoothing parameter or bandwidth. We propose two data-driven bandwidth selectors that maximize the power while keeping the size of the test at a given level, and show via simulations the significant improvement of the presmoothed logrank test over the classical one. The application of the new test with both bandwidth selectors is illustrated with several real data examples.

Palabras e frases chave: Censoring; Survival analysis; Two-sample problem; Bootstrap; Data-driven bandwidth

1. INTRODUCTION

The most used of all distribution-free methods for testing the equality of two survival distributions under right random censoring is the logrank (LR) test. It can be shown to have optimality properties for detecting differences in hazard rates assuming that the hazards are proportional. Even if this assumption proves to be reasonable in many settings, situations of patent nonproportionality may still be found. This is the case if, for example, short- and long-term effects of a treatment have opposite signs. Weighted LR tests have been proposed with the aim of increasing efficiency in the presence of nonproportional hazards. Under weak conditions, Gill (1980) proved the consistency of weighted LR tests both against ordered hazard alternatives and, if the weight function is monotone, against stochastic ordering alternatives. The $G^{\rho,\gamma}$ class (Fleming and Harrington 1991) of weighted LR statistics would allow to construct a test with maximal power against any given nonproportional alternative. However, since in practice it is unknown how the survival of two groups may differ, power optimality is not guaranteed. Variants showing greater robustness against a wider range of alternatives have been developed like, e.g., tests that use supremum versions of weighted LR statistics (see Fleming and Harrington 1991).

Since the LR test and its weighted versions are based on the comparison of the Nelson-Aalen (NA) estimates of the cumulative hazard functions, the power would increase, regardless of the weight, if a more efficient estimator of the cumulative hazard were used. Cao, López-de-Ullibarri, Janssen and Veraverbeke (2005) and Jácome and Cao (2007) have demonstrated the efficiency of the so-called presmoothed NA estimator. While the NA estimator is a step function jumping only at the uncensored observations, the presmoothed NA estimator jumps also at the censored observations. Thus, more information on the local behavior of the lifetime distribution is provided, and the accuracy of the estimation is increased. The improvement comes at the price of having to choose a smoothing parameter or bandwidth. Jácome and López-de-Ullibarri (2014) introduced a new LR-type test that, instead of the NA estimator, is based on its presmoothed counterpart. The presmoothed LR test has the proper size under the null hypothesis, while at the same time

improving the power over the classical LR test for a wide range of alternatives, mainly with highly skewed distributions, large censoring rates and/or moderate sample sizes. The good behavior of the presmoothed LR test strongly depends on the choice of two bandwidths (one for each sample).

In this paper the problem of automatic bandwidth selection for the presmoothed LR test is addressed. In Section 3 we describe the main characteristics of the presmoothed LR test. Two methods for bandwidth choice are proposed in Section 3. In Section 4 we present the results of a simulation study carried out to investigate the power of the presmoothed LR test with the proposed bandwidth selectors. Finally, in Section 5 the performance of the test is illustrated with the analysis of two real datasets.

2. THE PRESMOOTHED LOGRANK TEST

Let us suppose that we have independent samples of n_1 and n_2 individuals from two different populations. For $j = 1, 2$, and $i = 1, 2, \ldots, n_j$ the ith subject in population j has latent lifetime T_{ij}, with survival function S_j and cumulative hazard function Λ_j. According to the right random censoring model, there exist censoring time variables C_{ij}, independent of the T_{ij}, whose distribution function is denoted by G_j. The available data for sample j consist of the pairs $\{(Z_{ij}, \delta_{ij}), i = 1, 2, \ldots, n_j\}$ where $Z_{ij} = \min(T_{ij}, C_{ij})$ is the ith observed time, with distribution function H_j, and $\delta_{ij} = \mathbf{1}(T_{ij} \leq C_{ij})$ is the censoring indicator.

We are interested in testing the null hypothesis of no differences between the two survival functions, i.e., $H_0 : S_1 = S_2$. The classical weighted LR statistic (Fleming and Harrington 1991) is

$$G^{LR}(t) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \int_0^t W(v) \frac{Y_1(v)Y_2(v)}{Y_1(v) + Y_2(v)} (d\Lambda_1^{NA}(v) - d\Lambda_2^{NA}(v)), \tag{1}$$

where $W(v)$ is a weight function, $Y_j(v) = \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} \geq v)$ is the size of the risk set at time v in sample j, and $\Lambda_j^{NA}(v) = \int_0^v Y_j^{-1}(u) dN_j(u)$ is the NA estimate of $\Lambda_j(v)$, $N_j(u) = \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} \leq u, \delta_{ij} = 1)$ being the number of failures in sample j occurred by time u. The weighted LR test (1) can be interpreted as a weighted comparison of NA estimates which are computed by considering only the uncensored observations.

A key function in presmoothing is $p_j(t) = P(\delta_{ij} = 1 | Z_{ij} = t)$, the conditional probability that the ith observed time in sample j is not censored. The presmoothing approach consists in replacing $\mathbf{1}(\delta_{ij} = 1)$ in the expression of the NA estimator with a smooth estimation of $p_j(Z_{ij})$. Thus, all observations contribute information regardless of whether they are or not censored. Since $p_j(t) = \mathrm{E}(\delta_{ij} | Z_{ij} = t)$, it can be estimated with a regression estimator. To avoid model misspecification problems, we use a nonparametric regression estimator like the Nadaraya-Watson (NW) kernel estimator:

$$\widehat{p}_{j,b_j}(t) = \frac{\sum_{i=1}^{n_j} K_{b_j}(t - Z_{ij}) \delta_{ij}}{\sum_{i=1}^{n_j} K_{b_j}(t - Z_{ij})},$$

where $b_j > 0$ is the presmoothing bandwidth, K is a kernel function, and $K_{b_j}(t) = b_j^{-1} K(b_j^{-1} t)$ is the rescaled kernel. It is assumed that K is a symmetric, twice continuously differentiable density, with bounded variation and compact support on the interval $[-1, 1]$.

The presmoothed weighted LR test statistic (Jácome and López-de-Ullibarri 2014) is defined by:

$$G_{\boldsymbol{b}}^P(t) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \int_0^t W(v) \frac{Y_1(v)Y_2(v)}{Y_1(v) + Y_2(v)} (d\Lambda_{1,b_1}^P(v) - d\Lambda_{2,b_2}^P(v)),$$

where $\boldsymbol{b} = (b_1, b_2)$, $\Lambda_{j,b_j}^P(v) = \int_0^v Y_j^{-1}(u) dN_{j,b_j}^P(u)$ is the presmoothed estimate of $\Lambda_j(v)$, and $N_{j,b_j}^P(u) = \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} \leq u) \widehat{p}_{j,b_j}(Z_{ij})$ is the presmoothed estimate of the number of failures in sample j that have occurred by time u. Note that the presmoothed weighted LR statistic compares the presmoothed estimators of Λ_1 and Λ_2 at each observed (censored or not) time. Also note that when $b_j \simeq 0$, then $\widehat{p}_{j,b_j}(Z_{ij}) \simeq \mathbf{1}(\delta_{ij} = 1)$, so the classical LR test is recovered from the presmoothed LR test by taking $\boldsymbol{b} = (0, 0)$.

The degree of smoothing, which is controlled by the bandwidth b_j, may strongly influence the aspect of $\widehat{p}_{j,b_j}(t)$, and thereby the effect of presmoothing. In estimation problems, the literature

provides many methods for smoothing bandwidth selection. In general, they aim at minimizing an error criterion when true curves are approximated by smooth estimates. For presmoothed estimators, it has been proved that they reduce mean squared error performance if the asymptotically optimal bandwidth is used. The improvement in efficiency is only of second order for cumulative hazard function estimation (Cao et al. 2005), but it is of first order in the case of density estimation (Cao and Jácome 2004). Empirical evidence obtained from simulation studies suggests that bandwidth choice is not so crucial, since presmoothed estimators tend to outperform their classical counterparts not only with the optimal bandwidth but also in a neighborhood of it (Cao and Jácome 2004; Cao et al. 2005). The `survPresmooth` package (López-de-Ullibarri and Jácome 2013) provides an implementation in R (R Core Team 2014) of two different methods of bandwidth selection for the presmoothed estimators of the functions S, Λ, f and $\lambda = \Lambda'$. Correct bandwidth choice is also expected to increase the power of the presmoothed LR test over the classical one. A simple approach consists in using an estimation-based optimal bandwidths for the presmoothed LR test. Simulation evidence presented by Jácome and López-de-Ullibarri (2014) shows that the presmoothed LR test performs well with bandwidths selected in that way.

Nevertheless, the goal in testing problems is different to that in estimation, since interest is in constructing a powerful test rather than a good estimator. Consequently, estimation-based optimal bandwidths may be suboptimal in testing problems. An approach to bandwidth selection that tries to maximize power against a given alternative while at the same time preserving test size should be more suitable. Methods of bandwidth selection specifically tailored to testing problems have received relatively little attention in the literature. In the context of testing for the equality of regression curves, papers by Kulasekera and Wang (1997), Zhang (2003) and Gao and Gijbels (2008) deserve special mention. Using empirical likelihood methods, Cao and Van Keilegom (2006) proposed a test for the two-sample problem based on the comparison of two kernel density estimators. They showed that the bandwidths had a marked influence on power, and proposed to select them as those maximizing power in a grid of bandwidths. Martínez-Clambor and de Uña-Álvarez (2013b) compared different bandwidth selectors for a k-sample smooth test based on a generalized L_1 distance (Martínez-Clambor and de Uña-Álvarez 2009). In the next section, we show how the procedures of Martínez-Clambor and de Uña-Álvarez (2013b) can be adapted to the presmoothed LR test.

3. BANDWIDTH SELECTION

We consider two different approaches for selecting the bandwidths vector $\boldsymbol{b}$ of the presmoothed LR test $G^P_{\boldsymbol{b}}$. They differ with respect to the optimality measure used: the first approach concentrates on the maximization of the power under the alternative hypothesis, while the second considers the idea of minimizing the p-value. As Martínez-Clambor and de Uña-Álvarez (2013b) point out, the two approaches are strongly related.

In the algorithms below, the optimal bandwidths are chosen from the cartesian product $\mathcal{B}_1 \times \mathcal{B}_2$, where $\mathcal{B}_j = \{b_{1j}, b_{2j}, \ldots, b_{\ell_j j}\}$ is a grid of ℓ_j bandwidths used for sample j, $j = 1, 2$. Typically, the bandwidth is a function of sample size. Most of the theoretical results related with kernel estimates assume that, as sample size increases, the bandwidth must tend to zero, but not too fast. Specifically, in presmoothed estimation of the cumulative hazard function, the optimal bandwidth in terms of minimizing the asymptotic mean squared error is of order $O(n^{-1/3})$ (Cao et al. 2005). The optimal rate for testing purposes may be different to that of estimation and, in fact, might not exist for the presmoothed LR test. However, bandwidths of order $O(n^{-1/3})$, such as those given by the plug-in or bootstrap bandwidth selectors for presmoothed cumulative hazard estimation (see López-de-Ullibarri and Jácome 2013), besides being optimal for estimation, also guarantee that the classical and presmoothed LR tests have the same asymptotic normal distribution. In practice, for the jth sample we use a grid of bandwidths of the form $b_{ij} = d_i \widehat{b}^{PI}_j$, for $i = 1, 2, \ldots, \ell_j$, where the d_i's define an increasing numeric sequence ranging from $d_1 = 0$ to $d_{\ell_j} = 2$ and $\widehat{b}^{PI}_j$ is the plug-in bandwidth for the presmoothed estimate of the cumulative hazard function computed with the jth sample. In this way the selected bandwidth is of order $O(n_j^{-1/3})$.

3.1. Power maximization approach

An appealing idea, proposed by Kulasekera and Wang (1997) and subsequently developed by Cao and Van Keilegom (2006), Gao and Gijbels (2008) and Martínez-Camblor and de Uña-Álvarez (2013b), is to choose the bandwidth that maximizes test power.

Let $\boldsymbol{b}$ and $\mathcal{T}_{\boldsymbol{b}}$ be, respectively, a vector of smoothing parameters and a test statistic depending on $\boldsymbol{b}$. Also, let $T_{0,\boldsymbol{b}}$ and $T_{1,\boldsymbol{b}}$ be two independent random variables having the distribution of $\mathcal{T}_{\boldsymbol{b}}$ under the null and alternative hypotheses, respectively. Assuming, without loss of generality, that the null is rejected for large values of $\mathcal{T}_{\boldsymbol{b}}$, we will denote by $c_{\boldsymbol{b},\alpha}$ the left endpoint of the α-level rejection region, which verifies $P(T_{0,\boldsymbol{b}} \leq c_{\boldsymbol{b},\alpha}) \geq 1-\alpha$. The bandwidth that maximizes the power at a given α-level is:

$$\boldsymbol{b}_{M,\alpha} = \underset{\boldsymbol{b}}{\operatorname{argmax}}\ P(T_{1,\boldsymbol{b}} > c_{\boldsymbol{b},\alpha}).$$

The distribution of many test statistics that depend on a bandwidth is asymptotically normal, and this is also the case of the presmoothed LR test (Jácome and López-de-Ullibarri 2014). So, under the assumption that $T_{0,\boldsymbol{b}} \sim N(\mu_{0,\boldsymbol{b}}, \sigma_{0,\boldsymbol{b}})$ and $T_{1,\boldsymbol{b}} \sim N(\mu_{1,\boldsymbol{b}}, \sigma_{1,\boldsymbol{b}})$, we have $c_{\alpha,\boldsymbol{b}} = \mu_{0,\boldsymbol{b}} + \sigma_{0,\boldsymbol{b}} z_\alpha$, where z_α is the $100(1-\alpha)\%$ percentile of the standard normal, and

$$\begin{aligned}\boldsymbol{b}_{M,\alpha} &= \underset{\boldsymbol{b}}{\operatorname{argmax}}\ P(T_{1,\boldsymbol{b}} > c_{\alpha,\boldsymbol{b}}) = \underset{\boldsymbol{b}}{\operatorname{argmax}}\ P(N(\mu_{1,\boldsymbol{b}}, \sigma_{1,\boldsymbol{b}}) > \mu_{0,\boldsymbol{b}} + \sigma_{0,\boldsymbol{b}} z_\alpha) \\ &= \underset{\boldsymbol{b}}{\operatorname{argmax}} \left(-\frac{\mu_{0,\boldsymbol{b}} + \sigma_{0,\boldsymbol{b}} z_\alpha - \mu_{1,\boldsymbol{b}}}{\sigma_{1,\boldsymbol{b}}} \right).\end{aligned}$$

Since in practice the parameters of the distribution of the test under both the null and the alternative are unknown, they can be estimated using bootstrap methods by generating resamples from the pooled sample and separately from each of the two samples, respectively. The algorithm for computing $\boldsymbol{b}_{M,\alpha}$ adapts the 'Double Bootstrap' (DB) technique proposed by Cao and Van Keilegom (2006). In that paper, one of the two bootstrap procedures used for DB is required to approximate the distribution of the test statistic under the null and, specifically, the value of $c_{\alpha,\boldsymbol{b}}$. This is not needed for the presmoothed LR test since, from Theorem 2.1 in Jácome and López-de-Ullibarri (2014), for bandwidths of the form $b_j = a_j n_j^{-\beta_j} + o(n_j^{-\beta_j})$ with $1/4 < \beta_j < 1/2$, $\mu_{0,\boldsymbol{b}}$ is 0 and $\sigma^2_{0,\boldsymbol{b}}$ is σ^2_{LR}, the asymptotic variance of the LR test. The asymptotic distribution of the presmoothed LR test under both the null and the contiguous alternatives is the same as that of the classical LR test. Hence $c_{\alpha,\boldsymbol{b}}$ does not depend on $\boldsymbol{b}$, and can be estimated by $\widehat{\sigma}_{LR} z_\alpha$, with

$$\begin{aligned}\widehat{\sigma}^2_{LR} = \frac{n}{n_1 n_2} \int_0^\tau W^2(v) \frac{Y_1(v) Y_2(v)}{(Y_1(v) + Y_2(v))^2} \\ \times \left(1 - \frac{\triangle N_1(v) + \triangle N_2(v) - 1}{Y_1(v) + Y_2(v) - 1} \right) (dN_1(v) + dN_2(v)),\end{aligned} \tag{2}$$

where $\tau = \min(Z_{(n_1)1}, Z_{(n_2)2})$ and $Z_{(n_j)j}$ is the n_jth order statistic of the jth sample.

The steps of the DB algorithm for the presmoothed LR test are:

1. For sample j, $j = 1, 2$, define the grid of bandwidths $\mathcal{B}_j = \{b_{1j}, b_{2j}, \ldots, b_{\ell j}\}$, where b_{ij}, $i = 1, 2, \ldots, \ell$, is an equispaced, increasing numeric sequence such that $b_{1j} = 0$ and $b_{\ell j} = 2\widehat{b}_j^{PI}$, and $\widehat{b}_j^{PI}$ denotes the plug-in bandwidth for the presmoothed estimate of the cumulative hazard function computed with that sample.

2. For sample j, $j = 1, 2$, independently draw R bootstrap resamples $\{(Z^{*r}_{ij}, \delta^{*r}_{ij}), i = 1, 2, \ldots, n_j\}$, for $r = 1, 2, \ldots, R$, by resampling from $\{(Z_{ij}, \delta_{ij}), i = 1, \ldots, n_j\}$ with replacement.

3. For each $r = 1, 2, \ldots, R$, and for each bandwidths vector $\boldsymbol{b} \in \mathcal{B}_1 \times \mathcal{B}_2$:

   3.1. Compute the critical value $c^{*r}_\alpha = \widehat{\sigma}^{*r}_{LR} z_\alpha$, where $\widehat{\sigma}^{*r}_{LR}$ is the squared root of (2) computed with the rth resample.

   3.2. Compute the value of the test statistic $G^{P,*r}_{\boldsymbol{b}}$ based on the rth resample and $\boldsymbol{b}$.

4. For each $\boldsymbol{b} \in \mathcal{B}_1 \times \mathcal{B}_2$, calculate the power by:

$$\widehat{pow}(\boldsymbol{b}) = \frac{1}{R}\sum_{r=1}^{R} \mathbf{1}(G_{\boldsymbol{b}}^{P,*r} > c_\alpha^{*r}).$$

5. Estimate $\boldsymbol{b}_{M,\alpha}$ by

$$\widehat{\boldsymbol{b}}_{M,\alpha}^{DB} = \underset{\boldsymbol{b}\in\mathcal{B}_1\times\mathcal{B}_2}{\operatorname{argmax}}\ \widehat{pow}(\boldsymbol{b}).$$

3.2. p-value minimization approach

This approach is based on the idea that more evidence against the null hypothesis is obtained for smaller p-values. Hence, in order to construct a powerful test it makes sense to minimize the function:

$$\begin{aligned}\pi(\boldsymbol{b}) &= P\{\mathcal{T}_{0,\boldsymbol{b}} > \mathcal{T}_{1,\boldsymbol{b}}\} = E[\mathbf{1}\{\mathcal{T}_{0,\boldsymbol{b}} > \mathcal{T}_{1,\boldsymbol{b}}\}] = E[E[\mathbf{1}\{\mathcal{T}_{0,\boldsymbol{b}} > \mathcal{T}_{1,\boldsymbol{b}}|\mathcal{T}_{1,\boldsymbol{b}}\}]] \\ &= E[\pi_0(\mathcal{T}_{1,\boldsymbol{b}})],\end{aligned}$$

where $\pi_0(t)$ is the p-value of t, that is, the probability under the null hypothesis that the value of the test statistic is larger than t. The minimum p-value bandwidth $\boldsymbol{b}_m$ is given by

$$\boldsymbol{b}_m = \underset{\boldsymbol{b}}{\operatorname{argmin}}\ \pi(\boldsymbol{b}) = \underset{\boldsymbol{b}}{\operatorname{argmin}}\ E[\pi_0(\mathcal{T}_{1,\boldsymbol{b}})].$$

We can interpret $\boldsymbol{b}_m$ as the bandwidth that achieves maximal separation between the null and alternative hypotheses. Unlike $\boldsymbol{b}_{M,\alpha}$, $\boldsymbol{b}_m$ is not influenced by the particular significance level at which the test is performed.

Martínez-Camblor et al. (2008) introduced an algorithm implementing the p-value minimization approach in the context of nonparametric k-sample tests (see also Martínez-Camblor and de Uña-Álvarez 2009, 2013b). Although their algorithm was called 'Double Minimum' (DM) because it consists of two minimization steps, only one minimization is required when applied to the presmoothed LR test. The DM algorithm for the presmoothed LR test can be structured in five steps. We only give the last three steps, since the first two are the same as in the DB algorithm above:

3. For each $r = 1, 2, \ldots, R$, and for each bandwidths vector $\boldsymbol{b} \in \mathcal{B}_1 \times \mathcal{B}_2$:

   3.1. Compute $\widehat{\sigma}_{LR}^{*r}$.

   3.2. Compute $G_{\boldsymbol{b}}^{P,*r}$ the test statistic based on the rth resample and $\boldsymbol{b}$.

   3.3. Compute $\widehat{\pi}^{*r}(\boldsymbol{b}) = 1 - \Phi(|G_{\boldsymbol{b}}^{P,*r}|/\widehat{\sigma}_{LR}^{*r})$.

4. For each $\boldsymbol{b} \in \mathcal{B}_1 \times \mathcal{B}_2$, approximate the p-value by

$$\widehat{\pi}(\boldsymbol{b}) = \frac{1}{R}\sum_{r=1}^{R} \widehat{\pi}^{*r}(\boldsymbol{b}).$$

5. Estimate $\boldsymbol{b}_m$ by

$$\widehat{\boldsymbol{b}}_m^{DM} = \underset{\boldsymbol{b}\in\mathcal{B}_1\times\mathcal{B}_2}{\operatorname{argmin}}\ \widehat{\pi}(\boldsymbol{b}).$$

A modified minimum p-value criterion, based on the idea of minimizing the p-values of the test statistic that would be obtained under the alternative, and then averaging with respect to the distribution of $T_{1,\boldsymbol{b}}$, should produce a more powerful test. This criterion leads to the bandwidth

$$\boldsymbol{b}_{\overline{m}} = E[\underset{\boldsymbol{b}>\mathbf{0}}{\operatorname{argmin}}\ \pi(T_{1,\boldsymbol{b}})].$$

This idea of averaging along the alternative has been previously used in density comparison for paired samples (Martínez-Camblor 2010) and later by Martínez-Camblor and de Uña-Álvarez (2013a,b).

The 'Bootstrap Minimum' (BM) algorithm of Martínez-Camblor et al. (2008) is a modification of the DM algorithm which can be adapted to the estimation of $\boldsymbol{b}_{\overline{m}}$ in the presmoothed LR test. The last two steps of the BM method are the following (the first three steps are the same as in the DM algorithm):

4. For each $r = 1, 2, \ldots, R$, choose the bandwidths vector $\boldsymbol{b} \in \mathcal{B}_1 \times \mathcal{B}_2$ that minimizes the corresponding sequence of p-values computed in step 3:

$$\widehat{\boldsymbol{b}}^{*r} = \underset{\boldsymbol{b}\in\mathcal{B}_1\times\mathcal{B}_2}{\operatorname{argmin}}\ \widehat{\pi}^{*r}(\boldsymbol{b}).$$

5. Estimate $\boldsymbol{b}_{\overline{m}}$ by

$$\widehat{\boldsymbol{b}}_{\overline{m}}^{BM} = \frac{1}{R}\sum_{r=1}^{R}\widehat{\boldsymbol{b}}^{*r}.$$

In the context of k-sample tests based on kernel density estimator, Martínez-Clambor (2010) and Martínez-Clambor and de Uña-Álvarez (2013b) showed that the DM method was more conservative and less powerful than the BM method. These authors point out that the better trade-off between level and power that the BM algorithm achieves by averaging along the alternative distribution can be explained by regarding this algorithm as a bagged (or bootstrap aggregated; see Breiman 1996) version of the DM algorithm. Since, as expected, the BM method also performs better than the DM method for the presmoothed LR test, only the results for the first method are shown in the next sections.

4. SIMULATION STUDY

We have studied the empirical behavior of the presmoothed LR test with the proposed bandwidth selectors, comparing it with that of the classical LR test. Although, for simplicity, we have considered only unweighted LR tests, similar results should be expected for their weighted versions.

Two different scenarios with a variety of alternatives and percentages of censoring were simulated. For facilitating comparison, the simulated configurations were the same as in Jácome and López-de-Ullibarri (2014). Specifically, lifetimes T_j, $j = 1, 2$, were generated from two different distributions:

(a) Weibull with density

$$f_j(t) = \frac{1}{\sigma_j}\exp\left(-\frac{\mu_j}{\sigma_j}\right)t^{\frac{1}{\sigma_j}-1}\exp\left[-\exp\left(-\frac{\mu_j}{\sigma_j}\right)t^{\frac{1}{\sigma_j}}\right],\ \text{for } t>0.$$

(b) Lognormal with density

$$f_j(t) = \frac{1}{\sqrt{2\pi}\sigma_j t}\exp\left[-\frac{(\log t-\mu_j)^2}{2\sigma_j^2}\right],\ \text{for } t>0.$$

The scale parameter for T_1 was always $\mu_1 = 0$. The departure from the null hypothesis was controlled by μ_2, which is set to 0 for the null hypothesis and to 0.75 for the alternative. The degree of skewness and tail weight of the distributions were controlled by the common shape parameters $\sigma_1 = \sigma_2 \equiv \sigma > 0$. While values of σ smaller than 1 yield lightly skewed, narrow and tall distributions, values greater than 1 produce heavily skewed, broad and shallow ones. For light skewness we chose $\sigma = 0.75$, and for heavy skewness $\sigma = 2$ (Weibull) and $\sigma = 1.5$ (lognormal).

Censoring times were drawn from uniform distributions $U(0, c)$. Different censoring rates were achieved by setting $c = 4$ for light censoring and $c = 1$ for heavy censoring.

The power of the presmoothed and classical LR tests was assessed with two-sided tests at $\alpha = 0.05$ significance level, for different sample sizes. The tests were evaluated at $\tau = \min(Z_{(n_1)1}, Z_{(n_2)2})$,

where $Z_{(n_j)j}$ is the n_jth ordered observation of sample j. This guarantees that the interval in which the test is computed does not surpass the interval where the observations are confined. We took $R = 200$ bootstrap resamples in the algorithms for bandwidth selection, and the proportion of rejections of the null was computed from 1000 Monte Carlo replications. The grid for bandwidth search was $\mathcal{B}_1 \times \mathcal{B}_2$ where $\mathcal{B}_j = \{b_{1j}, b_{2j}, \ldots, b_{5j}\}$ defines an equispaced increasing numeric sequence from $b_{1j} = 0$ to $b_{5j} = 2\widehat{b}_j^{PI}$, and $\widehat{b}_j^{PI}$ denotes the plug-in bandwidth for the presmoothed estimate of the cumulative hazard function. If $\widehat{b}_j^{PI} = 0$, then $b_{5j} = Z_{(n_j)j} - Z_{(1)j}$. In a complementary simulation study, taking $b_{5j} = 5\widehat{b}_j^{PI}$ did not result in significant differences (results not shown).

Table 1 shows the proportion of rejections of the true null hypothesis for the presmoothed LR test with DB and BM bandwidths in the simulated scenarios. Also shown in Table 1 are the results for the classical LR test. The DB method tends to give an anticonservative test (see Cao and Van Keilegom 2006; Martínez-Camblor and de Uña-Álvarez 2013b). Although this seems to be also the case for the presmoothed LR test, both methods of bandwidth selection give very similar results in terms of test size, which is maintained very close to the nominal 0.05 level.

			Weibull			Lognormal		
Skew	Cens	(n_1, n_2)	LR	DB	BM	LR	DB	BM
Light	Light	(20,20)	0.052	0.057	0.057	0.057	0.061	0.060
		(50,50)	0.050	0.051	0.048	0.050	0.058	0.053
		(25,50)	0.063	0.066	0.064	0.036	0.044	0.039
		(100,100)	0.056	0.065	0.062	0.046	0.064	0.058
	Heavy	(20,20)	0.055	0.056	0.054	0.050	0.058	0.058
		(50,50)	0.037	0.048	0.048	0.040	0.061	0.058
		(25,50)	0.047	0.066	0.064	0.063	0.083	0.081
		(100,100)	0.048	0.063	0.061	0.047	0.077	0.076
Heavy	Light	(20,20)	0.054	0.059	0.059	0.049	0.062	0.058
		(50,50)	0.056	0.057	0.059	0.047	0.059	0.056
		(25,50)	0.061	0.064	0.064	0.041	0.057	0.051
		(100,100)	0.062	0.065	0.066	0.043	0.052	0.052
	Heavy	(20,20)	0.048	0.058	0.054	0.046	0.049	0.049
		(50,50)	0.045	0.052	0.049	0.048	0.055	0.054
		(25,50)	0.049	0.059	0.053	0.053	0.072	0.064
		(100,100)	0.053	0.061	0.056	0.049	0.065	0.049

Táboa 1: Simulation results. Proportion of rejections of the true H_0 at a nominal 0.05 level for the LR and presmoothed LR tests with DB and BM bandwidths. The censoring percentage (Cens) is the same for the two samples.

For all the simulated scenarios, Figure 1 shows the power of the presmoothed LR test with DB and BM bandwidths as a function of sample size, and compares it with the classical LR test. Although little difference is perceived between both algorithms of bandwidth selection, the power is always slightly greater with DB than with BM bandwidths. On the other hand, it is remarkable that the presmoothed LR test always outperforms the classical LR test, which is clearly the less powerful in all the scenarios, mainly in the presence of heavy skewness and/or heavy censoring. At the same time, the size of the presmoothed LR test remains very close to the nominal level independently of the bandwidth selector. While the bandwidths seem to have a negligible effect on the probability of rejecting a true null hypothesis, they are important in order to make effective the increased ability of the presmoothed test to detect false null hypotheses. The usefulness of presmoothing in combination with the proper choice of the bandwidths is clear.

5. APPLICATION TO A REAL DATA SET

In order to illustrate the use of the proposed methods of bandwidth selection in practical settings, we have applied them to two different clinical datasets. These datasets were already analyzed in Jácome and López-de-Ullibarri (2014).

Example 1. *Kidney data.* The goal of the study of the kidney dataset (Klein and Moeschberger 2003), was to determine if, in patients undergoing kidney dialysis, the time until the first catheter exit-site infection was different depending on whether the catheter was surgically or percutaneously placed. The data consist of the times (in months) until first exit-site infection of 119 patients, a variable coding if the observed times are or not censored, and the group number: 1 for a percutaneously placed catheter (76 patients and 85.52% of censoring) and 2 for surgically placed catheter (43 patients and 65.11% of censoring). The interest is in comparing the survival curves of the two groups.

At the 0.05 significance level, the classical LR test does not detect any differences between both curves: the value of the LR statistic is 1.590, resulting in a p-value of 0.112. As for the presmoothed LR test, the plug-in bandwidths are computed to be $(\widehat{b}_1^{PI}, \widehat{b}_2^{PI}) = (13.000, 2.027)$. We then take the rectangular grid of $20 \times 20 = 400$ bandwidths $\mathcal{B}_1 \times \mathcal{B}_2 = \{0, 1.368, \ldots, 26.000\} \times \{0, 0.213, \ldots, 4.054\}$. The selected bandwidths are $\widehat{\boldsymbol{b}}_{M,\alpha}^{DB} = (19.158, 3.414)$ and $\widehat{\boldsymbol{b}}_{\overline{m}}^{BM} = (10.030, 1.721)$, leading to non-significant p-values of 0.135 and 0.154, respectively. Thus, none of the LR-type tests find significant differences between groups. In fact, as Figure 2 shows, the presmoothed LR test does not reject the null hypothesis with any of the bandwidths of the grid, and it is also noticeable that the p-values are always greater than with the classical LR test.

Example 2. *PBC data.* Between 1974 and 1984, the Mayo Clinic conducted a double-blinded randomized trial in primary biliary cirrhosis of the liver (PBC), comparing the drug D-penicillamine with a placebo. The lifetime is the survival time (in years), and right censoring is caused by the end of follow-up or by liver transplantation. PBC is more common in women than men, and men appear to develop hepatocellular carcinoma later than women. Here we focus on the effect of sex on the survival rate of PBC.

In order to show the performance of the presmoothed LR test in a situation with a borderline p-value, the study is restricted to the 292 patients without edema, a subset in which the differences in survival by sex are not conclusive according to the classical LR test. There are 33 males and 259 females, with censoring percentages of 42.4% and 66.4%, respectively.

The classical LR test leads to a marginally significant p-value of 0.045. Figure 2 shows that, for a wide range of bandwidths, the presmoothed LR test seems to be more powerful in this case. This suggests that the null should be rejected at the 0.05 level, concluding in consequence that male and female patients have different survival distributions.

In this example the plug-in bandwidths are $(\widehat{b}_1^{PI}, \widehat{b}_2^{PI}) = (5.378, 2.988)$, and the rectangular grid $\mathcal{B}_1 \times \mathcal{B}_2$ consists of the $20 \times 20 = 400$ bandwidths $\{0, 0.566, \ldots, 10.756\} \times \{0, 0.315, \ldots, 5.976\}$. The selected bandwidths are $\widehat{\boldsymbol{b}}_{M,\alpha}^{DB} = (1.132, 4.428)$ and $\widehat{\boldsymbol{b}}_{\overline{m}}^{BM} = (4.554, 4.754)$, corresponding to p-values of 0.023 and 0.028, respectively. This example shows that in cases of borderline p-values of the classical LR test, the presmoothed LR test may succeed in taking a decision more unambiguously.

REFERENCIAS

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Cao R, López-de-Ullibarri I, Janssen P, Veraverbeke N (2005) Presmoothed Kaplan-Meier and Nelson-Aalen estimators. J Nonparametr Stat 17:31–56

Cao R, Jácome MA (2004) Presmoothed kernel density estimator for censored data. J Nonparametr Stat 16:289–309

Cao R, Van Keilegom I (2006) Empirical likelihood tests for two-sample problems via nonparametric density estimation. Canad J Stat 34:61–77

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York

Gao J, Gijbels I (2008) Bandwidth selection in nonparametric kernel testing. J Am Stat Assoc 484:1584–1594

Gill R (1980) Censoring and stochastic integrals. Mathematical Centre Tracts, 124. Mathematisch Centrum, Amsterdam

Jácome MA, Cao R (2007) Almost sure asymptotic representation for the presmoothed distribution and density estimators for censored data. Statistics 41:517–534

Jácome MA, López-de-Ullibarri I (2014) Logrank-type tests with presmoothing. Biom J 56:851–866

Klein JP, Moeschberger ML (2003) Survival analysis: techniques for censored and truncated data, 2nd edn. Springer, New York

Kulasekera KB, Wang J (1997) Smoothing parameter selection for power optimality in testing of regression curves. J Am Stat Assoc 438:500–511

López-de-Ullibarri I, Jácome MA (2013) survPresmooth: An R package for presmoothed estimation in survival analysis. J Stat Softw 54(11):1–26

Martínez-Camblor P (2010) Nonparametric k-sample tests based on kernel density estimator for paired design. Comput Stat Data Anal 54:2035–2045

Martínez-Camblor P, de Uña-Álvarez J, Corral N (2008) k-sample test based on the common area of kernel density estimator. J Stat Plan Inference 138:4006–4020

Martínez-Camblor P, de Uña-Álvarez J (2009) Nonparametric k-sample tests: density function vs. distribution function. Comput Stat Data Anal 53:3344–3357

Martínez-Camblor P, de Uña-Álvarez J (2013a) Density comparison for independent and right censored data via kernel density estimator. Comput Stat 28:269–288

Martínez-Camblor P, de Uña-Álvarez J (2013b) Studying the bandwidth in k-sample smooth tests. Comput Stat 28:875–892

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`

Zhang C (2003) Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. J Am Stat Assoc 98:609–629

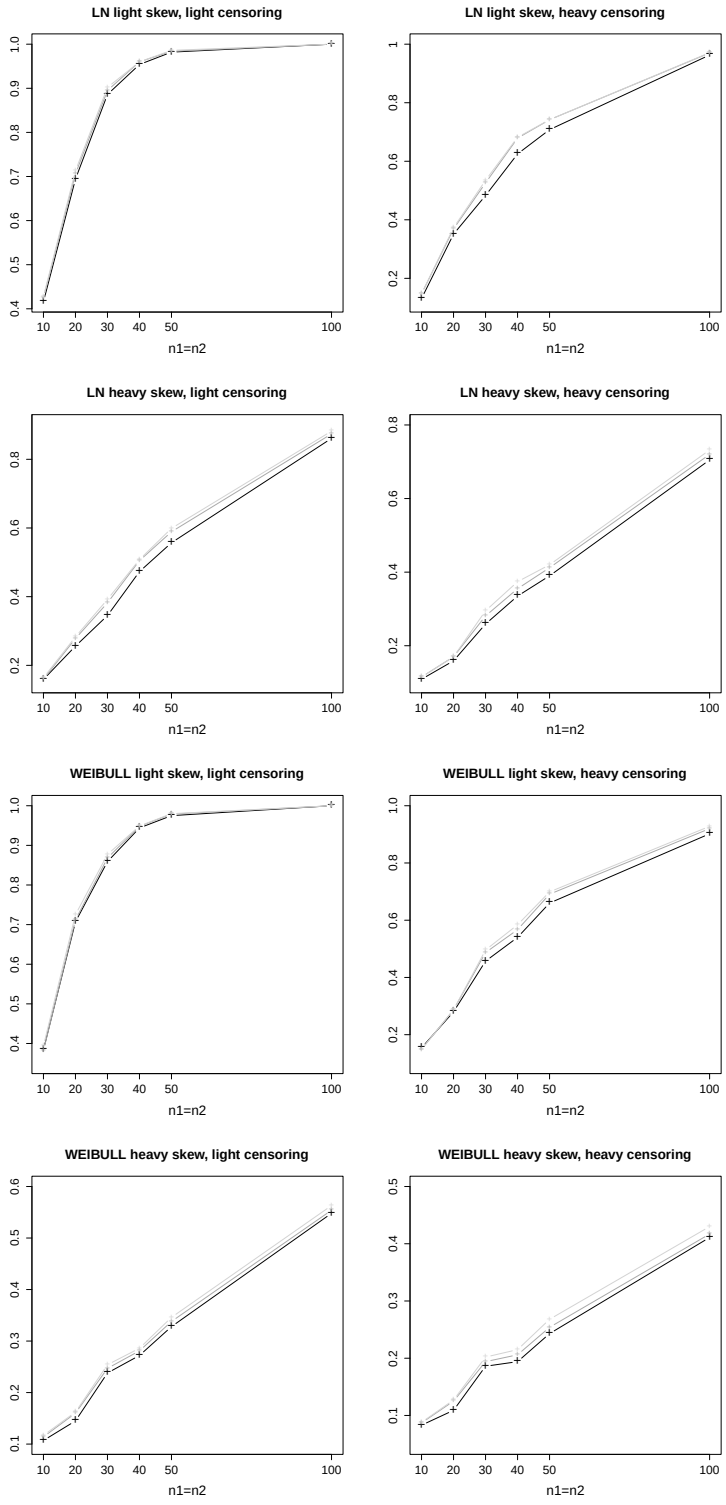


Figura 1: Proportion of rejections of H_0 under the alternatives considered for the classical (black line) and presmoothed LR tests with BM (dark grey line) and DB (light grey line) bandwidths.

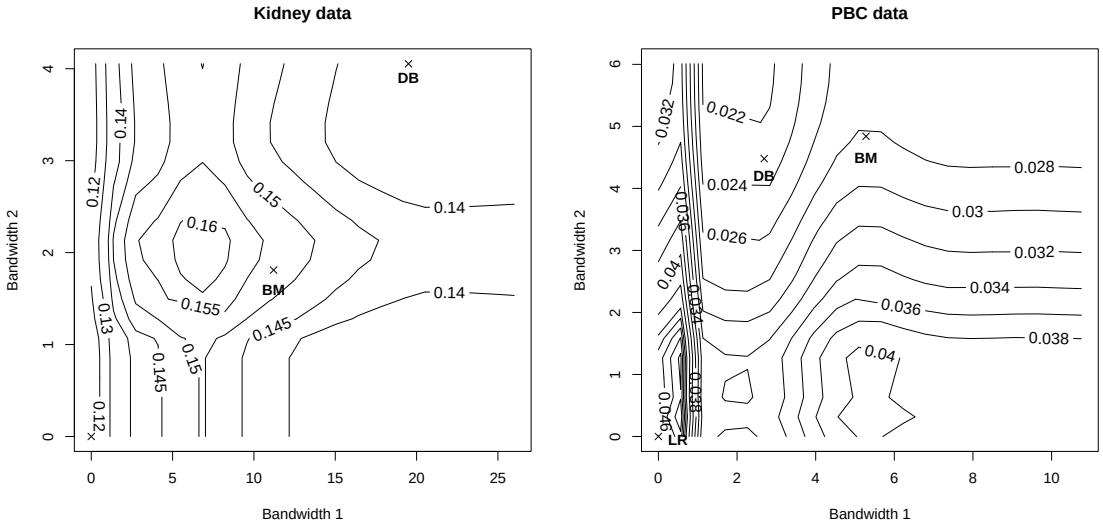


Figura 2: p-values of the presmoothed LR test in a grid of bandwidths for the kidney data (left) and PBC data (right). The p-values of the premoothed LR test with the DB and BM bandwidths, as well as that of the LR test (at the point $(0, 0)$) are marked with crosses.

Comunicacións orais

Sesión de Medias

SIMULACIÓN

Miguel Ángel Rodríguez-Gigirey Pérez[1] .

[1] Profesor de matemáticas no I.E.S. de Arzúa

RESUMO

Un dos temas non abarcados no currículo de Matemáticas nin de Métodos estatísticos e numéricos durante a etapa do bacharelato é a ***simulación***. Hai anos fixen a proba de introducilo na materia optativa citada con gran éxito de resultados, e puiden comprobar axuda á comprensión doutros temas por parte do meu alumnado.

Palabras e frases chave: *simulación, folla de cálculo, números aleatorios.*

1. INTRODUCIÓN

Mediante o uso de follas de cálculo, os rapaces e rapazas de 2º de bacharelato simulan lanzamentos de dados e moedas, a extracción de cartas, e outros fenómenos a través da xeración de gran cantidade de números aleatorios que transforman en ditos fenómenos. Posteriormente recontan resultados e realizan comparativas dos datos obtidos coas probabilidades teóricas, mediante parámetros estatísticos e gráficos.

Este tema, que non se atopa no currículo, esperta grande interese entre o alumnado e motiva a algúns a xerar novas situacións para aproximar probabilidades descoñecidas de fenómenos do seu entorno. Tamén se fan exemplos de selección de mostras aleatorias en temas posteriores.

2. PRIMEIROS PASOS

Comézase a unidade amosando as características da función ALEATORIO( ) das follas de cálculo, amosando como manipulala para obter valores de dados, de moedas, de cartas, de quinielas ou de números da primitiva.

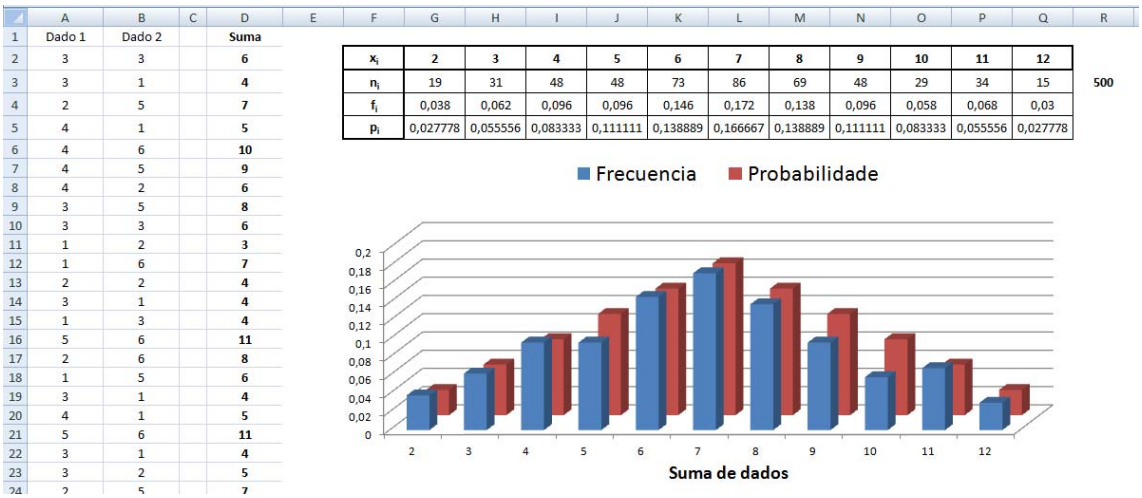
	A	B	C	D
1	**Obxectivo**	**Num. aleatorio**	**Comando**	**Resultado (Ex.)**
2	**Dado**		ENTERO(6*ALEATORIO()+1)	5
3	**Moeda**		SI(ALEATORIO()<0,5;"CARA";"CRUZ")	CARA
4	**Quiniela**	ALEATORIO( )	SI(B4<0,5;1;SI(B4<5/6;"X";2))	X
5	**Carta**	ALEATORIO( )	SI(B5=1;"AS";SI(B5=8;"SOTA";SI(B5=9; "CABALO";SI(B5=10;"REI";B5))))	7

Táboa 1: Comandos para xerar resultados básicos

Posteriormente trabállase con exercicios nos que se repiten 200, 500 ou 1000 veces cada unha destas extraccións ou lanzamentos, ás veces en series de 2, 3 ou 5 de cada vez e procédese ao reconto dos resultados mediante os comandos condicionais e a función CONTAR.SI (ou similar, segundo o software de folla de cálculo que se utilice)

Con estes datos realízanse táboas de frecuencias absolutas e relativas, tratando de ver con estas últimas como se aproximan aos valores das probabilidades teóricas.

Por exemplo, un dos exercicios que se realiza na aula consiste en simular o lanzamento de 500 parellas de dados e realizar a suma de ambos. En temas de probabilidade anteriores analizáronse as probabilidades das distintas sumas {2, 3, 4, ..., 11, 12}, e demostrouse así que a suma 7 é a máis probable, seguidas de 6 e 8, 5 e 9, etc., ata ver tamén que as sumas 2 e 12 son as máis improbables. Con esta experiencia, o alumnado comproba a "realidade" dunha realización simulada da proba e, cun gráfico pode comparar resultados teóricos e da experiencia.



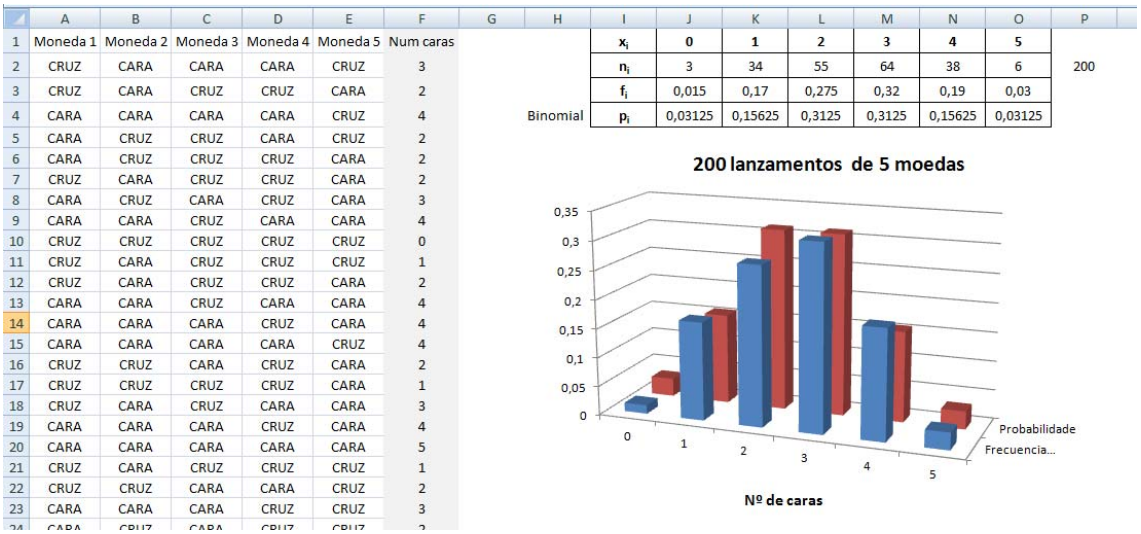
Dado 1	Dado 2	Suma
3	3	6
3	1	4
2	5	7
4	1	5
4	6	10
4	5	9
4	2	6
3	5	8
3	3	6
1	2	3
1	6	7
2	2	4
3	1	4
1	3	4
5	6	11
2	6	8
1	5	6
3	1	4
4	1	5
5	6	11
3	1	4
3	2	5

x_i	2	3	4	5	6	7	8	9	10	11	12	
n_i	19	31	48	48	73	86	69	48	29	34	15	500
f_i	0,038	0,062	0,096	0,096	0,146	0,172	0,138	0,096	0,058	0,068	0,03	
p_i	0,027778	0,055556	0,083333	0,111111	0,138889	0,166667	0,138889	0,111111	0,083333	0,055556	0,027778	

Figura 1. Posible resultado do experimento *lanzamento de dous dados e suma de ambos*.

3. DESENVOLVEMENTO DE PROBAS

Son moitos os experimentos que se simulan na aula. Cabe destacar outro no que se xeran 5 moedas (poden ser máis ou menos) e repítese a proba 200 veces. O reconto de *caras* permite construír unha táboa de frecuencias que visualiza a realización efectiva dunha variable aleatoria discreta, na que posteriormente observan que sigue unha distribución binomial. Neste reconto, as frecuencias relativas aproxímanse á función masa de probabilidade de dita distribución binomial, vista en temas anteriores e que sorprende ao alumnado.



Moneda 1	Moneda 2	Moneda 3	Moneda 4	Moneda 5	Num caras
CRUZ	CARA	CARA	CARA	CRUZ	3
CRUZ	CARA	CRUZ	CRUZ	CARA	2
CARA	CARA	CARA	CARA	CRUZ	4
CARA	CRUZ	CRUZ	CARA	CRUZ	2
CARA	CRUZ	CRUZ	CRUZ	CARA	2
CRUZ	CARA	CRUZ	CRUZ	CARA	2
CARA	CRUZ	CARA	CRUZ	CARA	3
CARA	CARA	CRUZ	CARA	CARA	4
CRUZ	CRUZ	CRUZ	CRUZ	CRUZ	0
CRUZ	CARA	CRUZ	CRUZ	CRUZ	1
CRUZ	CARA	CRUZ	CRUZ	CARA	2
CARA	CARA	CRUZ	CARA	CARA	4
CARA	CARA	CARA	CRUZ	CARA	4
CARA	CARA	CARA	CARA	CRUZ	4
CRUZ	CRUZ	CARA	CRUZ	CARA	2
CRUZ	CRUZ	CRUZ	CRUZ	CARA	1
CRUZ	CARA	CRUZ	CARA	CARA	3
CARA	CARA	CARA	CRUZ	CARA	4
CARA	CARA	CARA	CARA	CARA	5
CRUZ	CARA	CRUZ	CRUZ	CRUZ	1
CRUZ	CRUZ	CARA	CARA	CRUZ	2
CARA	CARA	CARA	CRUZ	CRUZ	3

	x_i	0	1	2	3	4	5	
	n_i	3	34	55	64	38	6	200
	f_i	0,015	0,17	0,275	0,32	0,19	0,03	
Binomial	p_i	0,03125	0,15625	0,3125	0,3125	0,15625	0,03125	

Figura 2. Posible resultado do experimento *lanzamento de 5 moedas e reconto de caras.*

Cando se rematan estas actividades na aula, pídese habitualmente que se cambie o número de probas por 1000 en vez de 200 para observar como a aproximación á distribución binomial é maior.

4. SIMULACIÓN AVANZADA

Unha vez que se considera que o alumnado ten suficiente soltura cos comandos e funcións da folla de cálculo e se realizaron probas que simulen situacións variadas, proponse un último exercicio que consiste nun xogo de simulación avanzada:

Simular un xogo no que en cada paso unha ficha pode avanzar/retroceder un máximo de 2 casillas, podendo tamén permanecer na mesma. Estudar o avance en 100 movementos nos seguintes casos:

a) A probabilidade de avanzar ou retroceder en cada caso é a mesma.

b) A probabilidade dos movementos ven dada pola función masa de probabilidade seguinte:

x_i	-2	-1	0	1	2
p_i	1/8	1/8	1/6	1/4	1/3

Realizar dous gráficos: un para estudar a posición da ficha en cada un dos 100 movementos e outro para analizar se o reconto de veces que se dá con cada un dos avances se aproxima á probabilidade teórica

Nesta simulación trabállase con case tódolos conceptos aprendidos na unidade e como novidade introdúcese un modelo diferente aos vistos de dados, cartas, moedas, …. Para que o alumnado vexa que é posible utilizar a simulación para experimentos variados.

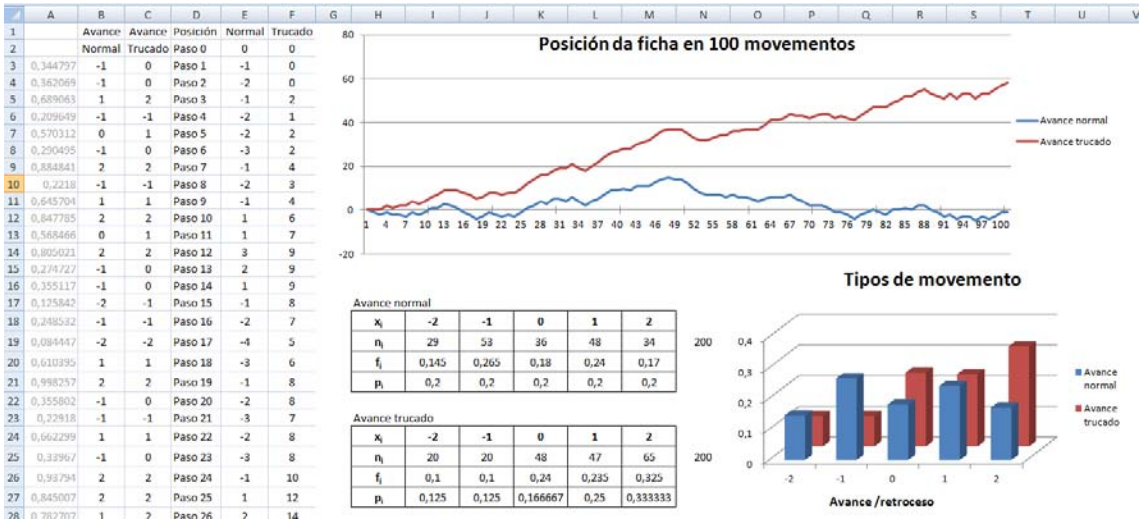


Figura 3. Posible resultado do experimento *avance da ficha no xogo.*

Tamén resulta moi práctico utilizar posteriormente esta ferramenta para elixir mostras aleatorias con poboacións coñecidas polo alumnado. Un exemplo pode ser seleccionar unha mostra estratificada de 50 alumnas e alumnos por curso no centro no que exerzo docencia. Unha vez explicados os conceptos clave de mostraxe, o alumnado sabe elixir de modo efectivo dita mostra para que sexa totalmente aleatoria.

5. CONCLUSIÓNS

Durante 5 cursos puiden constatar que na materia de Métodos estatísticos e numéricos, de 2º de bacharelato, é posible introducir unha unidade de simulación, a través de folla de cálculo que resulta moi útil para afianzar os conceptos de estatística descritiva, probabilidade, distribucións, mostraxe e incluso procesos estocásticos. Os profesores e profesoras de matemáticas tendemos moitas veces a dicir na aula que se poden simular con ordenador estas probas, pero rara vez se amosa cómo facelo ou se visualiza con exemplos. Damos a entender que consiste nun proceso difícil, cando realmente eles poden facelo cunhas poucas explicacións e observar facilmente os resultados.

O desenvolvemento desta unidade precisa de entre 6 e 8 sesións de clase, con ordenador para cada alumn@.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

LOMCE, hacia un cambio de estrategias en el aula

María Jesús Casado Barrio[1]

[1] Catedrática matemáticas de enseñanza secundaria

RESUMEN

La ciudadanía española está inmersa, una vez más, en un cambio de legislación educativo, el séptimo desde 1970. Cada uno de ellos introduce nuevas metodologías de aprendizaje vinculadas a las tendencias del momento y que no siempre llegan a ponerse en práctica. En este artículo se analizan algunas novedades de la LOMCE y su aportación a la mejora de la competencia estadística de los futuros ciudadanos.

Palabras y frases clave: Estándares de aprendizaje, pruebas externas, metodologías activas, Investigación estadística.

1. INTRODUCCIÓN

Los conocimientos estadísticos de la población y el manejo que puedan hacer de ellos están ligados, en buena parte, al aprendizaje recibido durante la etapa escolar. En nuestro país se está llevando a cabo un cambio legislativo que está fundamentado en un cambio de paradigma educativo, el **alumno como protagonista de su propio aprendizaje.** Se pretende que el alumno esté preparado para ser un ciudadano del siglo XXI, lo que implica convivir con una sociedad cambiante, muy tecnológica y con un marco social multicultural e inestable. **¿Qué partido pueden sacar nuestros aprendices de su bagaje estadístico? ¿cuáles son los conocimientos mínimos necesarios y el enfoque que deben darles?**

La LOMCE, Ley Orgánica para la Mejora de la Calidad Educativa, especifica seis elementos básicos en el currículo: Objetivos, contenidos, criterios de evaluación, **estándares de aprendizaje** (novedad), **competencias clave** (siete en total) y **metodología didáctica**.

2. ESTÁNDARES DE APRENDIZAJE

Los estándares de aprendizaje son especificaciones de los criterios de evaluación (elemento central del currículo), son referentes de la planificación curricular y la programación docente[i].

Su implantación en España resulta novedosa pero en otros lugares del mundo ya se hizo hace años con diferentes nomenclaturas. En Inglaterra, por ejemplo se llaman "Metas de Logro" y se les asignan niveles dentro de cada asignatura (adecuado, elemental, insuficiente). Su finalidad es definir de forma consensuada qué metas de aprendizaje deben adquirir todos los estudiantes, y forman parte de las pruebas nacionales e internacionales.

En 1996 la UNESCO[ii] encargó un informe a una comisión presidida por el prestigioso educador Jacques Delors. Por aquel entonces se planteaban cuatro aprendizajes fundamentales: **aprender a conocer**, **aprender a vivir juntos**, **aprender a hacer** y **aprender a ser**. La escuela de hoy, todavía sigue primando el primer aprendizaje: **conocer**. Es necesario superar este anclaje en la transmisión de conocimientos, de ahí que se promueva un cambio de metodología que se basa en un enfoque competencial, se busca la adquisición de competencias clave ligadas a **estándares de aprendizaje comunes a zonas del mundo cada vez más amplias**.

El currículo de matemáticas se estructura en una serie de bloques temáticos cuyo tratamiento en el aula se debe realizar de forma conjunta. El primer bloque, "Procesos, métodos y actitudes en matemáticas" es una declaración de intenciones sobre las herramientas básicas de pensamiento que el alumnado debe utilizar durante su aprendizaje. Los estándares cobran importancia vital en este bloque porque enlazan procesos cognitivos (semióticos y semánticos), con otros de tipo tecnológico, actitudinal o de experiencia vital. Estos se suman a los de los demás bloques para formar un currículo coherente con los cuatro aprendizajes.

Además, el bloque de estadística y probabilidad, especifica aspectos concretos del tratamiento de datos y la medida de la incertidumbre. Como muestra enumeramos los de la enseñanza primaria, no están

especificados por curso porque esta tarea fue delegada a las comunidades autónomas, así como la de asignarles a cada uno de ellos las competencias clave[iii].

1.1. Identifica datos cualitativos y cuantitativos en situaciones familiares.

2.1. Recoge y clasifica datos cualitativos y cuantitativos, de situaciones de su entorno, utilizándolos para construir tablas de frecuencias absolutas y relativas.

2.2. Aplica de forma intuitiva a situaciones familiares, las medidas de centralización: la media aritmética, la moda y el rango.

2.3. Realiza e interpreta gráficos muy sencillos: diagramas de barras, poligonales y sectoriales, con datos obtenidos de situaciones muy cercanas.

3.1. Realiza análisis crítico argumentado sobre las informaciones que se presentan mediante gráficos estadísticos.

4.1. Identifica situaciones de carácter aleatorio.

4.2. Realiza conjeturas y estimaciones sobre algunos juegos (monedas, dados, cartas, lotería…).

5.1. Resuelve problemas que impliquen dominio de los contenidos propios de estadística y probabilidad, utilizando estrategias heurísticas, de razonamiento (clasificación, reconocimiento de las relaciones, uso de contraejemplos), creando conjeturas, construyendo, argumentando, y tomando decisiones, valorando las consecuencias de las mismas y la conveniencia de su utilización.

5.2. Reflexiona sobre el proceso de resolución de problemas: revisando las operaciones utilizadas, las unidades de los resultados, comprobando e interpretando las soluciones en el contexto, proponiendo otras formas de resolverlo.

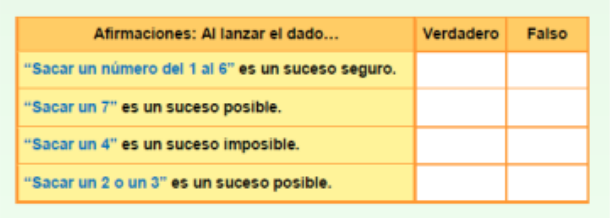
Una lectura reflexiva nos lleva a valorar el peso que en ellos tienen los procesos identificados dentro del marco de la resolución de problemas: representar, interpretar, argumentar ... y el bajo peso que se le dedica a la parte de cálculo, sobrevalorada en la metodología tradicional.

3. PRUEBAS EXTERNAS

Como se indicó anteriormente, los estándares son referentes de la planificación curricular, y de las **pruebas de evaluación externas**, otro de los puntos novedosos y controvertidos de la LOMCE. En primaria no tendrán efectos académicos pero en secundaria y bachillerato sí ya que su superación va ligada a la obtención del título correspondiente. El hecho de que haya una prueba externa condiciona a buena parte del profesorado a la hora de diseñar la programación y de priorizar contenidos. Una muestra es el receso importante que sufrió la estadística en la anterior reforma legislativa debido a su exclusión de la asignatura de matemáticas de segundo de bachillerato en la opción de ciencias y tecnología y como consecuencia de las pruebas de acceso a la universidad (selectividad); esto influyó no solo en el curso del que fueron eliminados sino también en los anteriores.

En el curso 2014-2015 se realizaron las primeras pruebas de 3º de primaria. En la página oficial del MECD podemos descargar ejemplos para este nivel. No están disponibles todavía las de secundaria y bachillerato. Estos son ejemplos de preguntas de pruebas externas de 3º de primaria[iv]

Evaluación Individualizada
Tercero de Educación Primaria

Competencia matemática	
TAREA 3	Trata de responder si son verdaderas o falsas las afirmaciones de la tabla. Marca con una "X" la casilla correspondiente. 
Bloque de contenido	Estadística y probabilidad
Criterio de evaluación	Observar y constatar que hay sucesos imposibles, sucesos que con casi toda seguridad se producen, o que se repiten, siendo más o menos probable esta repetición
Destreza	Resolución de problemas
Estándar de aprendizaje	Realiza conjeturas y estimaciones sobre algunos juegos (monedas, dados, cartas, lotería…) utilizando el vocabulario suceso seguro, suceso posible o suceso imposible

Afirmaciones: Al lanzar el dado…	Verdadero	Falso
"Sacar un número del 1 al 6" es un suceso seguro.		
"Sacar un 7" es un suceso posible.		
"Sacar un 4" es un suceso imposible.		
"Sacar un 2 o un 3" es un suceso posible.		

Figura 1: Ejemplo de prueba estándar de 3º de primaria

Evaluación de tercer curso de Educación Primaria			ÍTEM Nº: CM24
Competencia matemática			
TÍTULO DE LA UNIDAD DE EVALUACIÓN: **El resfriado**			
BLOQUE DE CONTENIDO	Incertidumbre y datos.		
PROCESO COGNITIVO	Aplicar y analizar. Análisis.		
DESTREZA	❒Cálculo	☒ Resolución de problemas	
DIFICULTAD ESTIMADA	☒Baja	❒Media	❒Alta
ESTÁNDAR DE APRENDIZAJE	CC.SS. Analiza informaciones y maneja imágenes, tablas, gráficos, esquemas, resúmenes y las TIC		

Respuesta: ❒ Cerrada ☒Semiconstruida ❒Construida ❒ Abierta

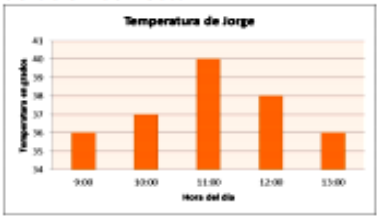
La temperatura que marca el termómetro que se pone Jorge está representada en la siguiente tabla:

Hora del día	Temperatura en grados
9:00	36
10:00	37
11:00	40
12:00	38
13:00	36

Completa el siguiente gráfico con los datos de la tabla:

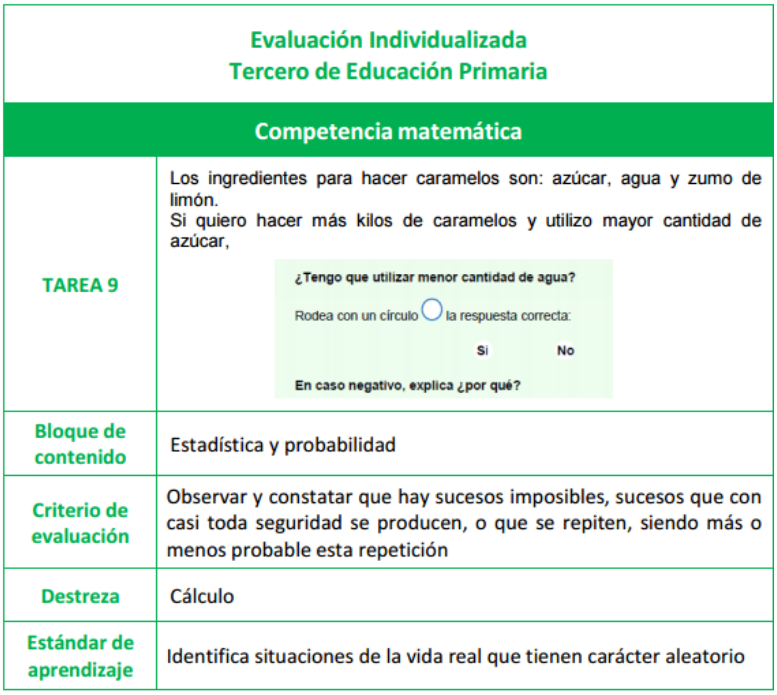


Solución correcta



CRITERIOS DE CORRECCIÓN	**Máxima puntuación:** Código 1: Dibuja bien las cuatro barras (se admiten líneas verticales). **Ninguna puntuación:** Código 0: Cualquier otra respuesta.

Figura 2: Ejemplo de prueba de 3º de primaria



Evaluación Individualizada Tercero de Educación Primaria	
Competencia matemática	
TAREA 9	Los ingredientes para hacer caramelos son: azúcar, agua y zumo de limón. Si quiero hacer más kilos de caramelos y utilizo mayor cantidad de azúcar, ¿Tengo que utilizar menor cantidad de agua? Rodea con un círculo ◯ la respuesta correcta: Si No En caso negativo, explica ¿por qué?
Bloque de contenido	Estadística y probabilidad
Criterio de evaluación	Observar y constatar que hay sucesos imposibles, sucesos que con casi toda seguridad se producen, o que se repiten, siendo más o menos probable esta repetición
Destreza	Cálculo
Estándar de aprendizaje	Identifica situaciones de la vida real que tienen carácter aleatorio

Figura 3: Ejemplo de prueba de 3º de primaria

Más información y ejemplos de pruebas de evaluación de 3º de primaria en la página web[v] del MECD

4. METODOLOGÍAS ACTIVAS: TRABAJOS DE INVESTIGACIÓN

En el currículo de algunas materias LOMCE, como Biología y Geología, existe un bloque llamado "proyecto de investigación" cuyo contenido principal es la realización de un proyecto de investigación en equipo. En el caso de matemáticas se menciona como "investigación matemática en distintos contextos", y se concreta para estadística en las "Fases y tareas de un estudio estadístico". No es fácil encontrar información en la red sobre cómo incorporar en el aula las actividades de este tipo. Un excelente recopilatorio es el de los concursos organizados por las sociedades españolas de estadística y los organismos colaboradores, como por ejemplo SGAPEIO o la SEIO. Se echa en falta un aporte de las editoriales o de los organismos oficiales, que en otros países ofrecen en sus webs recursos para el profesorado con ejemplos prácticos. Un caso es la web NZMaths del ministerio de educación de Nueva Zelanda[vi], la oficina australiana de estadísticas[vii] o la asociación americana de estadística[viii]

5. CONCLUSIONES

Cualquier reforma educativa, y esta no difiere de las anteriores, busca un cambio metodológico respecto a la forma de enseñar en las aulas. Podemos estar o no de acuerdo con el enfoque elegido pero no podemos relegar la lectura del currículo al listado de contenidos, porque la visión que tendremos será sesgada y probablemente nuestra conclusión será errónea.

La LOMCE enfatiza el aprendizaje por competencias desde todas las áreas y con distintos niveles de desempeño. Su grado de adquisición se medirá, utilizando los estándares de aprendizaje evaluables, y entre los instrumentos de medición se encuentran las pruebas externas, cuya aplicación se debate en la actualidad.

El bloque de Estadística y Probabilidad está presente en todas las etapas educativas dentro de la materia de matemáticas, aunque sigue figurando al final de la relación en todos los cursos. En su descripción cobran significado las investigaciones estadísticas asociadas a situaciones reales y el tratamiento de datos para extraer conclusiones y tomar decisiones.

Queda en manos del profesorado la interpretación de la ley como una oportunidad de incorporar nuevas situaciones de aprendizaje que conviertan a nuestro alumnado en ciudadanos estadísticamente

competentes o continuar con ejercicios rutinarios que empiezan y terminan en el aula y se traducen en un compendio más de prácticas algorítmicas sin mayor relevancia para sus vidas.

REFERENCIAS

[i] Se pueden consultar el Real Decreto 126/2014, de 28 de febrero, por el que se establece el currículo básico de la Educación Primaria y en el Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato

[ii] http://www.unesco.org/education/pdf/DELORS_S.PDF DELORS, Jacques y otros (1996). La educación encierra un tesoro: informe a la UNESCO de la Comisión Internacional sobre la Educación en el siglo XXI

[iii] Se pueden consultar el Real Decreto 126/2014, de 28 de febrero, por el que se establece el currículo básico de la Educación Primaria y en el Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato

[iv] http://www.mecd.gob.es/dctm/inee/evaluacionterceroprimaria/pruebas-oficiales-201415/extremadura/pautasdecorreccione.primariacm2015.pdf?documentId=0901e72b81e2aae9

[v] http://www.mecd.gob.es/inee/Evaluacion_tercero_Primaria.html

[vi] http://nzmaths.co.nz/resource/planning-statistics-investigation

[vii] http://www.abs.gov.au/education

[viii] http://www.amstat.org/education/stew/index.cfm

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

CARREIRAS POPULARES, APRENDIZAXE ESTATÍSTICA E VALORES SOCIAIS

Xosé Enrique Pujales Martínez

Catedrático de Instituto xubilado

RESUMO

O traballo que se presenta foi pensado para ser desenvolto co alumnado para o Concurso Incubadora de Sondaxes e Experimentos de 2014, pero problemas administrativos impediron a súa realización con el.

Partindo do auxe entre a cidadanía da actividade de correr e da súa participación en carreiras populares, proponse estudar: o incremento das carreiras de pago fronte as anteriormente predominantes carreiras de balde; a diferente proporción participante real / persoa inscrita en cada unha delas; as características que acompañan ás carreiras populares; a competitividade presente nesta actividade; o axuste a unha distribución normal dos tempos de chegada dos corredores e corredoras.

O estudo destas cuestións serve para aprender contidos conceptuais estatísticos, empregar distintos gráficos e, especialmente, desenvolver competencias, adquirir capacidades e destrezas, e traballar valores e actitudes positivos para a sociedade.

Palabras e frases chave: carreiras populares, gráficos estatísticos, test de bondade de axuste, valores sociais.

1. INTRODUCIÓN

Correr por paseos, camiños, montes,… está de moda. E as persoas que desenvolven esta saudábel actividade non se quedan niso senón que se animan a participar e competir nas carreiras populares que se organizan por toda a xeografía galega. Esta realidade permítenos formular cuestións nas que a Estatística pode aparecer como unha excelente ferramenta para entendelas. No traballo de investigación que acompaña á resolución das cuestións propostas, o alumnado terá a ocasión de coñecer e aplicar conceptos de estatística, empregar os gráficos estatísticos máis adecuados á situación estudada, á vez que aprender a buscar e organizar a información, manexar programas de computadoras, reflexionar sobre os resultados obtidos, sacar conclusións, traballar en equipo dun xeito responsábel, ser constante nas tarefas encomendadas, manter unha actitude crítica, etc.

2. CARREIRAS DE PAGO – CARREIRAS DE BALDE. PARTICIPACIÓN

Hai poucos anos as inscricións á inmensa maioría das carreiras populares eran de balde. Agora case todas (máis dun 80%) son de pago. A obtención da proporción actual é un traballo de investigación que obrigaría ao alumnado a estudar: qué medios debemos utilizar para obter esa información; qué escolleremos, a poboación ou unha mostra; cómo obter unha mostra representativa; cál é a opinión dos corredores e corredoras ante esta situación; qué reaccións se dan entre eles; qué nos parece a alternativa de correr sen dorsal;… Para estudar estas cuestións será importante, como en todo traballo de investigación, contar co asesoramento de expertos.

Nas carreiras hai que inscribirse previamente. Despois de inscritos, podemos presentarnos ou non á proba atlética. Algúns dos aspectos a investigar son os seguintes: qué porcentaxe de corredores se presentan respecto aos inscritos nas carreiras de pago; qué porcentaxe ás de balde; qué gastos provocan os corredores aos organizadores; cómo representar os resultados; cál é a valoración dos resultados; cómo valorar a actitude dos que non se presentan; cál é a posíbel consecuencia desta actitude;…

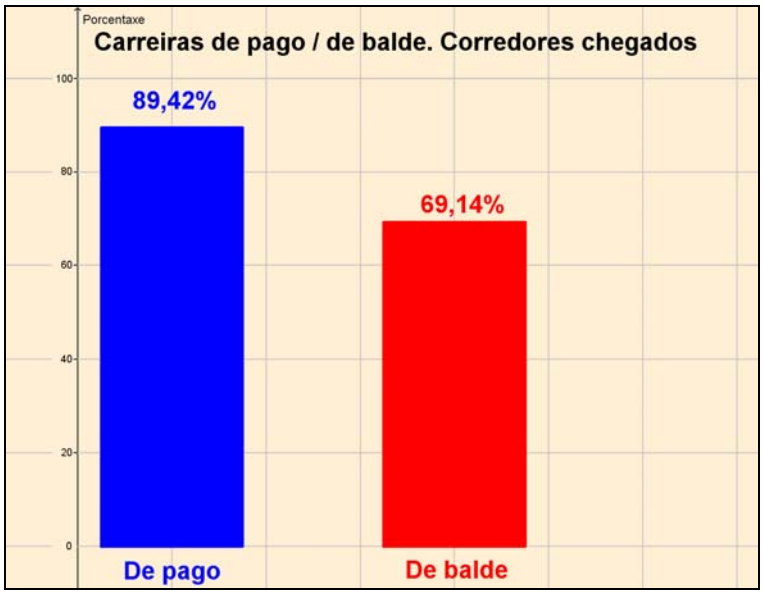


Figura 1. Proporción de asistentes nos dous tipos de carreiras



Figura 2. Despois de 37 edicións de balde, pasou a ser de pago pola abundante ausencia de corredores

Grazas a estas cuestións trabállanse contidos conceptuais e procedimentais, pero sobresae a potenciación de valores e actitudes (constancia para obter os datos inscritos –que se volatilizan en poucos días– esforzo, responsabilidade, actitude crítica,…)

3. CARREIRAS POPULARES. CARACTERÍSTICAS

É incorrecto pensar que ás carreiras populares realizadas ao largo de toda a xeografía de Galicia se inscriben os afeccionados e afeccionadas do concello organizador e dos arredores. Unha razón importante para que pase isto é a existencia de dúas ligas competitivas entre corredores, a de *Champion Chip Norte* e, sobre todo, a de *Correr en Galicia*. Ditas ligas fai que moitos corredores estean dispostos a desprazarse a

calquera lugar co fin de conseguir puntos para as ligas na que está inscrito. Por este motivo, os tempos conseguidos nas carreiras non son os esperados nunha carreira "popular". Isto nótase nunhas carreiras máis que noutras e, daquela, hai carreiras máis "populares" que outras. Recollendo datos de distintas carreiras ou de diferentes edicións da mesma, sería interesante que o alumnado tratara de tipificar as carreiras populares. Algúns dos padróns que se dan nestas son:

- Redución da porcentaxe de boas marcas.
- Incremento da Mediana.
- Incremento da porcentaxe de mulleres corredoras.

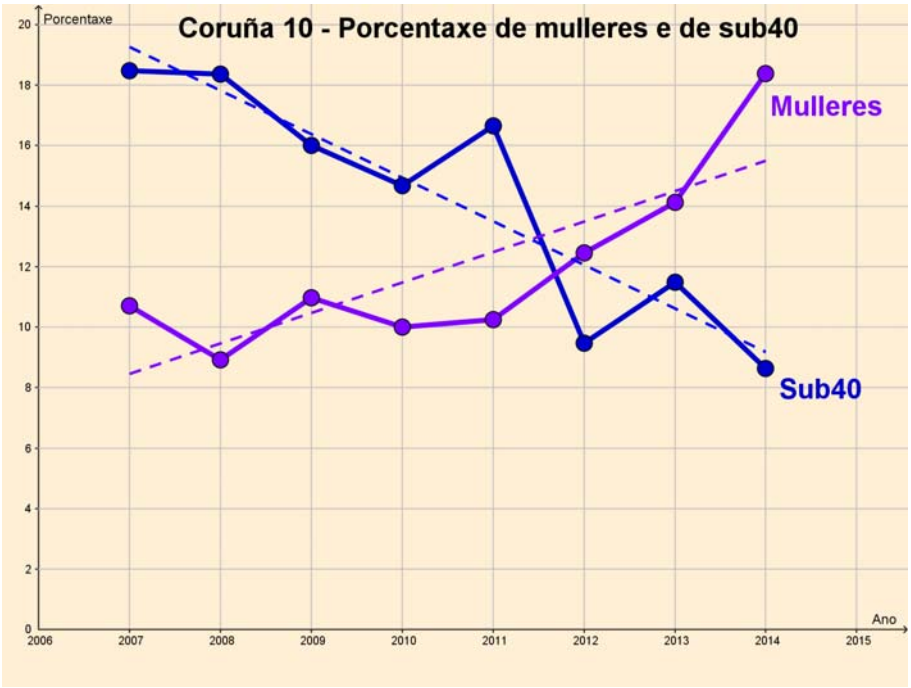


Figura 3. Evolución das porcentaxes de mulleres e de sub40 en Coruña 10

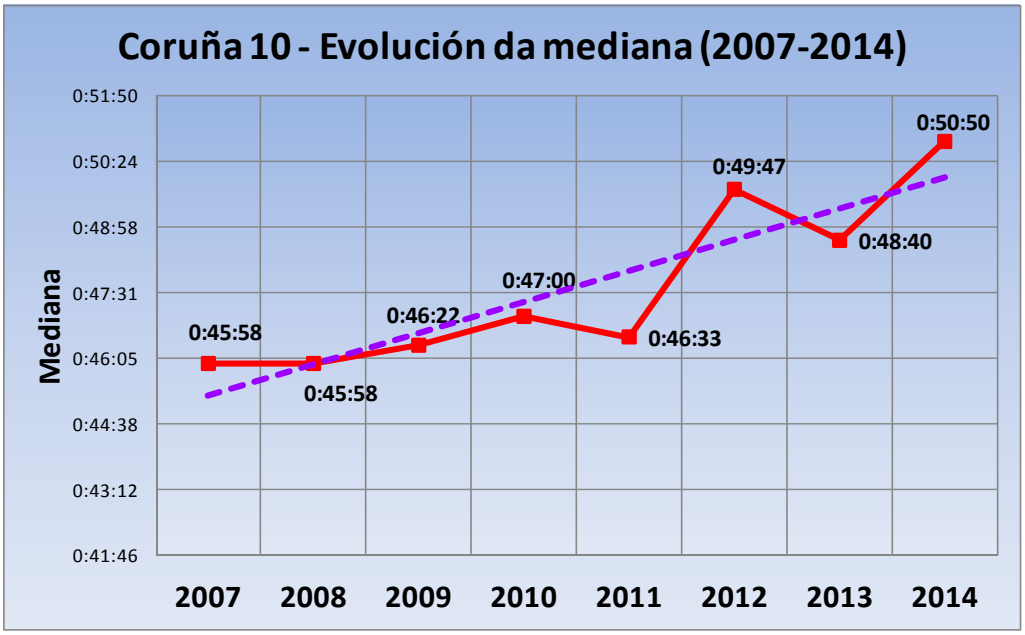


Figura 4. Incremento da mediana nas distintas edicións de Coruña 10

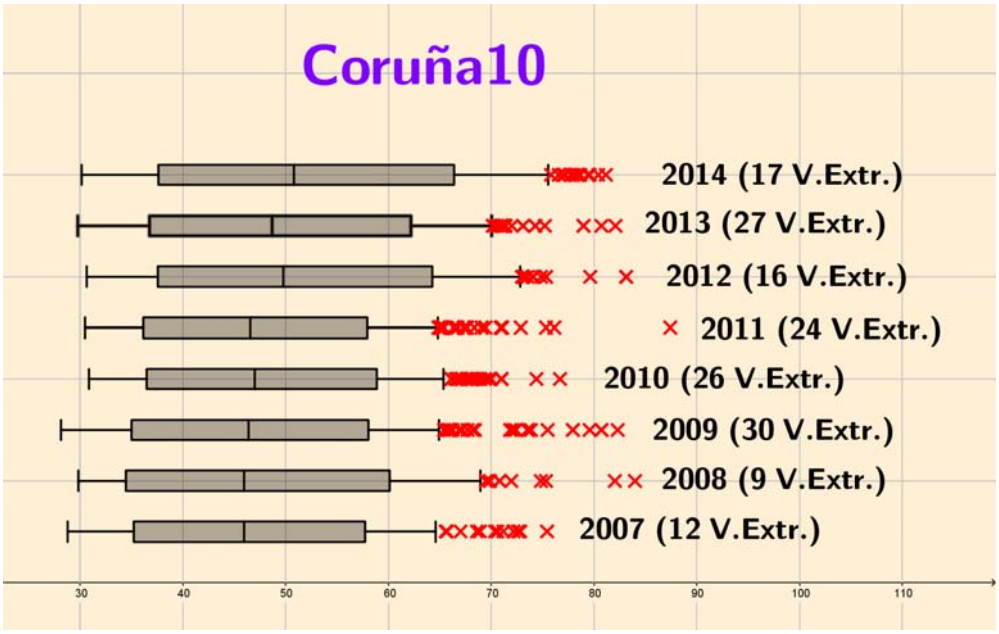


Figura 5. Evolución dos diagramas de caixa na carreira Coruña 10

Na figura 5 os diagramas de caixa permiten comparar unhas edicións con outras, podendo analizar as diferenzas entre elas e concluír que edición é a máis popular de todas. O traballo con este tipo de gráfico permite debater as vantaxes e os inconvenientes deste diagrama con outros, como por exemplo os histogramas. Por exemplo, na figura 6 vese que cando estudamos unha soa carreira, o histograma dá máis información.

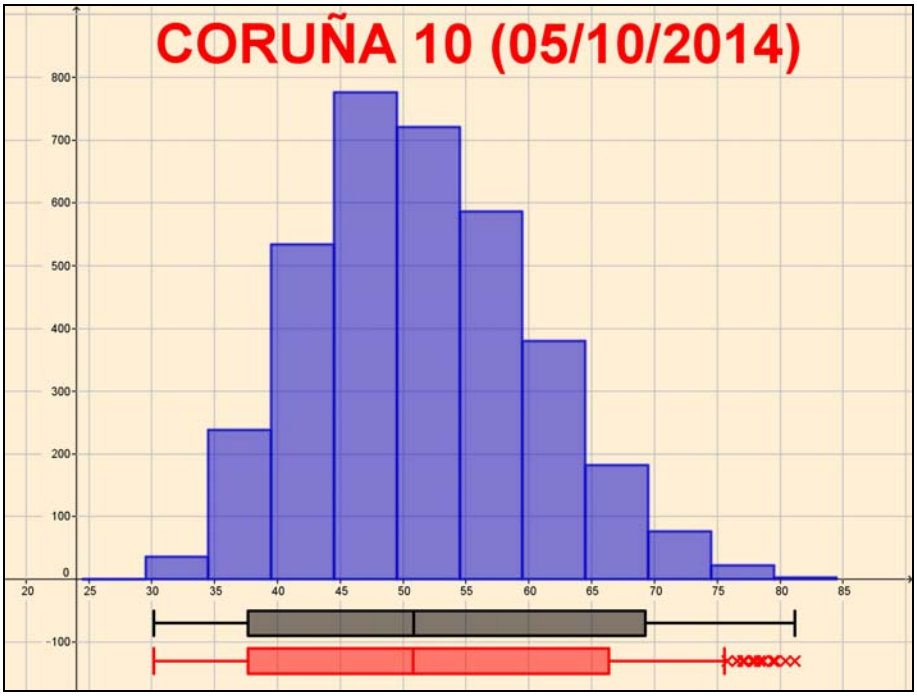


Figura 6. Histograma, diagrama de caixa sen valores extremos e con eles.

4. CORREDORES POPULARES E ESPÍRITO COMPETITIVO

Os afeccionados e afeccionadas a correr participan nas carreiras populares por varios motivos: reto persoal, afán de superación, confraternidade,… pero tamén van aparecendo nesta práctica aspectos non idílicos, produto dun espírito competitivo insán, que leva a facer trampas a si mesmo e aos demais. Un deses trazos que se pode estudar estatisticamente é a colocación dos corredores no lugar de saída das

carreiras. Cada participante debería situarse detrás da saída a unha distancia dela acorde ás súas capacidades. Non facer isto entraña perigos. E podemos estudar este aspecto representando unha nube de puntos onde cada corredor está representado por un punto correspondente ao seu posto na chegada e o tempo de retraso na saída.

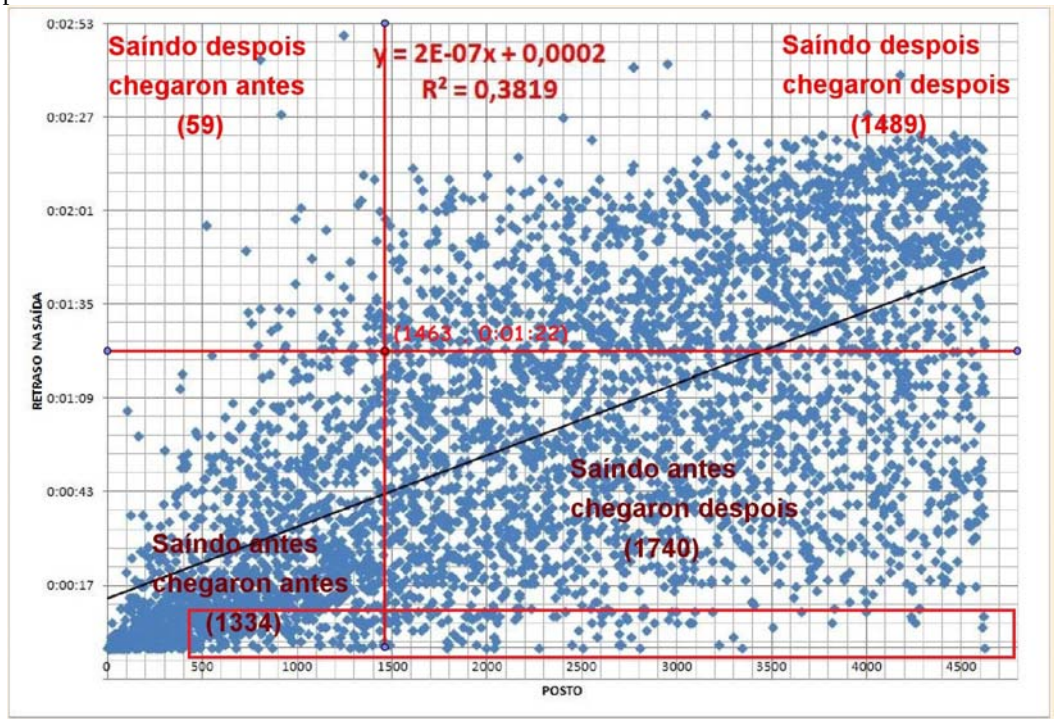


Figura 7. Nube de puntos de corredores na Vig-Bai de 2014

Facer un estudo deste tipo serve para aplicar o concepto de recta de regresión, incrementar a destreza para obter, representar e localizar información, mellorar a competencia dixital, profundar na destreza computacional, interpretar gráficas e potenciar a actitude crítica ante o comportamento humano.

5. TEMPOS DE CHEGADA DOS CORREDORES E TEST DE BONDADE

Toda persoa que contemplou a chegada dunha carreira popular terá observado que os primeiros clasificados chegan aos poucos, despois vai aumentando a súa frecuencia e máis adiante esta comeza a diminuír. Unha persoa que teña escoitado falar da curva de Gauss podería preguntarse se a frecuencia de persoas que chegan a meta axústase a unha distribución normal. E é importante potenciar a curiosidade, animar a que se formulen hipóteses como a anterior, que se busquen os medios de comprobación, que se utilicen e que se obteñan unhas conclusións baseadas nos datos non dun só caso (particular) senón nos dunha mostra representativa de carreiras.

Para a resolución desta cuestión podemos empregar, segundo o nivel do alumnado, métodos subxectivos e cualitativos: o histograma semella unha campá simétrica?; o diagrama de caixa é simétrica e sen valores extremos?; as poligonais dos valores obtidos e dos valores esperados se a distribución fose normal son semellantes?. Pero tamén podemos introducir e aplicar coa axuda dunha computadora un test de bondade de axuste, que nos dea un criterio obxectivo e cuantitativo da cuestión proposta.



Figura 8. Aplicación a unha carreira do test de bondade de axuste a unha distribución normal

REFERENCIAS

Pujales, X.E. (2014) Estatística, corredores e carreiras populares. XXVII Congreso de ENCIGA.

Triola, M.F. (2013) Estadística. 11ª edición. Pearson.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

PROPOSTAS DE TRABALLOS ESTATÍSTICOS PARA AS AULAS DE SECUNDARIA

Covadonga Rodríguez-Moldes Rey[1]

[1] Catedrática de Matemáticas. IES Mugardos (A Coruña)

RESUMO

A Estatística ocupa unha parte importante do currículo das matemáticas en secundaria aparecendo sempre como último bloque de coñecementos despois dos de números, álxebra, xeometría e análise. Posiblemente esta colocación sexa a causa de que non sempre se imparta este bloque coa atención que se merece a pesares da súa potencialidade didáctica e as aplicacións prácticas que ten. Nesta comunicación preséntanse varias propostas de traballos estatísticos levados a cabo nas aulas de secundaria, que van dende o estudo dos peixes do mar e a lonxa en 1º de ESO ata os resultados dunha degustación de chicle en 4º de ESO.

Palabras e frases chave: Secundaria, estatística, matemáticas, traballos, aula

1. INTRODUCIÓN

Nos currículo de matemáticas en secundaria a Estatística é un bloque de coñecementos que aparece en último lugar despois de Números, Álxebra, Xeometría e Análise. Así ven sendo en todas as leis educativas habidas en España dende hai máis de 30 anos. Invariablemente, o último bloque de contidos da materia Matemáticas na ESO é a Estatística e a Probabilidade. Posiblemente esta colocación ao final, xunto coa amplitude dos programas escolares, sexa a causa de que non sempre se imparte este bloque coa atención que se merece a pesares da súa enorme potencialidade didáctica e das aplicacións prácticas que ten. En moitos casos, máis dos recomendables, limítase a súa presenza na aula á explicación dos elementos e fórmulas básicas da estatística unidimensional e á realización de exercicios elementais nun curto espazo de tempo.

Na última reforma que estamos experimentando, a LOMCE, en Galicia o currículo xira arredor dos criterios de avaliación de xeito que a cada criterio se lle asocian contidos, estándares de aprendizaxe, competencias e obxectivos, polo tanto, si analizamos os criterios de avaliación do bloque Estatística e Probabilidade nos cursos da ESO, podemos darnos unha idea do tratamento que a LOMCE lle da a este bloque, que semella ter bastante presente no currículo á Estatística e á Probabilidade. Analicemos, pois, eses criterios de avaliación.

Criterios de avaliación LOMCE do bloque de Estatística e Probabilidade na ESO

Materia: Matemáticas
Cursos: 1º ESO e 2º ESO

- B5.1. Formular preguntas axeitadas para coñecer as características de interese dunha poboación e recoller, organizar e presentar datos relevantes para respondelas, utilizando os métodos estatísticos apropiados e as ferramentas adecuadas, organizando os datos en táboas e construíndo gráficas, calculando os parámetros relevantes e obtendo conclusións razoables a partir dos resultados obtidos.
- B5.2. Utilizar ferramentas tecnolóxicas para organizar datos, xerar gráficas estatísticas, calcular parámetros relevantes e comunicar os resultados obtidos que respondan ás preguntas formuladas previamente sobre a situación estudada.
- B5.3. Diferenciar os fenómenos deterministas dos aleatorios, valorando a posibilidade que ofrecen as matemáticas para analizar e facer predicións razoables acerca do comportamento dos aleatorios a partir das regularidades obtidas ao repetir un número significativo de veces a experiencia aleatoria, ou o cálculo da súa probabilidade.

- B5.4. Inducir a noción de probabilidade a partir do concepto de frecuencia relativa e como medida de incerteza asociada aos fenómenos aleatorios, sexa ou non posible a experimentación.

Materia: Matemáticas orientadas ás ensinanzas académicas (MAC)
Curso: 3º ESO

- B5.1. Elaborar informacións estatísticas para describir un conxunto de datos mediante táboas e gráficas adecuadas á situación analizada, xustificando se as conclusións son representativas para a poboación estudada.
- B5.2. Calcular e interpretar os parámetros de posición e de dispersión dunha variable estatística para resumir os datos e comparar distribucións estatísticas.
- B5.3. Analizar e interpretar a información estatística que aparece nos medios de comunicación, valorando a súa representatividade e a súa fiabilidade.
- B5.4. Estimar a posibilidade de que aconteza un suceso asociado a un experimento aleatorio sinxelo, calculando a súa probabilidade a partir da súa frecuencia relativa, a regra de Laplace ou os diagramas de árbore, e identificando os elementos asociados ao experimento.

Curso: 4º ESO

- B5.1. Resolver situacións e problemas da vida cotiá aplicando os conceptos do cálculo de probabilidades e técnicas de reconto axeitadas.
- B5.2. Calcular probabilidades simples ou compostas aplicando a regra de Laplace, os diagramas de árbore, as táboas de continxencia ou outras técnicas combinatorias.
- B5.3. Utilizar o vocabulario axeitado para a descrición de situacións relacionadas co azar e a estatística, analizando e interpretando informacións que aparecen nos medios de comunicación e fontes públicas oficiais (IGE, INE, etc.).
- B5.4. Elaborar e interpretar táboas e gráficos estatísticos, así como os parámetros estatísticos máis usuais, en distribucións unidimensionais e bidimensionais, utilizando os medios máis axeitados (lapis e papel, calculadora ou computador), e valorando cualitativamente a representatividade das mostras utilizadas.

Materia: Matemáticas orientadas ás ensinanzas aplicadas (MAP)
Curso: 3º ESO

- B5.1. Elaborar informacións estatísticas para describir un conxunto de datos mediante táboas e gráficas adecuadas á situación analizada, e xustificar se as conclusións son representativas para a poboación estudada.
- B5.2. Calcular e interpretar os parámetros de posición e de dispersión dunha variable estatística para resumir os datos e comparar distribucións estatísticas.
- B5.3. Analizar e interpretar a información estatística que aparece nos medios de comunicación, e valorar a súa representatividade e fiabilidade.

Curso: 4º ESO

- B5.1. Utilizar o vocabulario axeitado para a descrición de situacións relacionadas co azar e a estatística, analizando e interpretando informacións que aparecen nos medios de comunicación e fontes públicas oficiais (IGE, INE, etc.).
- B5.2. Elaborar e interpretar táboas e gráficos estatísticos, así como os parámetros estatísticos máis usuais, en distribucións unidimensionais, utilizando os medios máis axeitados (lapis e papel, calculadora, folla de cálculo), valorando cualitativamente a representatividade das mostras utilizadas.
- B5.3. Calcular probabilidades simples e compostas para resolver problemas da vida cotiá, utilizando a regra de Laplace en combinación con técnicas de reconto como os diagramas de árbore e as táboas de continxencia.

Ademáis deste bloque específico adicado á Estatística e á Probabilidade existe un primeiro bloque común a toda a ESO, é o Bloque 1, chamado **Procesos, métodos e actitudes en matemáticas** que *debe desenvolverse de xeito transversal e simultaneamente ao resto de bloques, constituíndo o fío condutor da materia; articúlase sobre procesos básicos e imprescindibles no quefacer matemático: resolución de problemas, proxectos de investigación matemática para realizar de maneira individual ou en grupo, matematización e modelización, actitudes adecuadas para desenvolver o traballo científico, e utilización*

de medios tecnolóxicos. Entre os criterios de avaliación deste bloque están os seguintes relacionados coa Estatística e a Probabilidade:

- B1.3. Describir e analizar situacións de cambio, para atopar patróns, regularidades e leis matemáticas, en contextos numéricos, xeométricos, funcionais, estatísticos e probabilísticos, valorando a súa utilidade para facer predicións.
- B1.5. Elaborar e presentar informes sobre o proceso, resultados e conclusións obtidas nos procesos de investigación.
- B1.6. Desenvolver procesos de matematización en contextos da realidade cotiá (numéricos, xeométricos, funcionais, estatísticos ou probabilísticos) a partir da identificación de problemas en situacións problemáticas da realidade.
- B1.11. Empregar as ferramentas tecnolóxicas adecuadas, de xeito autónomo, realizando cálculos numéricos, alxébricos ou estatísticos, facendo representacións gráficas, recreando situacións matemáticas mediante simulacións ou analizando con sentido crítico situacións diversas que axuden á comprensión de conceptos matemáticos ou á resolución de problemas.
- B1.12. Utilizar as tecnoloxías da información e da comunicación de maneira habitual no proceso de aprendizaxe, procurando, analizando e seleccionando información salientable en internet ou noutras fontes, elaborando documentos propios, facendo exposicións e argumentacións destes, e compartíndoos en ámbitos apropiados para facilitar a interacción.

Con esta introdución preténdese xustificar a conveniencia de impulsar a realización de traballos e proxectos de investigación estatística con alumnado de ESO comprobando que son actividades plenamente integradas no actual currículo. Nesta comunicación preséntanse traballos destas características realizados con alumnado de ESO coordinados pola autora. Estes traballos traballos non tipo concurso Incubadora de Sondaxes e Experimentos pois este concurso limita o número de persoas por traballo, son traballos nos que participou todo o alumnado dun grupo –ás veces de máis dun grupo- coordinados pola autora da comunicación.

2. PROPOSTAS DE TRABALLOS ESTATÍSTICOS

A continuación descríbense proxectos realizados con distintos grupos de ESO nos que se traballou a estatística con todo o alumnado do grupo arredor dun tema específico. Como característica común está a exposición e presentación dos resultados en público.

Título: Guía para un consumo responsable do chicle
Nivel: 3º-4º ESO
Data de realización: 2000 (4º PDC)

Nos grupos chamados Programas de Diversificación Curricular, PDC, de 3º e 4º de ESO –que coa LOMCE transfórmanse en Programas de Mellora- dase o habitat ideal para realizar proxectos e traballos de investigación porque hai un profesor ou profesora que se encarga dun ámbito, neste caso o científico-tecnolóxico, durante oito sesións lectivas á semana. Son grupos con poucos alumnos cun perfil especial no que voluntade de mellorar prima sobre outras cualidades e que coa atención casi individualizada da que dispoñen, logran acadar os obxectivos da ESO e graduarse con alta probabilidade.

Este traballo tivo cinco partes ben diferenciadas coas que se buscaba abarcar un amplo abano de posibilidades de traballar contidos científico-matemáticos e de valores e actitudes vitais. Foron estas:
1ª: O chicle, historia e curiosidades
2ª: De onde sae o chicle?. Como se fai?
3º: O chicle e a saúde
4ª: Resultados dunha degustación
5ª: Chicle si...¡pero non así!

Centrarémonos no apartado 4ª que involucrou un alto número de alumnos e alumnas da ESO nunha degustación de chicle deseñada e levada a cabo polo alumnado de 4º PDC. Así foi como contaron o comezo:

"O primeiro era selecciona-lo sabor (porque non se poden comparar chicles de distintos sabores, sería un lío). O sabor elixido foi o de menta, xa que é o máis consumido debido a que disimula o alento a tabaco. Collimos para o noso estudio os catro chicles de menta de máis venta no instituto.
A continuación había que decidir a quen lle faríamos a enquisa-degustación. Pensamos que unha boa mostra sería un grupo de cada curso; así fixemo-la a 92 alumnos e alumnas de 1ºA, 2ºB, 3ºB e 4ºB
Tivemos que usa-lo ordenador para elaborar o formulario de respostas
Había que puntuar de 1 a 5 os seguintes elementos: sabor, textura e globos. Como o sabor parécenos o máis importante, tería que ter máis peso; decidimos que os pesos serían : sabor=3, textura=1, globos=1; e chegamos á nosa fórmula do chicle que é unha media ponderada: Nota do chicle = (3. sabor + 1.textura + 1.globos)/5"

A continuación veu a organización e preparación das degustacións coa a compra de materiais, a petición dos permisos necesarios, o establecemento do calendario..... E fixeronse as degustacións con enorme expectación pola novidade que supoñía esta actividade.

Os datos obtidos foron procesados informaticamente nunha folla de cálculo na que o alumnado tivo que traballar con moitos datos, efectuar cálculos estatísticos sinxelos e elaborar bastantes gráficas para resumir todos estes datos. Finalmente realizaron a ficha técnica e obtiveron conclusións como a porcentaxe de consumo de chicle e os chicles preferidos entre o alumnado do centro.

Cando todos os apartados estiveron completados - o quinto consistiu na recopilación de imaxes de malos hábitos no consumo do chicle-, houbo que presentar o traballo a todos os compañeiros e preparar un dossier que foi premiado no certame do Xove Consumidor Europeo.

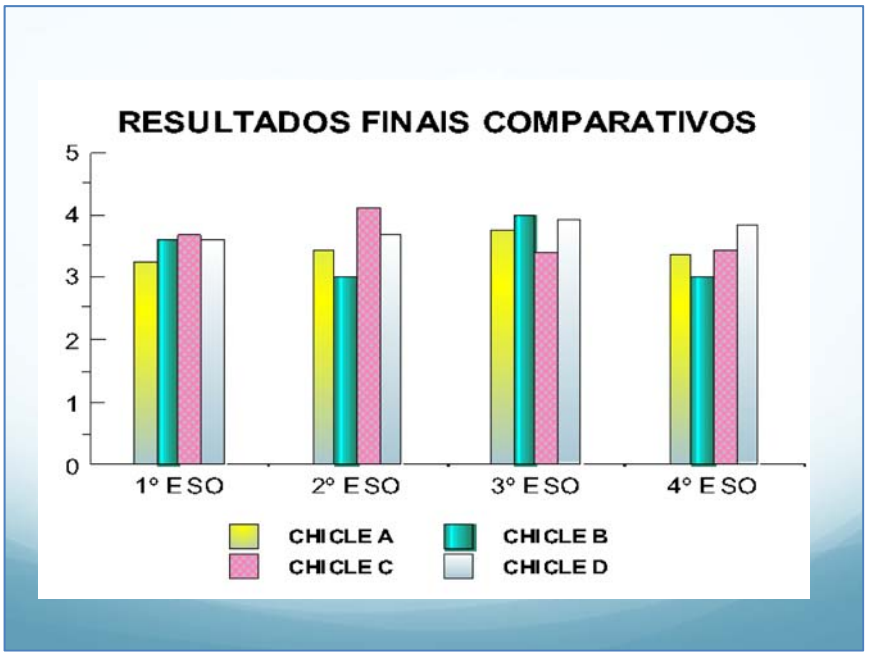


Ilustración 1. Unha das gráficas resumo da degustación de chicle

Título: Alimentarse ou *alimentirse*
Nivel: 3º-4º ESO
Data de realización: 2001 (3º ESO)

Realizamos un estudo dos hábitos alimentarios do alumnado do IES Mugardos para o que houbo que recoller datos, procesalos e obter resultados; posteriormente eses resultados presentáronse a un especialista en nutrición e dietética, o presidente da Sociedade de Nutrición e Dietética de Galicia, Dr Diego Bellido, co obxecto de obter un diagnóstico e detectar posibles desequilibrios nutricionais no

alumnado do centro. Finalmente, todo o traballo presentouse aos compañeiros do centro e aos pais e nais para que foran conscientes dos erros e intentaran mellorar a dieta á vez que se promovían accións saudables no centro como a incentivación do consumo de froita.

Pretendíamos estudar a alimentación do alumnado dende que se levantaba ata as tres da tarde, hora de volta a casa do colexio. Estes datos recolléronse de dúas formas, por un lado a través dunha enquisa que o alumnado pasou ao alumnado nas aulas e por outro, como queríamos estudar o que se consumía na cafetería do centro, realizaron unha recollida exhaustiva de datos organizados en parellas de rapaces que se turnaban para anotar nun formulario o consumo na cafetería en cada recreo.

Houbo que procesar moitos datos e realizar moitas gráficas para obter os resultados que se agruparon en dúas categorías: resultados sobre o almorzo na casa do alumnado e resultados do consumo na cafetería do instituto.

Cos resultados da casa elaboramos o almorzo típico do alumnado do IES Mugardos: nada de froita, leite con cacao, galletas ou cereais e descubríronse problemas nutricionais. E cos resultados da cafetería descubrimos que no primeiro recreo a consumición preferida son ós bocatas pequenos mentras que no segundo recreo consúmense con diferencia moitas máis golosinas que outras cousas.

Despois de ser analizados os resultados polo Dr Bellido, un grupo de alumnos e alumnas entrevistouno no seu despacho e desa entrevista elaboraron un documento de nove puntos con pautas para mellorar os hábitos alimentarios no IES Mugardos.

Este traballo mereceu un premio do Instituto Galego de Consumo e además, o equipo do Dr Bellido presentou os resultados no congreso nacional de nutrición.

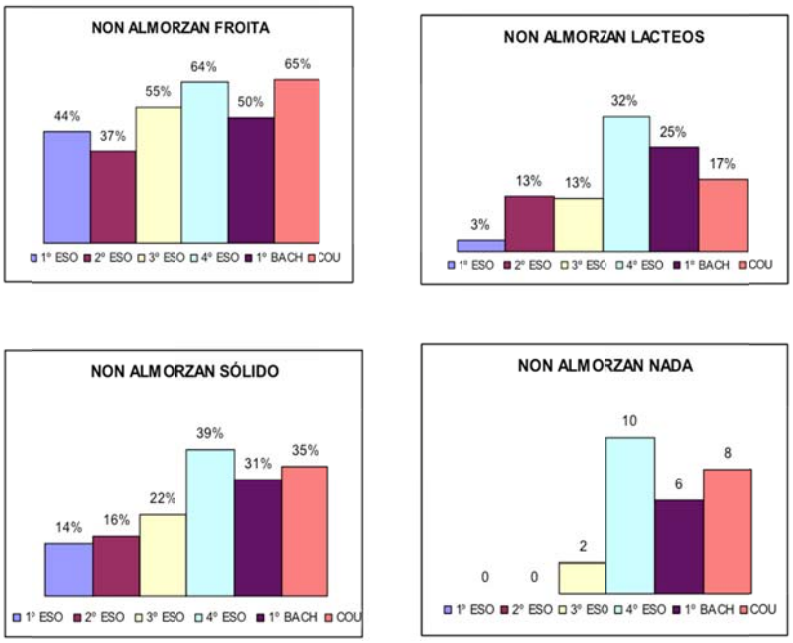


Ilustración 2. Datos do almorzo do alumnado

Tlitulo: Os peixes do mar e a pesca
Nivel: 1º-2º ESO
Data de realización: 2007

Este traballo pretendía unha acción educativa interdiciplinar en 1º de ESO consistente en desenvolver diversos contidos relacionados coas áreas de Matemáticas, Ciencias Naturais e temas transversais a partires da información que baixo o epígrafe "marítima" aparecía diariamente en La Voz de Galicia.

Unha parte desa información era a cotización máxima e mínima das especies subastadas o día anterior nas lonxas de A Coruña e a de Vigo. Centrámonos na de A Coruña e repartimos as especies que se subastaban entre o alumnado de 1º de ESO. Tódolos días ao chegar ao centro, había alumnado encargado de levar á aula a información marítima e cada alumno ou alumna debía comprobar a presencia da especie da que

estaba a facer o seguemento e, despois de calcular o promedio dos precios máximo e mínimo, recollía nunha *ficha* o prezo de cotización e ía construíndo unha gráfica de puntos; ademáis anotaría este dato nun *panel* deseñado para facer un seguemento a todas as especies.

Na clase de Matemáticas, cos datos das cotizacións, cada alumno elaborou unha gráfica para observar a evolución dos prezos da especie da que facía o seguemento e calculou as medidas de centralización. Ademáis, cos datos de tódalas especies ordenaron e construiron pirámides de prezos.

Esta actividade matemática foi unha parte dun proxecto interdisciplinar que levou ese ano o mundo do mar ao instituto rematando no acto cultural de portas abertas titulado "Os príncipes das mareas" no que o alumnado deu conta á comunidade do traballo realizado e rendiu homenaxe aos mariñeiros de Mugardos. Finalmente o traballo foi premiado nos premios Prensa-Escola que organiza La Voz de Galicia.

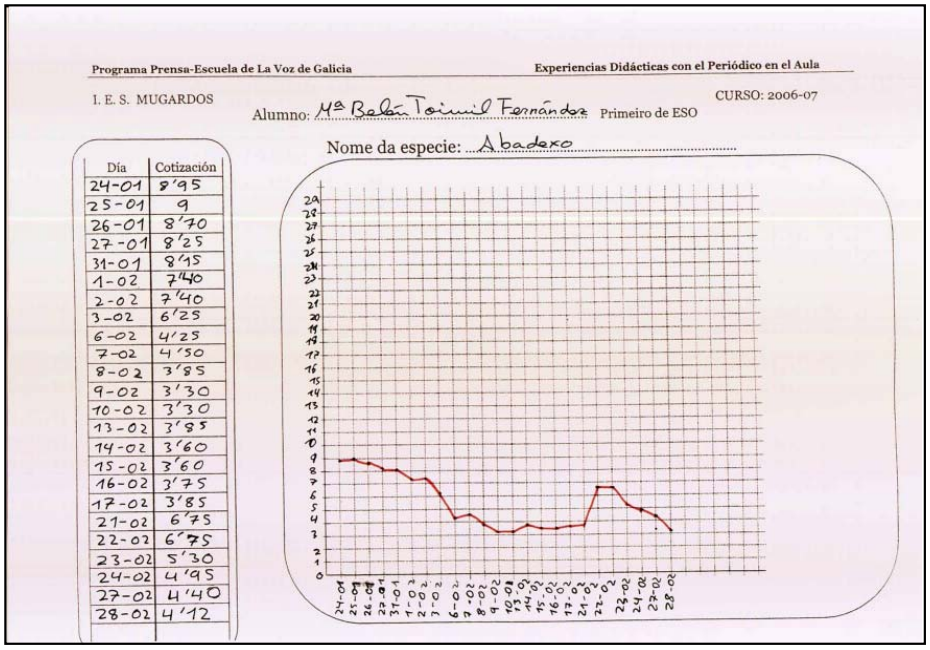


Ilustración 3. Seguemento do prezo do abadexo

Título: O tempo, está cambiando?
Nivel: 1º-2º ESO
Data de realización: 2008

O recoñecemento ao traballo anterior e o aplauso recibido polo alumnado cando contou a experiencia no acto de portas abertas, levou a que a comezo do curso seguinte o ambiente para iniciar un novo proxecto era altamente favorable. Así foi que iniciamos un novo proxecto interdisciplinar no que participou o alumnado de 1º e 2º de ESO e o profesorado de Matemáticas, Sociais e Ciencias.

O comentario: "O tempo...¿está cambiando?" deu lugar a unha conversa na que se mencionaron conceptos como tempo, clima, cambio climático, quencemento global, desastres naturais....con puntos de vista diferentes e conceptos nada claros. Os catro profesores deste proxecto pensamos que é necesario educar ao alumnado para que esté en condicións de emitir opinión crítica e que esta xorde a partires do coñecemento e estudo dun tema, analizándoo dende varios puntos de vista.

O proxecto tiña varias fases que se poden resumir en: estudar o tempo, recoller a opinión dos maiores, analizar informacións e escoitar aos expertos.

Comezamos a estudar o tempo co uso da prensa na aula, esta vez coa páxina meteorolóxica, e repartimos entre o alumnado lugares do mundo dos que aparecían as temperaturas máxima e mínima tentando elixir lugares representativos de distintas zonas xeográficas: ecuador, hemisferio norte e

hemisferio sur. Tamén se recollían as temperaturas máxima e mínima de Galicia e de España. Para esta recollida deseñáronse unhas táboas e estableceuse una rutina diaria: alumnado encargado de traer a prensa á aula e de preocuparse da recollida de datos dos demáis.

Día a día aumentaba o número de datos e comenzáronse a obter os primeiros resultados: a diferenza entre as temperaturas máxima e mínima en Las Palmas é menor que en Madrid, tamén se observaba como en lugares próximos ao ecuador estas diferencias eran pequenas. Sabíamos tamén cales eran os lugares máis frios de Galicia e España, e moitas máis cousas.

Polo que respecta ao alumnado de 1º de ESO, a medida que recollía os datos elaboraba unha gráfica individual dos lugares que estaba seguindo (a modo de serie temporal) e climogramas; por grupos construían mapas climáticos tanto de Galicia como de España, Europa e resto do mundo que colocaban en paneis nos corredores do centro.

Dende o punto de vista da estatística, como o alumnado de 2º de ESO xa sabía facer gráficas e analizar datos, decidimos traballar a competencia dixital traballando os datos recollidos cunha folla de cálculo e aproximando a nube de puntos a unha función cuadrática, desa forma aprenderon como para facer un estudo científico poden sustituirse moitos puntos por expresións matemáticas que permiten estudos rigurosos, relacionados coa probabilidade, por exemplo (pódese calcular así a probabilidade de alcanzar determinadas temperaturas ou de que sucedan fenómenos estraños).

O traballo foi moi amplo e, entre outras actividades, o alumnado recolleu a opinión dos seus familiares maiores para saber como era o clima cando eran nenos, e asistiron a charlas de expertos da Universidade de Santiago. Finalmente contaron todo o traballo realizado á comunidade escolar no acto cultural dese ano.

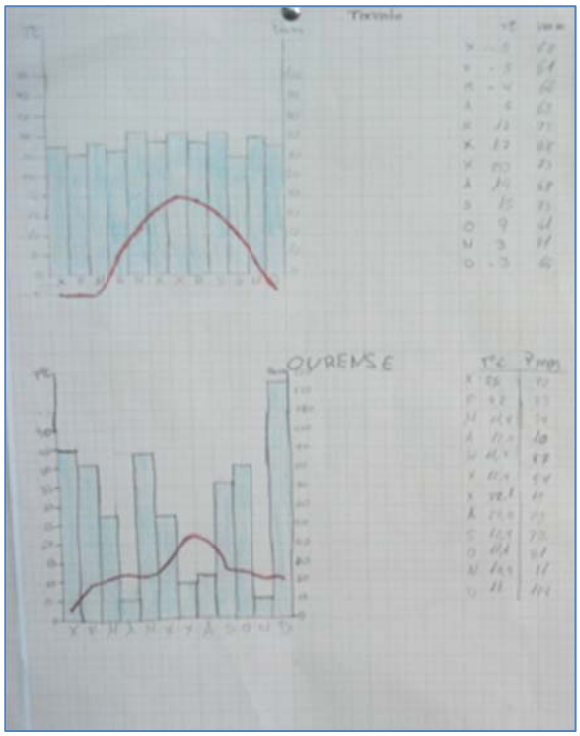


Ilustración 4. Un climograma

3. CONCLUSIÓNS

A experiencia de coordinar traballos estatísticos con alumnado de ESO é altamente satisfactoria a pesares de que existen momentos puntuais nos que se teñen dúbidas do que se está a facer, outros nos que semella ser un traballo excesivo, outros nos que se sinte impotencia...pero finalmente imponse a sensación do deber cumprido sobre todo cando se percibe a satisfacción e o orgullo do alumnado por ser quen de levar a cabo unha investigación de ser capaces de transmitila.

Comunicacións orais

Sesión 2 de Premio SGAPEIO

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

PROBLEMAS DE ASIGNACIÓN EN EXTINCIÓN DE INCENDIOS FORESTALES: UNA APLICACIÓN AL CONTROL DE LOS RECURSOS AÉREOS

J. Rodríguez Veiga[1], I. Gómez-Costa[2], M. J. Ginzo-Villamayor[3] y B. Casas-Méndez[4]

[1] ITMATI, Technological Institute of Industrial Mathematics, Santiago de Compostela, Spain, jorge.rodriguez.veiga@gmail.com
[2] INDRA, Technology and Consulting, A Coruña, Spain, ivan.gomez.cs@gmail.com
[3] Modestya Research Group, Department of Statistics and Operations Research, University of Santiago de Compostela, Santiago de Compostela, Spain, mariajose.ginzo@usc.es
[4] Modestya Research Group, Department of Statistics and Operations Research, University of Santiago de Compostela, Santiago de Compostela, Spain, balbina.casas.mendez@usc.es

RESUMEN

El fenómeno de los incendios forestales se ha convertido en uno de los mayores problemas que nuestros bosques están sufriendo, debido a la alta frecuencia y a la intensidad que han adquirido en las últimas décadas. Es esencial en el control de estos incendios catastróficos una toma de decisiones eficientes, debido a que el presupuesto para los recursos destinados a la extinción del fuego es limitado. En este trabajo, se hace uso de técnicas de investigación operativa que permiten las asignaciones óptimas de aeronaves a las norias de extinción y a las bases de repostaje, que son dos importantes tareas a realizar desde el control de medios aéreos en un incendio forestal

Palabras y frases chave: gestión de incendios forestales, asignación de recursos aéreos, programación lineal entera, norias de extinción, bases de repostaje.

1. INTRODUCCIÓN

El diseño de los sistemas de apoyo a las decisiones en el ámbito de la logística es un área muy activa para la investigación y la realización de aplicaciones dentro de la moderna Investigación Operativa. En el marco del control de incendios forestales, es esencial tomar decisiones eficientes debido a que el presupuesto destinado a los recursos encaminados a la extinción del fuego es limitado. En este sentido, vale la pena mencionar que la teoría económica juega un papel central en la gestión de los incendios forestales. Precursores en el estudio económico de los incendios forestales fueron Headley (1916) y Sparhawk (1925), que describen la forma de establecer un programa óptimo de manejo del fuego.

Un marco teórico utilizado para identificar la forma más eficiente de administrar los costes de un incendio forestal es el *Cost Plus Net Value Change* (C + NVC) (Gorte y Gorte, 1979). Bajo este enfoque, se pretende minimizar el coste por el uso de los recursos en la lucha de un incendio, además de un coste producido por las hectáreas de tierra quemados, donde no sólo hay que tener en cuenta las pérdidas materiales producidas por el fuego (árboles, bienes urbanos, etc.), sino también la reposición o reconstrucción de estas zonas.

En el caso de España, en 2010, surgió el llamado proyecto *Prometeo* con el objetivo de mejorar la eficiencia en la lucha contra incendios. *Prometeo* fue uno de los mayores proyectos de investigación aplicada adjudicados a un consorcio de empresas en España en el ámbito de la lucha contra incendios forestales. El proyecto involucró a más de 16 empresas y la participación activa de los gobiernos hacia la consecución de los siguientes objetivos: optimizar los recursos de los que dispone el Gobierno para reducir al mínimo el riesgo para los bosques, para mitigar los daños al medio ambiente en caso de incendio y para reducir el número y tamaño de los grandes incendios forestales, reduciendo al mínimo los

dispositivos de extinción de seguridad.

Siguiendo el proyecto *Prometeo*, aparecen en 2013 y 2015 los llamados proyectos *Lumes* y *Enjambre*, respectivamente, en los que nuevamente participan diversas empresas de los sectores público y privado. El principal objetivo de estos proyectos es el desarrollo de nuevas tecnologías avanzadas para la lucha integral contra los grandes incendios forestales, lo que permite reducir el número de éstos y la superficie, y la generación de un recinto de seguridad en las operaciones que reducen significativamente la tasa de accidentes de los participantes (técnicos, bomberos y pilotos).

Diversas tareas son de gran interés para ejecutar proyectos de esta magnitud y características. Estas tareas van desde el procesamiento de imágenes que proporcionan información sobre la estructura de la vegetación y de la evolución del fuego, el estudio de la viabilidad de los aviones no tripulados en este contexto, los algoritmos y estrategias para garantizar la seguridad de las brigadas terrestres, el análisis de las operaciones nocturnas de extinción y la coordinación del tráfico aéreo.

En este trabajo, llevamos a cabo una investigación en el marco descrito, centrada en la coordinación del tráfico aéreo. En situaciones de grandes incendios forestales, donde se encuentran diferentes puntos de carga y descarga de agua y la presencia de hasta 50 aviones de trabajo a la vez (algunos de ellos procedentes de otras administraciones), aumenta el tráfico aéreo en gran cantidad. Esto significa que se pueden producir situaciones en las que existe riesgo de colisión entre diferentes aeronaves (por ejemplo, en casos en los que dos circuitos de noria comparten la misma toma de agua, pero que tienen diferentes puntos de descarga), o situaciones en las que algunos aviones están en espera de órdenes, o es probable que se roce el tiempo máximo de trabajo. Por lo tanto, la presencia de una coordinación del tráfico aéreo que rige cada pedido de aviones, control de la noria que se asigna, su tiempo de trabajo, y el riesgo de colisión entre aeronaves, es necesario. Inicialmente, nos concentramos en dos tareas que son la asignación de aeronaves a las norias y la asignación de aeronaves a las bases de repostaje. Además, programamos y resolvemos los modelos correspondientes con el software libre *R* (cf. http://www.r-project.org/) y mostramos algunos de los resultados numéricos y computacionales obtenidos.

El objetivo es automatizar estas tareas, con el fin de ganar tiempo y ahorrar costes. Todo esto se realiza utilizando modelos y técnicas tomadas de la investigación operativa. La idea nace de una empresa en el sector de la seguridad y el enfoque es integrar esta herramienta en una aplicación más amplia que incluye diversas utilidades.

Cabe destacar que el problema de asignación es uno de los problemas clásicos de la programación lineal, habiendo aparecido con el trabajo de Votaw y Orden (1952) y se hizo más relevante, con la publicación del método húngaro para resolverlo (cf. Kuhn, 1955). Una revisión reciente de este problema y sus generalizaciones se debe a Pentico (2007). Vale la pena mencionar que el primer modelo que consideraremos está cerca del problema de asignación en tres dimensiones (cf. Geetha y Vartak, 1994) porque tratamos de asignar los aviones a las norias, que se encuentran a su vez asignadas a los diferentes frentes de fuego. Sin embargo, una situación un poco más general surge aquí porque se consideran varios objetivos simultáneamente (cf. Geetha y Nair, 1993), ya que se trata de maximizar la descarga de agua en los frentes y minimizar las distancias entre bases de aeronaves y norias, a la vez que se introducen restricciones específicas en los aviones y los puntos de agua relacionados con la capacidad, así como las preferencias en el porcentaje de agua recibida en los frentes.

Por otro lado, hay que destacar la existencia de diferentes referencias en la literatura relacionadas con los sistemas de apoyo a las decisiones en la lucha contra incendios, y de revisiones sobre el empleo de técnicas de investigación operativa en la lucha contra incendios en determinadas situaciones de desastre. Por ejemplo, Keramitsoglou et al. (2004) utilizan algoritmo de Dijkstra con el fin de optimizar las operaciones que involucran rutas de unidades de respuesta enviadas a atacar frentes de fuego desde varios puntos. Otros trabajos son Dimopoulou y Giannikos (2004), Galindo y Batta (2013), Mavsar et al. (2013), y Minas et al. (2012).

2. SOBRE LOS MODELOS

En la actualidad, el número de recursos aéreos que pueden coincidir en el tiempo y en el espacio en un incendio forestal ha aumentado considerablemente, en comparación con lo que ocurría en un pasado relativamente reciente, cuando el número de aeronaves no era muy grande, tenían poca autonomía y pocas bases de repostaje asociadas. En general, la organización del ataque aéreo no planteaba complicaciones y se gestionaba entre los propios pilotos de aviones. Hoy en día, no tener una función de coordinación de aire adecuada aumentaría sustancialmente el riesgo para la seguridad de vuelo y supondría una reducción grave en la eficacia.

2.1. LA ASIGNACIÓN DE AERONAVES A LAS NORIAS

Es muy común en un incendio forestal, que el trabajo de las aeronaves se organice a través de las llamadas norias, de forma que un conjunto de aeronaves está volando sobre el fuego formando un circuito desde el que cada aeronave accede a un punto de toma de agua para posteriormente realizar la descarga sobre un frente del incendio. Naturalmente, si estas operaciones se realizan de modo desorganizado nos encontramos ante una operación de alto riesgo, ya que si no está claro el estado de todas las aeronaves de forma continua, la posibilidad de colisión es alta. Las ventajas de este tipo de trabajo a través de las norias son principalmente: una ocupación más corta del espacio aéreo en el entorno del fuego, lo que significa que, además de no perder el tiempo las aeronaves, se hace el máximo daño al fuego en el menor tiempo posible, se consigue una mayor resistencia en zonas de altas llamas, y se consigue la posibilidad de líneas de defensa más largas y más precisas.

Si suponemos que previamente se ha determinado el número y tipo de aeronaves para ser enviado al incendio, un problema para el coordinador es distribuir de manera óptima los recursos a su disposición. Queremos construir una función en lenguaje *R*, que proporciona una asignación óptima de los recursos y puede servir de apoyo a la toma de decisiones por el coordinador durante el incendio. Por lo tanto, vamos a proponer un modelo de programación lineal entera cuyo propósito es maximizar la producción por hora de aeronaves que están actuando en el fuego, es decir, el objetivo es maximizar la cantidad de agua descargada en el fuego. Todo esto se hace a través de una adecuada asignación de aeronaves entre las norias, sujetos a la restricción de que no dejamos frentes de fuego desatendidos, que el número máximo de aeronaves por noria sea respetado (este número se determina previamente), y que el porcentaje de agua para cada frente, elegido por el coordinador, también se respeta.

2.2. LA ASIGNACIÓN DE AERONAVES A LAS BASES DE REPOSTAJE

Cuando los aviones asignados a un fuego comienzan el periodo de descanso, tienen que realizar el reabastecimiento de combustible. La tarea de asignar estos aviones a bases de reabastecimiento de combustible no es sencilla, ya que hay que tener en cuenta diversos factores, como la hora de llegada a estas bases, la cantidad de combustible disponible en cada punto, etc. El modelo de programación lineal entera que hemos construido proporciona la asignación óptima de los aviones a los puntos de reabastecimiento de combustible, de tal manera que el tiempo total utilizado para todas las aeronaves en la operación es mínimo. El modelo tiene en cuenta los siguientes aspectos. En primer lugar, considera el número de aeronaves que pueden abastecerse de combustible al mismo tiempo en la misma base. Por ejemplo, un punto de reabastecimiento de combustible puede estar constituido por un camión cisterna en el medio de un área abierta, siendo único dicho camión cisterna y constar de una sola manguera; en este caso el suministro simultáneo de varias aeronaves se hace imposible. Los litros de combustible de cada base y la capacidad de combustible de cada aeronave también son relevantes. También podría suceder que una base de reabastecimiento de combustible esté disponible muy próxima al fuego, pero el combustible disponible allí es menos que el requerido por el total de los aviones. Esto causaría que no todos los aviones podrían ser asignados a la base, a pesar de la cercanía de la misma. Por otra parte, los aviones pueden preferir esperar a que otro avión complete su repostaje en una base y luego recibir el suministro, que tener que ir a una base más distante (perdiendo así más tiempo). Además, se pretende que, una vez que se encontró la asignación óptima, haya una advertencia de la nueva capacidad de los puntos de reabastecimiento de combustible, de modo que, si es necesario, se vuelva a cargar la base de combustible.

3. CONCLUSIONES

En el trabajo aquí resumido, presentamos dos modelos de programación lineal entera que solucionan dos problemas reales con los que se puede encontrar el coordinador de control de tráfico en un servicio de incendios forestales.

Se pueden solicitar a los autores los modelos matemáticos detallados, junto con ejemplos y resultados numéricos y un estudio de simulación con instancias grandes en el que se compara la ejecución utilizando diferentes *solvers* (*lpSolveAPI* y *Gurobi*) y métodos de resolución (de tipo lexicográfico, entre otros).

Para resolver estos problemas reales, aplicamos una metodología novedosa que constituye un enfoque innovador tomado de la investigación operativa. Estos modelos se programan y se resuelven exactamente utilizando una herramienta de software libre como *R*. En varios ejemplos totalmente inspirados en situaciones reales vemos que la resolución del modelo es rápida. Estas características indican, en nuestra opinión, que este material puede ser útil como una herramienta para apoyar la toma de decisiones realizada por los coordinadores del control de tráfico en este tipo de incendio. De hecho, junto con otras herramientas que nuestro equipo de trabajo ha creado, como por ejemplo los algoritmos anticolisión, este material se integrará adecuadamente en un sistema más complejo, completo y fácil de usar, de apoyo a la decisión, que también incorpora métodos modernos de tratamiento y procesado de imágenes .

En los modelos que hemos definido, se realiza la asignación de recursos a las norias y a los puntos de repostaje. Es importante tener en cuenta que una vez hecha esta asignación de acuerdo a ciertos objetivos y restricciones, las aeronaves han de integrarse eficazmente en la tarea de extinción y ahora vamos a tener la necesidad de tener en cuenta una asignación temporal que incluye los horarios de trabajo de pilotos, períodos de descanso, el tiempo de vuelo desde el suelo, etc. Actualmente, estamos trabajando en una asignación eficiente de este tiempo contemplando las diversas restricciones y calculando los costes asociados con el trabajo de extinción de acuerdo a la filosofía *C + NVC* que se menciona en la introducción de este trabajo.

AGRADECIMIENTOS

Los autores desean agradecer las interesantes propuestas, sugerencias y comentarios, así como el apoyo en materia computacional, debido a W. González-Manteiga, B. Pateiro-López, A. Riera-Álvarez, y J. L. Saiz-Díaz. Esta investigación ha recibido apoyo financiero del Ministerio de Ciencia e Innovación y el Ministerio de Economía y Competitividad a través de los proyectos MTM2011-27731-C03 y MTM2014-53395-C3-2-P, respectivamente, y de ITMATI, *Technological Institute of Industrial Mathematics*, Santiago de Compostela, a través de los proyectos *Lumes* y *Enjambre*.

REFERENCIAS

Dimopoulou, M. and Giannikos, I. (2004) Towards an integrated framework for forest fire control. European Journal of Operational Research 152, 476-486.

Galindo, G. and Batta, R. (2013) Review of recent developments in OR/MS research in disaster operations management. European Journal of Operational Research 230, 201-211.

Geetha, S. and Nair, K. P. K. (1993) A variation of the assignment problem. European Journal of Operational Research 68, 422-426.

Geetha, S. and Vartak, M. N. (1994) The three-dimensional bottleneck assignment problem with capacity constraints. European Journal of Operational Research 73, 562-568.

Gorte, J. K. and Gorte, R. W. (1979) Application of economic techniques to fire management -a status review and evaluation. General Technical Report INT-53. Ogden, UT, USDA Forest Service, Intermountain Forest and Range Experiment Station, 26 p.

Headley, R. (1916) Fire Suppression District 5. USDA Forest Service, Washington, 58 p.

Keramitsoglou, I., Kiranoudis, C. T., Sarimveis, H., and Sifakis, N. (2004) A multidisciplinary decision support system for forest fire crisis management. Environmental Management 33, 212-225.

Kuhn, H. W. (1955) The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 83-97. Republished in 2005.

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.

Minas, J. P., Hearne, J. W., and Handmer, J. W. (2012) A review of operations research methods applicable to wildfire management. International Journal of Wildland Fire 21, 189-196.

Pentico, D. W. (2007) Assignment problems: a golden anniversary survey. European Journal of Operational Research 176, 774-793.

Sparhawk, W. N. (1925) The use of liability ratings in planning forest fire protection. Journal of Agricultural Research 30, 693-792.

Votaw, D. F. and Orden, A. (1952) The personnel assignment problem. Symposium on Linear Inequalities and Programming. SCOOP (Scientific Computation of Optimum Programs Project) 10 US Air Force, 155-163.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Programación lineal sucesiva sin región de confianza

Ángel M. González-Rueda[1], Julio González-Díaz[1] e María P. Fernández de Córdoba[1]

[1]Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela.

RESUMEN

La programación lineal sucesiva (SLP, Successive Linear Programming) es un enfoque muy utilizado para resolver problemas complejos de optimización no lineal. Informalmente, consiste en un proceso iterativo en el cual, en cada iteración, se resuelve una aproximación lineal del problema original entorno a la solución obtenida en la iteración anterior. De esta forma, en cada iteración restringe su búsqueda a una cierta región de confianza entorno al punto donde se linealiza. En este trabajo estudiamos una modificación natural del SLP, que denominamos programación lineal sucesiva sin región de confianza (SLP-NTR, Successive Linear Programming with No Trust Region), donde a la hora de resolver las aproximaciones lineales en cada iteración, no se limita su región de búsqueda. Estudiamos en qué medida esta modificación del algoritmo conserva las propiedades teóricas del algoritmo SLP clásico. Desde un punto de vista práctico, cabe destacar que esta metodología ya ha sido aplicada en un problema de la vida real de optimización de redes de transporte de gas obteniendo resultados satisfactorios.

Palabras e frases chave: programación lineal sucesiva, programación no lineal entera mixta, redes de transporte de gas.

1. INTRODUCCIÓN

La *programación lineal sucesiva* SLP resuelve problemas no lineales usando sucesiones de problemas de programación lineal. Los métodos basados en SLP surgieron de manera natural antes de la década de los 60 aplicados a problemas reales, especialmente en la industria petrolera. La primera referencia publicada sobre el SLP es el trabajo (Griffith y otros, 1961) desarrollado en Shell Development Company.

Las principales ventajas del método SLP son que es un algoritmo fácil de implementar, las linealizaciones del problema son flexibles y es capaz de resolver problemas muy grandes, debido a que en cada iteración se resuelve un problema lineal que, incluso teniendo muchas variables, resulta sencillo de resolver. En la actualidad existen numerosos solvers que son capaces de resolver problemas de programación lineal muy grandes (millones de variables y restricciones) en apenas unos segundos. Las desventajas surgieron al investigar el SLP sobre experimentos computacionales. En algunos tipos de problemas, el algoritmo resulta lento por moverse siempre dentro de regiones de confianza que ralentizan la convergencia, e incluso en algunos casos, no se obtiene la convergencia.

Los algoritmos basados en SLP son especialmente eficientes en problemas no lineales con casi tantas restricciones como variables. Si además el óptimo es un vértice de la región factible (es decir, el número de restricciones activas en el punto es mayor o igual que el de variables), la convergencia del algoritmo SLP es cuadrática (véase (Palacios,1980)). En general, los modelos de refinería suelen ser de esta naturaleza y muchos de ellos han sido resueltos con éxito usando la filosofía SLP.

Una de las primeras formalizaciones del SLP viene descrita en (Palacios,1982) donde se define el algoritmo detalladamente y de forma general, y se analizan algunos aspectos teóricos sobre su convergencia. La linealización que consideran es la aproximación de Taylor de primer orden, por

lo que se requiere la diferenciabilidad de las funciones que definen el problema. A continuación definiremos formalmente la idea general de los algoritmos SLP.

Consideremos el problema de programación no lineal NP definido como

$$\begin{array}{rll} \text{minimizar} & f(x) & \\ \text{sujeto a} & g_i(x) \leq 0 & i \in I^m \\ & h_j(x) = 0 & j \in I^l \\ & l \leq x \leq u, & \end{array}$$

siendo $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ para $i \in I^m$ y $h_j : \mathbb{R}^n \to \mathbb{R}$ para $j \in I^l$ funciones continuamente diferenciables y $l, u \in \overline{\mathbb{R}}^n$ con $\overline{\mathbb{R}} = [-\infty, +\infty]$.

Definición 1. *Se define la aproximación de Taylor de primer orden* TLP$(\overline{x}, s)$ *de* NP *entorno a* $\overline{x}$ *con paso* $s \in \mathbb{R}^n$ *como el problema de programación lineal*

$$\begin{array}{rll} \textit{minimizar} & f(\overline{x}) + \nabla f(\overline{x}) \cdot (x - \overline{x}) & \\ \textit{sujeto a} & g_i(x) + \nabla g_i(\overline{x}) \cdot (x - \overline{x}) \leq 0, & i \in I^m, \\ & h_j(\overline{x}) + \nabla h_j(\overline{x}) \cdot (x - \overline{x}) = 0, & j \in I^l, \\ & l \leq x \leq u, & \\ & -s \leq x - \overline{x} \leq s. & \textit{(Región de confianza)} \end{array}$$

Denotaremos por TLP$(\overline{x})$ *a la aproximación de Taylor de primer orden de* NP *entorno a* $\overline{x}$ *en el que eliminamos las restricciones asociadas a la región de confianza.*

Obsérvese que $d = x - \overline{x}$ es la dirección en la que nos movemos partiendo de $\overline{x}$ y que la última restricción nos indica que el paso d está limitado por s, es decir, nos movemos dentro de una *región de confianza* definida entorno al punto $\overline{x}$.

El algoritmo SLP sigue el siguiente procedimiento: parte de un punto dado $x_0 \in \mathbb{R}^n$ (no necesariamente factible para NP) y una cota de paso $s \in \mathbb{R}^n$, y resuelve el problema TLP(x_0, s) obteniendo una solución x_1. El candidato x_1 se acepta si verifica cierto criterio y, en ese caso, se toma $x_0 = x_1$, se incrementa la cota de paso s y se repite el proceso. Si el candidato x_1 no cumple el criterio considerado, se rechaza y se vuelve a resolver TLP(x_0, s) pero actualizando la cota de paso s.

En general, la solución x_1 de TLP(x_0, s) es un punto factible para el problema linealizado, pero no para el problema NP. No obstante, se puede demostrar que, si $x_1 = x_0$, entonces x_1 es factible para NP y además se trata de un punto que verifica las condiciones de Karush-Khun-Tucker (KKT) de NP. Luego, tiene sentido tomar como *criterio de parada* la condición de que la distancia $\|x_k - x_{k-1}\|$ sea suficientemente pequeña.

Una debilidad que presenta el algoritmo anterior es que puede tener problemas de factibilidad, en el sentido de que el problema linealizado que se resuelve en cada iteración podría no ser factible. Para solventar dicho inconveniente, en (Zhang y otros, 1985) se introduce el algoritmo de *programación lineal sucesiva con penalización PSLP* (Penalty Successive Linear Programming). La idea consiste en modificar el problema no lineal, penalizando el incumplimiento de las restricciones no lineales en la función objetivo.

El PLSP resulta más robusto y eficiente que versiones anteriores del SLP y fue aplicado con éxito a problemas reales de gran tamaño de la industria química y petrolera. Por ejemplo, en (Baker y otros, 1985), se muestran los resultados obtenidos al resolver problemas reales no lineales de gran tamaño de la empresa Exxon con el algoritmo PSLP. Este algoritmo, al igual que otras versiones del SLP, es especialmente eficiente en aquellos problemas con casi tantas restricciones como variables.

Otra ventaja que presenta el PSLP es que se tiene un resultado teórico de convergencia, según el cual cualquier punto de acumulación de la sucesión dada por el algoritmo es un punto KKT del problema no lineal original.

Antes de detallar el pseudocódigo del algoritmo PSLP, es necesario introducir algunos conceptos.

Definición 2. *Dado el problema no lineal* NP*, se define el problema penalizado* PNP *con parámetro de penalización* μ *como*

$$\begin{aligned} \textit{minimizar} \quad & p(x) = f(x) + \mu \left[\sum_{i \in I^m} \text{máx}\{g_i(x), 0\} + \sum_{j \in I^l} |h_j(x)| \right] \\ \textit{sujeto a} \quad & l \leq x \leq u. \end{aligned}$$

Notemos que el problema penalizado PNP también se puede formular de forma equivalente como:

$$\begin{aligned} \text{minimizar} \quad & p(x) = f(x) + \mu \left[\sum_{i \in I^m} y_i + \sum_{j \in I^l} (p_j^+ + p_j^-) \right] & \\ \text{sujeto a} \quad & g_i(x) \leq y_i, & i \in I^m \\ & h_j(x) = p_j^+ - p_j^-, & j \in I^l \\ & l \leq x \leq u & \\ & -s \leq x \leq s & \text{(Región de confianza)} \\ & y_i, p_j^+, p_j^- \geq 0 & i \in I^m, j \in I^l. \end{aligned}$$

Obsérvese que se trata de un problema factible siempre que $l \leq u$. El parámetro de penalización μ puede considerarse como un vector $(\mu_i, \overline{\mu}_j)$ de $m + l$ variables de manera que podamos asociar a cada restricción una penalización distinta.

El siguiente resultado establece las relaciones existentes entre el problema no lineal original NP y el problema penalizado PNP que acabamos de describir.

Teorema 1. *Sea* NP *un problema no lineal y* PNP *el problema penalizado asociado con parámetro* μ*. Las siguientes afirmaciones son ciertas:*

1. *Si* $\overline{x}$ *es un punto KKT de* NP *con multiplicadores de Lagrange* (u_i, v_j) *para* $i \in I^m$*,* $j \in I^l$ *tales que* $\mu_i > |u_i|$ *para* $i \in I^m$ *y* $\overline{\mu}_j > |v_j|$ *para* $j \in I^l$*, entonces* $\overline{x}$ *es un punto KKT para el problema* PNP*.*

2. *Si* $\overline{x}$ *es un punto KKT para* PNP *y* $\overline{x}$ *es factible para* NP*, entonces* $\overline{x}$ *es un punto KKT de* NP*.*

La idea del algoritmo PSLP será resolver en cada iteración una linealización del problema penalizado PNP. Definamos esto formalmente.

Definición 3. *Dado el problema penalizado* PNP *con parámetro de penalización* μ*, se define la aproximación lineal de Taylor de primer orden de* PNP *entorno a un punto* $\overline{x}$ *con paso* s *como el problema* TLPP$(\overline{x}, s)$ *definido por*

$$\begin{aligned} \textit{minimizar} \quad & p_L(x) = \nabla f(\overline{x}) \cdot x + \mu \left[\sum_{i \in I^m} y_i + \sum_{j \in I^l} (p_j^+ + p_j^-) \right] & \\ \textit{sujeto a} \quad & g_i(\overline{x}) + \nabla g_i(\overline{x}) \cdot (x - \overline{x}) \leq y_i, & i \in I^m \\ & h_j(\overline{x}) + \nabla h_j(\overline{x}) \cdot (x - \overline{x}) = p_j^+ - p_j^-, & j \in I^l \\ & l \leq x \leq u & \\ & -s \leq x - \overline{x} \leq s & \textit{(Región de confianza)} \\ & y_i, p_j^+, p_j^- \geq 0 & i \in I^m, j \in I^l. \end{aligned}$$

El pseudocódigo del PSLP puede verse en la descripción del algoritmo 1 . Los valores habituales de los parámetros, así como la demostración del teorema anterior, pueden consultarse en (Bazaraa y otros, 2006).

La sucesión $\{x_k\}$ obtenida con el algoritmo PSLP no siempre es convergente, sin embargo tenemos asegurado que cualquier subsucesión convergente lo hace a un punto KKT del problema penalizado PNP. La prueba puede verse en (Zhang y otros, 1985).

Teorema 2. *Sea* $\{x_k\}$ *la sucesión obtenida al aplicar el algoritmo PSLP al problema* PNP. *Si el conjunto* $C = \{x \in \mathbb{R}^n : l \leq x \leq u, p(x) \leq p(x_0)\}$ *es acotado, entonces cualquier punto de acumulación de* $\{x_k\}$ *es un punto KKT de* PNP. *Si además el punto límite es factible para* NP, *entonces es un punto KKT de* NP.

Algoritmo 1 Penalty Successive Linear Programming (PSLP)

1: Inicializar $x_0 \in \mathbb{R}^n$ cumpliendo las restricciones de cota del problema, y s_0. Fijar los parámetros μ, $0 < \rho_0 < \rho_1 < \rho_2 < 1$, $0 < \beta < 1$ y α. Sea $k = 0$.
2: Resolver TLPP(x_k, s_k). Sea $\overline{x}_k$ la solución.
3: Calcular $\Delta p = p(x_k) - p(\overline{x}_k)$ y $\Delta p_L = p_L(x_k) - p_L(\overline{x}_k)$
4: **if** $\Delta p_L = 0$ **then**
5: STOP: Devolver $\overline{x}_k$
6: **else**
7: Calcular $r_k = \frac{\Delta p}{\Delta p_L}$
8: **end if**
9: **if** $r_k < \rho_0$ **then**
10: $s_{k+1} = \beta s_k$
11: **else**
12: $x_{k+1} = \overline{x}_k$ y actualizar $s_{k+1} = \begin{cases} \beta s_k & \text{si} \quad r_k < \rho_1 \\ \frac{s_k}{\beta} & \text{si} \quad r_k > \rho_2 \\ s_k & \text{en otro caso} \end{cases}$
13: **end if**
14: Tomar $s_{k+1} = \max\{s_{k+1}, \alpha\}$, $k = k + 1 \rightarrow$ ir a 2

2. Programación lineal sucesiva sin región de confianza

En esta sección definimos y estudiamos la *programación lineal sucesiva sin región de confianza SLP-NTR* (Successive Linear Programming with No Trust Region). Como su propio nombre indica, se trata de aplicar el método SLP a problemas no lineales pero sin acotar los pasos en la búsqueda de direcciones de mejora. La idea consiste en linealizar el problema entorno a un punto, resolverlo sin región de confianza, y repetir el proceso con la solución obtenida.

Cabe destacar que la motivación de este estudio surge fruto de un proyecto de colaboración entre la Universidad de Santiago de Compostela y la empresa Reganosa. El objetivo de dicho proyecto era la resolución de un problema relacionado con la optimización de redes de transporte de gas. Por la estructura y tamaño del problema parecía adecuado la utilización de metodologías SLP, pero la presencia de elementos controlables en la red (tales como válvulas de control de presión, estaciones de compresión,..) que exigían la introducción de variables binarias en el problema, motivó la definición del SLP-NTR. En este contexto, el carácter discreto de la región factible es incompatible con la definición de una región de confianza pequeña, lo que hace que las versiones de SLP clásicas no sean fáciles de adaptar a problemas discretos.

El algoritmo SLP-NTR no pretende competir con el PSLP introducido anteriormente. Lógicamente, en algunos casos, el algoritmo SLP-NTR presenta oscilaciones mientras que la región de confianza (acotación del paso) del PSLP evita ese inconveniente.

Con todo, el SLP-NTR presenta algunas ventajas con respecto al PSLP que consideramos interesantes y que pueden resultar útiles para abordar algunos tipos de problemas reales como el problema del transporte de gas en el que ya hemos observado resultados satisfactorios. Por ejemplo, es muy sencillo de implementar, no es necesario reescalar los parámetros de penalización ni las cotas de paso, y en algunos casos rastrea la región factible de una manera "más global"(puesto que no acotamos la región de búsqueda). Más interesante, como ya comentamos, es la facilidad de adaptación que presenta el SLP-NTR frente a problemas de programación no lineal con variables enteras o binarias, dando lugar a un algoritmo heurístico para la resolución de problemas de programación no lineal entera mixta.

El algoritmo SLP-NTR sigue el siguiente procedimiento: dado $\overline{x} \in \mathbb{R}^n$, considera una aproximación lineal $\mathrm{LP}(\overline{x})$ del problema no lineal NP, que por ejemplo podría ser $\mathrm{TLP}(\overline{x})$ (la aproximación de Taylor de primer orden de NP sin región de confianza). El algoritmo SLP-NTR parte de un punto $x_0 \in \mathbb{R}^n$ (no necesariamente factible para NP), y resuelve el problema $\mathrm{LP}(x_0)$ obteniendo como solución x_1. Si se verifica que la distancia entre x_1 y x_0 es suficientemente pequeña, el algoritmo para. En caso contrario, se vuelve a resolver el problema linealizado entorno a x_1. El pseudocódigo del SLP-NTR aparece descrito en la descripción del algoritmo 2.

Algoritmo 2 Successive Linear Programming with No Trust Region (SLP-NTR)

1: Inicializar $x_0 \in \mathbb{R}^n$. Sea $k = 0$.
2: Resolver LP (x_k). Sea $\overline{x}_k$ la solución obtenida.
3: **if** $\|\overline{x}_k - x_k\|_\infty < \varepsilon$ **then**
4: STOP: Devolver $\overline{x}_k$
5: **else**
6: $x_{k+1} = \overline{x}_k$, $k = k + 1 \rightarrow$ ir a 2
7: **end if**

Puesto que la implementación del SLP-NTR es muy sencilla y flexible, a la hora de enfrentarnos a cada problema particular, podemos buscar la estrategia más acertada para evitar las posibles infactibilidades del problema linealizado. Por ejemplo, en el problema de optimización de redes de gas que describiremos más adelante, basta con ampliar un poco las restricciones de cota descritas para las presiones y penalizarlas en la función objetivo, para obtener factibilidad. Si se prefiere asegurar la factibilidad en cada iteración de manera general, se puede implementar el *problema penalizado sin región de confianza* (PSLP-NTR), penalizando las restricciones no lineales en la función objetivo, para el cual todos nuestros resultados teóricos son perfectamente aplicables.

A continuación presentamos los resultados teóricos asociado al algoritmo SLP-NTR, para lo cual necesitamos previamente introducir una serie de conceptos.

Consideremos un problema de programación lineal LP dado por

$$\begin{array}{ll} \text{minimizar} & c \cdot x \\ \text{sujeto a} & a_i \cdot x \leq b_i, \quad i \in I^m \\ & q_j \cdot x = d_j, \quad j \in I^l. \end{array}$$

El problema LP viene completamente caracterizado por el *vector de parámetros*

$$\overline{\pi} = (c, a, b, q, d) \in \mathbb{R}^n \times (\mathbb{R}^n)^m \times \mathbb{R}^m \times (\mathbb{R}^n)^l \times \mathbb{R}^l.$$

Definición 4. *Se define aplicación $\mathcal{S} : \Pi \rightarrow \mathcal{P}(\mathbb{R}^n)$ que asigna a cada $\pi \in \Pi$ el conjunto factible $\mathcal{S}(\pi)$ del correspondiente problema de programación lineal. Análogamente, se define la aplicación $\mathcal{S}_o : \Pi \rightarrow \mathcal{P}(\mathbb{R}^n)$ que asigna a cada $\pi \in \Pi$ el conjunto de óptimos $\mathcal{S}_o(\pi)$ del correspondiente problema de programación lineal.*

Definición 5. *Sea $\overline{\pi} = (c, a, b, q, d) \in \mathrm{dom}(\mathcal{S}_o)$. Se dice que $\mathcal{S}$ es semicontinua inferiormente en $\overline{\pi}$ si, para cada conjunto abierto $\mathcal{U} \subset \mathbb{R}^n$ con $\mathcal{S}(\overline{\pi}) \cap \mathcal{U} \neq \emptyset$, existe un entorno $\mathcal{V}$ de $\overline{\pi}$ tal que, para cada $\pi \in \mathcal{V}$, $\mathcal{S}(\pi) \cap \mathcal{U} \neq \emptyset$.*

El siguiente teorema garantiza que, en el caso favorable de que el algoritmo SLP-NTR converja, la solución obtenida es, desde el punto de vista teórico, igual de buena que la que pueda proporcionarnos el algoritmo PSLP. Además, el hecho de que el SLP-NTR no se restrinja a movimientos locales puede llevar a un mejor rastreo de la región factible.

Teorema 3. *Sea $\{x_k\}$ la sucesión generada por SLP-NTR. Supongamos que $\{x_k\}$ converge a un punto $\overline{x}$ y que $\mathcal{S}$ es semicontinua inferiormente en el problema* $\mathrm{LP}(\overline{x})$*. Las siguientes afirmaciones son ciertas,*

1. *el punto $\overline{x}$ es solución de* $\mathrm{LP}(\overline{x})$*,*
2. *el punto $\overline{x}$ es factible para* NP*,*

3. si LP($\overline{x}$)=TLP($\overline{x}$)*, entonces* $\overline{x}$ *es un punto KKT de* NP.

Como ya mencionamos previamente, el hecho de no imponer región de confianza en el SLP-NTR puede provocar oscilaciones que hagan que la sucesión no converja. Por ello, una pregunta natural es si, aún en el caso de no converger, sigue teniendo sentido considerar como criterio de parada que la distancia entre puntos consecutivos de la sucesión sea pequeña. La respuesta a esta cuestión la analizamos a continuación, donde demostramos que ese criterio de parada nos garantiza encontrar un punto *casi-factible* y *casi-KKT* del problema no lineal original.

Consideremos un problema de programación no lineal NP definido como:

$$\begin{array}{rl} \text{minimizar} & f(x) \\ \text{sujeto a} & g_i(x) \leq 0, \quad i \in I^m \\ & h_j(x) = 0, \quad j \in I^l \\ & x \in K, \end{array}$$

donde K es un subconjunto compacto de $\mathbb{R}^n$, $f : K \to \mathbb{R}$, $g_i : K \to \mathbb{R}$ para $i \in I^m$ y $h_j : K \to \mathbb{R}$ para $j \in I^l$ son funciones continuamente diferenciables en K. Sea $S \subseteq K$ la región factible de NP.

Denotamos por $\mathcal{I}$ al conjunto de funciones infinitesimales siguiente:

$$\mathcal{I} = \{\, t : \mathbb{R}^+ \to \mathbb{R}^+ \text{ tal que } \lim_{x \to 0} t(x) = 0\}.$$

Definimos a continuación el concepto de punto *casi-factible* y de restricción de igualdad *casi-activa*.

Definición 6. *Sea* $x \in K$ *y* $\varepsilon > 0$*. Entonces,* x *es* ε*-factible si, para cada* $i \in I^m$ *y cada* $j \in I^l$*,*

$$g_i(x) \leq \varepsilon, \quad y \quad |h_j(x)| \leq \varepsilon.$$

Se define el conjunto de restricciones de desigualdad ε*-activas en* x*, como*

$$I_x^\varepsilon = \{i \in I^m : |g_i(x)| \leq \varepsilon\}$$

En el siguiente teorema probamos que si un punto x es casi-factible, hay un punto factible cercano $\overline{x}$ de manera que las restricciones casi activas de x son restricciones activas para $\overline{x}$. Para ello, necesitamos suponer que la región factible del problema no lineal está acotada. Es habitual en los problemas reales que las variables estén acotadas superior e inferiormente, por lo que dicha restricción no es muy exigente.

Teorema 4. *Dado el problema no lineal* NP*. Existe un número real* $M > 0$ *y una función infinitesimal* $\hat{t} \in \mathcal{I}$ *tales que, si* x *es un punto* ε*-factible para* $0 < \varepsilon < M$*, entonces existe un punto* $\overline{x} \in S$ *tal que*

a) $d(x, \overline{x}) < \hat{t}(\varepsilon)$*,*

b) $I_x^\varepsilon \subseteq I_{\overline{x}}$*.*

Para el siguiente resultado necesitamos introducir el concepto de *punto casi-KKT* que, intuitivamente, es un punto que está cerca de un punto factible $\overline{x}$ que casi verifica las condiciones de Karush-Khun-Tucker. En (Han y otros, 2010), (Haeser, 2010), (Haeser y otros, 2011), se introduce el concepto de punto "almost-KKT". Aunque las distintas definiciones propuestas son parecidas, no son exactamente iguales, y el contexto y la finalidad para la que se introduce esta noción son diferentes.

Definición 7. *Se dice que un punto* $x \in K$ *es un punto* $(\varepsilon_1, \varepsilon_2)$*-KKT del problema no lineal* NP *para* $\varepsilon_1, \varepsilon_2 \geq 0$ *si*

1. x *es* ε_1*-factible,*

2. para cada $i \in I_x^{\varepsilon_1}$*, existen* $u_i \geq 0$ *y, para cada* $j \in I^l$*, existen* $v_j \in \mathbb{R}$*, tales que*

$$\|\nabla f(x) + \sum_{i \in I_x^{\varepsilon_1}} u_i \nabla g_i(x) + \sum_{j \in I^l} v_j \nabla h_j(x)\| < \varepsilon_2.$$

El próximo teorema establece formalmente que si dos puntos de la sucesión generada por SLP-NTR están a distancia δ, entonces el segundo es un punto $(\varepsilon_1, \varepsilon_2)$-KKT, donde ε_1 y ε_2 tienden a cero cuando δ tiende a cero.

Teorema 5. *Sean g_i y h_j funciones $\mathcal{C}^2(K)$ y sea $\{x_k\}$ la sucesión generada por el algoritmo SLP-NTR aplicado al problema* NP. *Entonces, existen $t_1, t_2 \in \mathcal{I}$ tales que, si x_k verifica que $0 < \|x_k - x_{k-1}\| < \delta$, entonces x_k es un punto $(\varepsilon_1, \varepsilon_2)$-KKT de* NP *con $\varepsilon_1 = t_1(\delta)$ y $\varepsilon_2 = (1 + \sum_{i \in I^L_{x_k}} u_i + \sum_{j \in I^l} v_j) t_2(\delta)$, donde u_i y v_j son los multiplicadores de Lagrange de x_k para el problema lineal* TLP(x_{k-1}).

Como ya destacamos previamente, el SLP-NTR da lugar a un algoritmo heurístico para resolver problemas de programación no lineal con variables enteras. Para resolver dichos problemas es muy frecuente en la literatura emplear enfoques en dos etapas de forma que:

- Etapa 1: se resuelve una versión simplificada del problema original obteniendo un valor para las variables enteras.
- Etapa 2: se afina la solución resolviendo el problema continuo resultante (fijando las variables enteras con el valor obtenido en la etapa anterior).

De esta forma surge de manera muy natural el algoritmo en dos etapas que denominamos *2-SLP*, que básicamente consiste en aplicar en la etapa 1 el SLP-NTR y en la etapa 2 el PSLP, de manera que el primero proporciona una configuración óptima de las variables discretas y a continuación el segundo afina el valor de las variables continuas restantes. Este es el enfoque que empleamos en la práctica para resolver el problema de transporte de gas obteniendo resultados satisfactorios.

Debemos tener en cuenta que como el SLP-NTR es menos estable en cuanto a convergencia que el PSLP, tiene sentido aplicar el algoritmo 2-SLP incluso en el caso de que tengamos un problema no lineal continuo (sin variables discretas). Una buena estrategia sería lanzar en primer lugar el SLP-NTR y, en caso de que no converja, aplicar en PSLP para afinar la solución anterior.

Las principales características que podemos destacar del enfoque 2-SLP son las siguientes. En primer lugar, en la etapa 1 resolvemos el problema original sin necesidad de considerar una versión simplificada del modelo. Además, tenemos una serie de resultados teóricos que avalan la utilización del SLP-NTR en la primera etapa. Desde el punto de vista aplicado, para el problema de optimización de la red de transporte de gas española, hemos observado un buen comportamiento del algoritmo obteniendo reducciones significativas de los costes comparándolos con los esquemas de funcionamiento reportados por el operador de transmisión del sistema (Transmission System Operator).

Para finalizar esta sección, presentaremos los resultados computacionales obtenidos donde se llevó a cabo una comparación de los 3 algoritmos: SLP-NTR, PSLP y 2-SLP. Para realizar este estudio hemos aplicado los algoritmos al problema no lineal consistente en la optimización de redes de transporte de gas que explicamos a continuación.

Una red de transporte de gas consiste básicamente en un número controlable de elementos (puntos de suministro, puntos de consumo, estaciones de compresión, válvulas de control, etc...) conectados mediante tuberías. A medida que el flujo de gas avanza por las tuberías, la presión del mismo va disminuyendo debido a la fricción que se produce contra la pared de las mismas. La pérdida de presión hace que sea más difícil garantizar la seguridad de suministro, esto es, satisfacer la demanda de los puntos de consumo con el gas proveniente de los puntos de abastecimiento, respetando los límites de presión permitidos por la propia configuración de la red.

Por lo tanto, es necesario emplear estaciones de compresión para contrarrestar esa pérdida de presión en el flujo de gas. De esta forma, dichas estaciones aumentan la presión del gas que circula por ellas facilitando su circulación y que alcance los puntos de consumo con la presión requerida. No obstante, para realizar ese proceso es necesario el consumo de una parte del gas que entra en la estación de compresión, de manera que supone un coste considerable para la gestión de la red. Es por lo tanto importante gestionar la red de manera óptima y eficiente, para reducir el autoconsumo de gas en las estaciones de compresión.

Si modelamos la red de transporte de gas mediante un grafo dirigido $\mathcal{G} = (N, A)$, donde N representa un conjunto finito de nodos y E un conjunto finito de aristas, este problema puede ser expresado matemáticamente de la siguiente forma:

$$\text{minimizar} \quad \sum_{k \in A^c} g_k$$

sujeto a

$$(1) \quad \underline{p}_i \leq p_i \leq \bar{p}_i \qquad \forall i \in N$$

$$(2) \quad \underline{q}_k \leq q_k \leq \bar{q}_k \qquad \forall k \in A$$

$$(3) \quad \sum_{k \in A_i^{ini}} q_k - \sum_{k \in A_i^{fin}} q_k = c_i \qquad \forall i \in N^c$$

$$(4) \quad 0 \leq \sum_{k \in A_i^{ini}} q_k - \sum_{k \in A_i^{fin}} q_k \leq c_i \qquad \forall i \in N^s$$

$$(5) \quad p_i^2 - p_j^2 = K_1 \frac{L_k}{D_k^5} |q_k| q_k + K_2 (p_i^2 + p_j^2)(h_j - h_i) \qquad \forall k = (i,j) \in A$$

$$(6) \quad g_k = K_3 \left(\left(\frac{p_j^2}{p_i^2} \right)^{\frac{\gamma - 1}{\gamma}} - 1 \right) q_k \qquad \forall k = (i,j) \in A^c$$

Esto nos lleva a la formulación de un problema de programación matemática no lineal (no convexo) donde las principales variables son: las presiones en los nodos, p_i, y el caudal en cada arista, q_k. La función objetivo será minimizar el gas consumido en las estaciones de compresión. En cuanto a las restricciones, se pueden dividir en cuatro tipos:

- Las restricciones de cota sobre las presiones y los caudales del gas, dadas por (1) y (2). En este caso, $\underline{p}_i$ y $\bar{p}_i$ son las cotas inferiores y superiores para las presiones, y $\underline{q}_k$ y $\bar{q}_k$ las cotas inferiores y superiores para los caudales.
- Las restricciones de conservación de flujo, dadas por (3) y (4), habituales en todos los problemas de flujo en redes. En cuanto al parámetro c_i, en la ecuación (3) denota la cantidad de gas requerida por el punto de consumo i, y en la ecuación (4) denota la capacidad máxima del punto de suministro i.
- Las restricciones físicas de pérdida de presión, dadas por (5). Dicha ecuación está relacionada con la longitud de la tubería, L_k, con el diámetro de la misma, D_k, así como la altura de su nodos inicial y final, h_i y h_j. Además aparecen dos constantes K_1 y K_2, que dependen de las propiedades del gas y de cada tubería.
- Las restricciones para calcular el consumo producido por cada compresor de la red, que aparece en (6) denotado por g_k. Como podemos observar, dicho consumo es proporcional al ratio de compresión ($\frac{p_j}{p_i}$) y al caudal (q_k).

Una descripción detallada de la formulación matemática del problema puede verse en (Bermúdez y otros, 2015).

Para llevar a cabo los experimentos computacionales hemos considerado 75 escenarios distintos sobre la red de transporte de gas española. La diferencia entre los distintos escenarios reside en la distribución de las demandas. Por ejemplo, los escenarios con bajas demandas se corresponden con días estivales, mientras que aquellos con altas demandas se dan en los días de invierno. Nuestro modelado de la red de gas española consta de 485 nodos y 623 aristas, con lo que tendremos un problema de optimización no lineal con más de 1000 variables y más de 500 restricciones no lineales. En la figura 1 podemos ver la red de transporte de gas de alta presión española que se ha considerado.



Figura 1: Red de transporte de gas española.

En la figura 2 comparamos las funciones objetivo obtenida por los tres algoritmos para los 75 escenarios. En concreto, para cada escenario se presenta el porcentaje en el que empeora la función objetivo de cada algoritmo con respecto a la mejor obtenida. Por lo tanto, valores próximos a cero indican que el algoritmo encuentra mejores funciones objetivo. Con una línea vertical azul señalamos los 26 escenarios para los que el SLP-NTR no ha convergido (un poco más de la tercera parte del total de escenarios) y por lo tanto podría ser que la solución obtenida por este algoritmo no fuese físicamente factible y carecer de sentido la comparación. Cabe destacar que tanto el PSLP como el 2-SLP han convergido para todos los escenarios. Volviendo a las funciones objetivo, se observa que en la mayoría de los escenarios los tres algoritmos encuentran prácticamente la misma solución. Además, se observa que en muchas situaciones donde el SLP-NTR no ha convergido el algoritmo 2-SLP no solo obtiene mejor función objetivo que el PSLP, sino que la mejora considerablemente en cuanto al orden de magnitud de mejora. Entrando en detalle, podemos decir que en 24 escenarios el 2-SLP obtiene mejor función objetivo que el PSLP y en 18 de ellos la mejora en las dos primera cifras decimales.

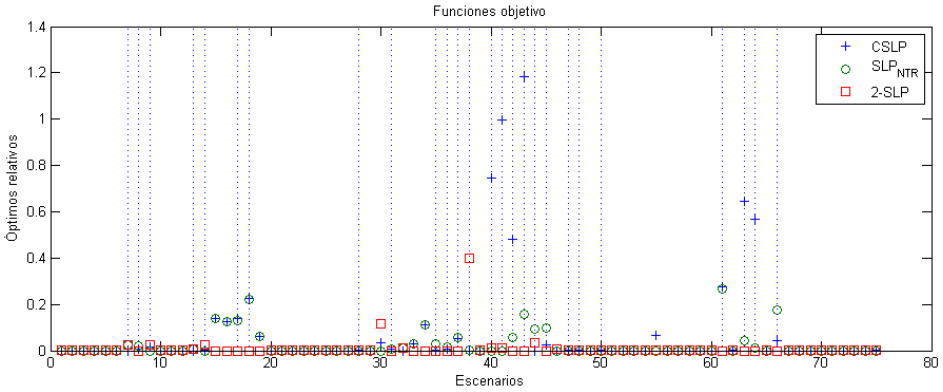


Figura 2: Comparación relativa de funciones objetivo.

Por último, en la figura 3 presentamos la función de densidad de los tiempos relativos de computación de los algoritmos. De forma análoga a las funciones objetivo, se representa la función de densidad de los porcentajes en los que se incrementa el tiempo de computación con respecto al algoritmo más rápido. Por lo tanto, valores próximos a cero indican que el algoritmo es más rápido. A la vista del gráfico se observa claramente que los algoritmos más rápidos son el SLP-NTR y el 2-SLP. En concreto, hay 61 escenarios ($\approx 81\,\%$ de los escenarios) donde el algoritmo SLP-NTR es más rápido que el PSLP, y 44 escenarios ($\approx 59\,\%$ de los escenarios) donde el 2-SLP es más rápido que el PSLP. Esto sugiere que una buena estrategia para enfrentarse al problema es, primero lanzar el algoritmo SLP-NTR, y después lanzar el PSLP para refinar la solución y asegurar la convergencia. Esta segunda etapa puede evitarse cuando la solución que proporciona el SLP-NTR es satisfactoria (buena convergencia, por ejemplo).

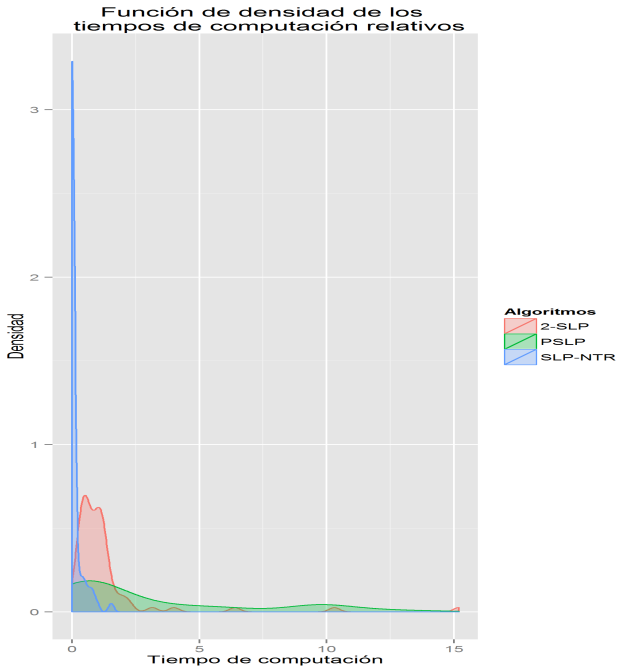


Figura 3: Comparación relativa de los tiempo de computación.

3. CONCLUSIONES

En este trabajo se ha introducido un nuevo algoritmo denominado programación lineal sucesiva sin región de confianza (SLP-NTR) que surge de una modificación natural del algoritmo clásico de programación lineal sucesiva. La motivación principal de este estudio surge fruto de la necesidad de resolver el problema de optimización de la red transporte de gas española que planteó la empresa Reganosa S.A. en un proyecto de investigación con la Universidad de Santiago de Compostela. Aunque en la práctica se había observado un buen comportamiento del algoritmo, se trató de buscar un fundamento teórico viendo en qué medida se conservan las propiedades teóricas de los algoritmos de programación lineal sucesiva clásicos.

En cuanto a los resultados teóricos, hemos demostrado que cuando el algoritmo SLP-NTR converge se obtiene un punto KKT del problema no lineal original, conservando las propiedades de los algoritmos clásicos. Por otro lado, debido a la eliminación de la región de confianza, el SLP-NTR presenta menos estabilidad en cuanto a convergencia. En este caso, cuando el algoritmo no converge, podemos garantizar que si dos puntos de la sucesión generada por el algoritmo están suficientemente próximos, entonces están cerca de un punto KKT del problema original. Además, se propone el algoritmo en dos etapas denominado 2-SLP que combina el SLP-NTR y el PSLP, que evita los posibles problemas de convergencia.

A la vista de las pruebas computacionales sobre la red de transporte de gas española, vemos que los resultados obtenidos tanto por el SLP-NTR como el 2-SLP son realmente satisfactorios

y competitivos con los obtenidos por el algoritmo clásico PSLP. De hecho, teniendo en cuenta tanto la convergencia como la función objetivo y los tiempos de computación, consideramos que la elección del 2-SLP a la hora de abordar un problema no lineal puede ser una buena estrategia. Si al lanzar el SLP-NTR obtenemos una solución aceptable, ya no es necesario lanzar el PSLP, y ahorramos el esfuerzo de reescalar el parámetro de penalización y el manejo de las cotas de la región de confianza. En el caso de que la solución obtenida con el SLP-NTR no nos resulta aceptable, siempre podemos lanzar en una segunda etapa el PSLP.

Por último cabe destacar la versatilidad del SLP-NTR para poder resolver problemas que incluyan variables enteras, surgiendo de forma muy natural el algoritmo en dos etapas 2-SLP como un heurístico para la resolución de problemas de programación no lineal entera mixta. En la primera etapa el SLP-NTR se encarga de buscar un valor óptimo para las decisiones discretas, y en la segunda etapa el PSLP resuelve el problema continuo resultante afinando la solución.

REFERENCIAS

Baker, T. E. and Lasdon, L. S. (1985) Successive linear programming at exxon. Management Science 31 , 264–274.

Bazaraa, M. S., Sherali, H. D. and Shetty, C. M. (2006) Nonlinear programming: Theory and algorithms. John Wiley and Sons.

Bermúdez, A., González-Díaz, J., González-Diéguez, F. J. González-Rueda, A. M. and Fernández de Córdoba, M. P. (2015) Simulation and optimization models of steady-state gas transmission networks. Energy Procedia 64, 130–139.

Griffith, R. and Stewart, R. A. (1961) A nonlinear programming technique for the optimization of continuous processing systems. Management Science 7, 379–392.

Haeser, G. (2010) On the global convergence of interior-point nonlinear programming algorithms. Computational & Applied Mathematics 29, 125–138.

Haeser, G. and Schuverdt, M. L. (2011) On approximate kkt condition and its extension to continuous variational inequalities. Journal of Optimization Theory and Applications 149, 528–539.

Han, D., Jian, J. and Xu, Q. (2010) A norm-relaxed method of quasi-multiplier-strongly subfeasible direction for general constrained optimization international conference on computer and computational intelligence 2010. Proceedings of the International Conference on Computer and Computational Intelligence.

Palacios-Gómez, F. and Lasdon, L. and Engquist, M. (1982) Nonlinear optimization by successive linear programming source. Management Science 10 , 1106–1120.

Palacios-Gómez, F.E. (1980) The solution of nonlinear optimization problems using successive linear programming: theoretical and computational results. University of Texas at Austin.

Zhang, J., Kim, N. H. and Lasdon, L. S. (1985) An improved successive linear programming algorithm. Management Science 31(10) , 1312–1331.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

OPTIMIZACIÓN DE LA GESTIÓN FORESTAL DE RODALES REGULARES

Manuel Arias-Rodil[1], Ulises Diéguez-Aranda[1] y Miguel E. Vázquez-Méndez[2]

[1]Departamento de Enxeñaría Agroforestal, Universidade de Santiago de Compostela. Escola Politécnica Superior, R/ Benigno Ledo, 27002 Lugo (Spain).

[2]Departamento de Matemática Aplicada, Universidade de Santiago de Compostela. Escola Politécnica Superior, R/ Benigno Ledo, 27002 Lugo (Spain).

RESUMEN

En este trabajo se propone un modelo para la optimización de la gestión forestal de rodales regurales monoespecíficos. Dicho modelo está compuesto por una serie de variables de estado, dadas por funciones inicialmente regulares (de clase C^∞), que dejan de serlo debido a las discontinuidades que provocan tratamientos selvícolas instantáneos. Pese a ello, el modelo propuesto sigue siendo diferenciable con respecto a las variables de decisión (instantes, tipos e intensidades de los tratamientos selvícolas, y edad de la corta final). Eso permite formular el problema de la gestión óptima de un rodal regular como un problema de optimización diferenciable con restricciones lineales, posibilitando el uso de algoritmos de optimización con derivadas, que resultan muy eficientes.

Palabras y frases clave: modelo dinámico, simulación forestal, programa selvícola, optimización diferenciable.

1. INTRODUCCIÓN

La predicción del crecimiento y la producción de las masas forestales y su respuesta a las actuaciones selvícolas han sido objetivos fundamentales de la investigación forestal, ya que la clave para una correcta gestión de los montes radica en un profundo conocimiento de los procesos de desarrollo de las especies arbóreas. Para tal fin, se han construido modelos matemáticos de diferentes características que, implementados en programas informáticos, permiten simular los resultados de diferentes programas selvícolas y evaluar así las consecuencias que tendrían sobre el sistema forestal (Diéguez-Aranda *et al.*, 2009). Un programa selvícola define una estrategia sobre las prácticas selvícolas que se pueden realizar en una masa forestal para alcanzar un determinado fin: preparación del suelo, plantación, fertilización, podas, claras (corta de determinados árboles para favorecer el desarrollo de los pies remanentes), etc.

La gestión de cualquier recurso requiere una fase de planificación que conlleva tomar decisiones. En el caso forestal, se pueden considerar al menos cuatro niveles: árbol, rodal, monte y paisaje (Bettinger *et al.*, 2009). Un rodal es un espacio forestal con características de estación (clima, suelo y fisiografía) y de masa homogéneas, mientras que un monte puede considerarse como un agregado de rodales. El nivel de rodal es el primero con verdadero significado en la toma de decisiones, ya que el crecimiento de un árbol está condicionado por el de los que lo rodean, y consideraciones logísticas y económicas hacen inviable casi siempre centrar la gestión en árboles concretos, excepto en aquellos de gran valor. En ese primer nivel, la planificación de cada rodal se realiza de forma individual, buscando que se alcancen los objetivos de su dueño (Clutter *et al.*, 1983), siendo la forma más lógica de gestionar propiedades forestales de pequeña extensión. Sin embargo, cuando se considera una perspectiva más amplia como el nivel de monte o el de paisaje, hay que tener en cuenta los posibles sacrificios que es preciso cometer a nivel de rodal para alcanzar uno o varios objetivos más globales (e.g., estabilidad en las cortas, en los ingresos o en las superficies de regeneración de un monte, o restricciones espaciales de los aprovechamientos).

Los rodales coetáneos o regulares, aquellos con árboles de la misma edad o clase de edad, se suelen gestionar mediante sistemas de turno de corta (RFM, del inglés *Rotation Forest Management*), en los que se aplican repetidamente ciclos selvícolas conformados por varias fases: una primera de renovación mediante plantación o regeneración natural, una segunda en la que se pueden realizar una o varias claras y diversos tratamientos culturales de mejora, y una tercera de aprovechamiento final cuando se llega al turno o edad de corta del rodal (Diéguez-Aranda *et al.*, 2009). En ese tipo de rodales, los modelos matemáticos más utilizados para la planificación de su gestión son empíricos, predictivos y generalmente determinísticos, proporcionando los valores de las principales variables del rodal en cualquier instante. Dado que el sistema forestal es dinámico, los modelos de rodal más avanzados se han desarrollado basándose en el enfoque del espacio de estados, cuyo principio debe buscarse en la teoría matemática de sistemas (García, 1994).

La gestión a nivel rodal se puede optimizar utilizando un simulador de crecimiento forestal (un programa informático que incluye el conjunto de funciones interrelacionadas y algoritmos que imitan el desarrollo del rodal) y un algoritmo de optimización. El simulador proporciona los resultados de los posibles programas de gestión, que se definen por el número, tipo, tiempo e intensidad de las operaciones selvícolas, así como por la edad de la corta final. Las alternativas de gestión se priorizan según diferentes criterios, dependiendo de los intereses del propietario forestal (e.g., crecimiento medio en volumen o beneficio económico).

Dado que el número de opciones en un problema de gestión es ilimitado, es necesario utilizar algún procedimiento de búsqueda automática (el algoritmo de optimización) para generar y evaluar alternativas con respecto al valor de la función objetivo especificada. En optimización a nivel rodal se ha empleado una gran variedad de algoritmos de optimización, que se diferencian principalmente en si requieren o no la discretización de las variables de decisión, reduciendo en caso afirmativo el número de soluciones a evaluar.

En los primeros, la búsqueda se reduce a la red de nodos definidos en un espacio discreto de soluciones, dentro del cual el algoritmo alcanza el máximo global. El problema radica en lograr un equilibrio entre la pérdida de información debida a la discretización y el tiempo de computación, que aumenta exponencialmente con el número de variables de decisión (Valsta, 1993). Los métodos de este tipo utilizados en optimización a nivel rodal han sido programación dinámica (e.g. Arimizu, 1958; Amidon y Akin, 1968; Díaz-Balteiro y Rodríguez, 2006) y búsqueda en profundidad (Arias-Rodil *et al.*, 2015a), el primero de los cuales requiere además discretizar las variables de estado en cada etapa.

Los métodos que no requieren una discretización de las variables de decisión que se han empleado en optimización a nivel rodal han sido fundamentalmente métodos de búsqueda directa (e.g., Cao, 2010; Pukkala y Kellomäki, 2012; Tahvonen *et al.*, 2013), y presentan el inconveniente de que no garantizan alcanzar el óptimo global (Valsta, 1993), aunque suelen proporcionar resultados razonablemente buenos en tiempos de computación reducidos y no requieren el uso de derivadas. Dentro de este grupo, se puede diferenciar entre los métodos que usan un único vector solución y aquellos que utilizan varios vectores solución, los llamados métodos basados en poblaciones (Pukkala, 2009).

En este trabajo se analiza la optimización de la gestión de rodales regurales de especies forestales para las que se ha desarrollado un modelo dinámico de crecimiento. Dicho modelo está compuesto por una serie de variables de estado, dadas por funciones inicialmente regulares (de clase C^{∞}), que dejan de serlo cuando se simulan tratamientos selvícolas. Como es bien sabido, estos tratamientos suponen, en los instantes en los que se aplican, cambios bruscos en los valores de las variables de estado, provocando que éstas presenten discontinuidades en esos instantes. Pese a ello, y tal y como se mostrará en la sección 3, el modelo sigue siendo diferenciable con respecto a las variables de decisión (instantes, tipos e intensidades de los tratamientos selvícolas, y edad de la corta final). Eso nos permite, tal y como haremos en la sección 4, formular el problema de la gestión óptima de un rodal regular como un problema de optimización diferenciable con restricciones lineales. Esta nueva formulación abre un enorme abanico de posibilidades a la hora de resolver el problema, ya que, como es bien sabido, el uso de algoritmos de optimización con derivadas resultan

muy eficientes en este tipo de problemas. Finalizaremos el trabajo, sección 5, presentando algunas discusiones sobre esos temas.

2. SISTEMAS DINÁMICOS

Un modelo de crecimiento predice valores futuros de ciertos *outputs* o salidas (e.g., el volumen de madera) a partir de ciertos *inputs* o entradas (e.g., el instante, tipo e intensidad de las claras). La idea de los sistemas dinámicos es caracterizar el estado del sistema de tal modo que, dado el estado actual, el futuro no dependa del pasado. A grandes rasgos, el estado del sistema en un instante dado es el conjunto de variables que recogen la información necesaria para determinar el comportamiento del sistema de ese instante en adelante (García, 1994). Por ejemplo, en rodales regulares monoespecíficos sometidos a diferentes tratamientos selvícolas, el estado puede caracterizarse por su altura dominante, el número de árboles por hectárea y el área basimétrica, y cualquiera otra variable de interés (salida) puede estimarse a partir de esas tres. Se asume entonces que dos rodales con los mismos valores para esas tres variables se comportarán esencialmente de la misma forma, independientemente de cómo han alcanzado ese estado.

Teniendo en cuenta que los tratamientos selvícolas normalmente ocurren en instantes determinados y provocan cambios de estado instantáneos, podemos asumir que las entradas (de ahora en adelante variables de control) del sistema vienen dadas por un vector $\mathbf{u}$ y denotamos por $\mathbf{x}(t)$ la función vectorial cuyas componentes son el conjunto de variables de estado del sistema y por $\mathbf{y}(t)$ la función vectorial que recoge las salidas del sistema. Atendiendo a la estructura general del enfoque del espacio de estados apuntada por García (1994), el comportamiento de un sistema queda determinado por dos funciones:

1. La función de transición $\mathbf{F}$ que verifica que

$$\mathbf{x}(t) = \mathbf{F}[\mathbf{x}(t_0), \mathbf{u}, t - t_0], \ \ \forall t \geq 0, \tag{1}$$

2. La función de salida $\mathbf{g}$ que cumple

$$\mathbf{y}(t) = \mathbf{g}[\mathbf{x}(t)], \ \ \forall t \geq 0. \tag{2}$$

La función de transición (1) proporciona el estado el cualquier instante t en función del estado en algún otro instante t_0, de las variables de control ($\mathbf{u}$) y del tiempo transcurrido entre t_0 y t, mientras que la función de salida (2) proporciona los valores de cualquier otra variable de interés (salida) a partir de las variables de estado.

La función de transición se obtiene, en muchos casos, al resolver el sistema de ecuaciones diferenciales que describe la *tasa de cambio* de las variables de estado y, en cualquier caso, el uso de funciones de transición y salida es muy frecuente en la modelización del crecimiento y la producción en el campo forestal. En la siguiente sección vamos a proponer un modelo dinámico general para el crecimiento de rodales regulares. Mostraremos como, en ese caso, la función de transición (y la de salida) es diferenciable con respecto a las variables de control y daremos una expresión explícita de la salida en función de las variables de control, lo que resulta de gran utilidad para calcular sus derivadas a la hora de buscar la gestión óptima del rodal.

3. MODELO DINÁMICO DE CRECIMIENTO PARA RODALES REGULARES

Según García (1994), un modelo suficientemente general para rodales coetáneos o regulares monoespecíficos, en los que se llevan a cabo podas y claras de diferente tipo e intensidad (siempre que no sean muy intensas), y para los que se quiere predecir el volumen de madera clasificado según diferentes destinos comerciales e incluso parámetros de distribuciones diamétricas, considera un vector de estado que incluye las variables altura dominante ($H(t)$, definida como la altura media de los árboles del estrato dominante), número de árboles por hectárea ($N(t)$) y área basimétrica

($G(t)$, suma de la sección normal –a 1,3 m del suelo– del tronco de todos los árboles que hay en una hectárea) (siguiendo la notación introducida en la sección anterior, $\mathbf{x}(t) = (H(t), N(t), G(t))$). En ese modelo, las variables de estado vienen dadas a partir de unas funciones conocidas, $h(t_k, H_k, t)$, $n(t_k, N_k, t)$ y $g(t_k, G_k, t)$, de manera que, si para algún $t_0 \in [a, b]$ se conocen los valores de $H(t_0) = H_0$, $N(t_0) = N_0$ y $G(t_0) = G_0$ y, si en ese intervalo de tiempo no se realiza ninguna *actuación* sobre el rodal, se verifica que, para todo $t \in [a, b]$,

$$H(t) = h(t_0, H_0, t), \tag{3}$$
$$N(t) = n(t_0, N_0, t), \tag{4}$$
$$G(t) = g(t_0, G_0, t). \tag{5}$$

Las funciones h, n y g suelen ser continuas y derivables (habitualmente h, n, $g \in C^\infty(\mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+)$) y, consecuentemente, las funciones $H(t)$, $N(t)$ y $G(t)$ definidas en $[a, b]$ también lo son.

Además, como ya hemos indicado, a partir de estas funciones puede estimarse cualquier otra variable de interés (salida) del problema. Por ejemplo, para calcular el volumen de fuste hasta un diámetro en punta delgada mayor que un cierto valor $d > 0$, se conoce otra función $v \in C^\infty(\mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^+)$ de manera que, si $V^d(t)$ denota el volumen de madera de diámetro mayor que d que hay en el rodal en el instante t, entonces

$$V^d(t) = v(H(t), N(t), G(t), d), \qquad t \geq 0.$$

Pensemos ahora que en el instante t_1 se realiza una primera clara sobre el rodal, de intensidad (proporción de árboles a extraer) $I_1 \in [0, 1]$ y tipo de clara definido por una relación de extracción (variable que relaciona la proporción de área basimétrica y la de árboles a extraer) $R_1^e \in [0, 1]$. Eso provoca que en ese instante las funciones $N(t)$ y $G(t)$ (no así $H(t)$, ya que se asume que no se ve afectada por la densidad del rodal al considerarse que las claras no afectan al estrato de árboles dominantes), sufran una discontinuidad de salto, verificándose que

$$H(t_1^+) = H(t_1^-) = H(t_1), \tag{6}$$
$$N(t_1^+) = (1 - I_1)N(t_1^-), \tag{7}$$
$$G(t_1^+) = (1 - I_1 R_1^e)G(t_1^-), \tag{8}$$

donde $H(t_1^+), N(t_1^+)$ y $G(t_1^+)$ son los valores de las funciones después de la clara y $H(t_1^-), N(t_1^-)$ y $G(t_1^-)$ los valores antes de clara.

De este modo, si hasta el instante $t_2 > t_1$ no se realiza ninguna otra actuación sobre el rodal, combinando (3)-(5) con (7)-(8) tenemos que, para todo $t \in [t_1, t_2]$,

$$N(t) = n(t_1, N(t_1^+), t) = n(t_1, (1 - I_1)N(t_1^-), t) = n(t_1, (1 - I_1)n(t_0, N_0, t_1), t),$$

$$G(t) = g(t_1, G(t_1^+), t) = g(t_1, (1 - I_1 R_1^e)G(t_1^-), t) = g(t_1, (1 - I_1 R_1^e)g(t_0, G_0, t_1), t).$$

Conviene observar que si bien las funciones $N(t)$ y $G(t)$ no son continuas en $[0, t_2]$ (tienen una discontinuidad de salto en t_1), las funciones

$$H(t) = h(t_0, H_0, t), \qquad t \in \mathbb{R}^+, \tag{9}$$
$$N_0(t) = n(t_0, N_0, t), \qquad t \in D_0, \tag{10}$$
$$N_1(t_1, I_1, t) = n(t_1, (1 - I_1)n(t_0, N_0, t_1), t), \qquad (t_1, I_1, R_1^e, t) \in D_1, \tag{11}$$
$$G_0(t) = g(t_0, G_0, t), \qquad t \in D_0, \tag{12}$$
$$G_1(t_1, I_1, R_1^e, t) = g(t_1, (1 - I_1 R_1^e)g(t_0, G_0, t_1), t), \quad (t_1, I_1, R_1^e, t) \in D_1, \tag{13}$$

sí son continuas y diferenciables (tienen la misma regularidad que n y g) en sus respectivos dominios de definición: $D_0 = \{t \in \mathbb{R}, t \geq 0\}$, $D_1 = \{(t_1, I_1, R_1^e, t) \in \mathbb{R}^3, t \geq t_1 \geq 0, I_1 \in [0, 1], R_1^e \in [0, 1]\}$.

Es fundamental destacar que este tipo de funciones son las que verdaderamente interesan en la gestión de un rodal: $H(t)$ proporciona la altura dominante en cualquier instante $t > 0$, $N_0(t)$ y $G_0(t)$ dan número de pies y área basimétrica hasta la primera clara, y $N_1(t_1, I_1, t)$ y $G_1(t_1, I_1, R_1^e, t)$ proporcionan número de pies y área basimétrica a partir de la primera clara ($t \geq t_1$), pero teniendo en cuenta que esa primera clara se hizo en el instante t_1, con una intensidad I_1 y un tipo de clara dado por R_1^e.

Este tipo de funciones son las que constituyen la función de transición del modelo y, como ya hemos señalado anteriormente, a partir de ellas pueden calcularse las otras variables de interés (salidas) del problema. Por ejemplo, el volumen de madera de diámetro mayor que d, en distintas situaciones:

1. En cualquier instante anterior a la primera clara,

$$V_0^d(t) = v(H(t), N_0(t), G_0(t), d), \qquad t \in D_0. \tag{14}$$

2. En cualquier instante posterior a la primera clara, pero teniendo en cuenta que ésta se hizo en t_1 con una intensidad I_1,

$$V_1^d(t_1, I_1, R_1^e, t) = v(H(t), N_1(t_1, I_1, t), G_1(t_1, I_1, R_1^e, t), d), \qquad (t_1, I_1, R_1^e, t) \in D_1. \tag{15}$$

3. Extraído en la primera clara,

$$V_{ext,1}^d(t_1, I_1, R_1^e) = V_0^d(t_1) - V_1^d(t_1, I_1, R_1^e, t_1), \qquad (t_1, I_1, R_1^e) \in \mathbb{R}^+ \times [0,1] \times [0,1]. \tag{16}$$

Obviamente la regularidad de estas funciones viene marcada por la regularidad de v (y de h, n y g). En particular, si v, h, n y g son continuas y diferenciables, entonces V_0^d, V_1^d y $V_{ext,1}^d$ son continuas y diferenciables en sus respectivos dominios de definición.

4. GESTIÓN ÓPTIMA DE UN RODAL

La gestión óptima de un determinado rodal pasa por elegir el programa selvícola que maximice la función objetivo especificada (e.g., el crecimiento medio, el valor actual neto o el valor esperado del suelo). Para ello, se busca el número de claras que deben hacerse ($N_C \in \mathbb{N}$), el instante ($t_i \geq 0$, $i = 1, \ldots, N_C$) en el que debe aplicarse cada clara, su intensidad ($I_i \in [0,1]$, $i = 1, \ldots, N_C$), el tipo de cada clara, definido por su relación de extracción ($R_i^e \in [0,1]$, $i = 1, \ldots, N_C$), y finalmente el instante de la corta final ($t_{N_C+1} \geq 0$). Restricciones de tipo tecnológico, logístico o económico limitan el número de claras posible y, por ello, teniendo en cuenta que N_C debe ser un número natural *pequeño* (generalmente inferior a 5), puede pensarse en eliminarlo de la variable de decisión y maximizar la función objetivo para cada valor de N_C aceptado (en lugar de un problema tendremos que resolver varios, pero en todos ellos N_C será un valor dado).

De este modo, fijado el valor de N_C, el vector que recoge las variables de control del problema es $\mathbf{u} = (t_1, I_1, R_1^e, \ldots, t_{N_C}, I_{N_C}, R_{N_C}^e, t_{N_C+1}) \in \mathbb{R}^{3N_C+1}$. De nuevo existen restricciones de los tipos indicados que determinan tiempos mínimos hasta la primera clara ($l_1 > 0$), entre la clara i y la $i+1$ ($l_i > 0$), y entre la última clara y la corta final ($l_{N_C+1} > 0$). Además, existen cota superior para la corta final ($T > 0$) y cotas inferiores ($\underline{I}_i \in (0,1)$, $\underline{R}_i^e \in (0,1)$) y superiores ($\bar{I}_i \in (0,1)$, $\bar{R}_i^e \in (0,1)$) para las intensidades y relaciones de extracción de cada clara. Con todo eso, el conjunto de actuaciones admisibles viene dado por

$$U_{ad} = \left\{ \mathbf{u} \in \mathbb{R}^{3N_C+1} \text{ tales que } \begin{array}{ll} t_i - t_{i-1} \geq l_i & i = 1, \ldots, N_C + 1, \\ I_i \in [\underline{I}_i, \bar{I}_i] & i = 1, \ldots, N_C, \\ R_i^e \in [\underline{R}_i^e, \bar{R}_i^e] & i = 1, \ldots, N_C, \\ t_{N_C+1} \leq T. & \end{array} \right\}$$

que obviamente resulta cerrado, acotado y convexo.

La gestión óptima del rodal consiste en encontrar el vector de variables de control $\mathbf{u} \in U_{ad}$ que maximice la función objetivo especificada. Ésta puede referirse, por ejemplo, a la renta en especie (e.g., crecimiento medio en volumen de madera) o a algún indicador económico. En este último caso suele emplearse el Valor Esperado del Suelo (VES), que es apropiado para comparar alternativas con diferente horizonte temporal, y que viene dado por

$$J(\mathbf{u}) = \frac{B(\mathbf{u}) - C(\mathbf{u})}{1 - \frac{1}{(1+r)^{t_{N_C+1}}}}, \tag{17}$$

donde $B(\mathbf{u})$ y $C(\mathbf{u})$ son, respectivamente, los beneficios y los costes de gestión actualizados al año cero, y $r \in [0, 1]$ es un parámetro conocido que representa la Tasa Anual Equivalente (TAE). Las funciones $B(\mathbf{u})$ y $C(\mathbf{u})$ vienen dadas a través de operaciones elementales entre $H(t)$, dada por (3) y las funciones, $N_i(t_1, I_1, \ldots, t_i, I_i, t)$, $G_i(t_1, I_1, R_1^e \ldots, t_i, I_i, R_i^e, t)$ y $V_i^d(t_1, I_1, R_1^e \ldots, t_i, I_i, R_i^e, t)$ definidas como en (11), (13) y (15) respectivamente (ver detalles en Arias-Rodil *et al.*, 2015b). De este modo, la regularidad de J viene marcada por la regularidad de h, n, g y v. En particular, si esas funciones son continuas y diferenciables, J también lo es y el problema de la gestión óptima del rodal se formula como el siguiente problema,

$$\max_{\mathbf{u} \in U_{ad}} J(\mathbf{u}), \tag{18}$$

que puede resolverse utilizando cualquier método de optimización numérica pensado para problemas diferenciables con restricciones lineales (ver, por ejemplo, Luenberger, 1989).

5. DISCUSIÓN Y CONCLUSIONES

En este trabajo se ha propuesto un modelo para la gestión de rodales forestales regulares que permite el uso de métodos de optimización diferenciable para su resolución. Las ventajas que presenta esto, con respecto a los métodos previamente utilizados en el campo forestal, son:

- Al contrario de los métodos de programación dinámica y búsqueda en profundidad, ahora no se require una discretización de los valores de las variables de decisión, por lo que potencialmente puede alcanzar mayores valores de la función objetivo.
- Los métodos de optimización diferenciable no realizan una búsqueda ciega en todo el espacio de soluciones (como puede ocurrir con los métodos de búsqueda directa) si no que hacen uso del vector gradiente (dirección de máximo crecimiento), reduciendo de forma significativa los tiempos de computación
- Para resolver el problema (18), los métodos de optimización diferenciable son mucho menos sensibles al punto inicial y, en la experiencia que tenemos (Arias-Rodil *et al.*, 2015b), proporcionan siempre los mismos resultados, que son al menos iguales a los de la mejor optimización de los anteriores.
- El aumento del vector de variables de decisión incrementa enormemente el tiempo de computación de los métodos de programación dinámica y búsqueda en profundidad, y ralentiza considerablemente los de búsqueda directa, mientras que los de tipo gradiente se muestran mucho más efectivos.

REFERENCIAS

Amidon, E., y Akin, G. (1968). Dynamic programming to determine optimum levels of growing stock. Forest Science, 14(3), 287–291.

Arias-Rodil, M., Pukkala, T., González-González, J. M., Barrio-Anta, M., y Diéguez-Aranda, U. (2015a). Use of depth-first search and direct search methods to optimize even-aged stand management: a case study involving maritime pine in Asturias (NW Spain). Canadian Journal of Forest Research (disponible online).

Arias-Rodil, M., Diéguez-Aranda, U., y Vázquez-Méndez, M.E. (2015b). Optimization of even-aged stand management using a differentiable model. Submitted.

Arimizu, T. (1958). Regulation of the cut by dynamic programming. Journal of the Operations Research Society of Japan, 1(4), 175–182.

Bettinger, P., Boston, K., Siry, J. P., y Grebner, D. L. (2009). Forest management and planning. Amsterdam: Elsevier.

Cao, T. (2010). Silvicultural decisions based on simulation-optimization systems. Dissertationes Forestales. Finnish Forest Research Institute.

Clutter, J. L., Fortson, J. C., Pienaar, L. V., Brister, G. H., y Bailey, R. L. (1983). Timber management: a quantitative approach. Colorado: Wiley.

Díaz-Balteiro, L., y Rodríguez, L. C. E. (2006). Optimal rotations on Eucalyptus plantations including carbon sequestration—A comparison of results in Brazil and Spain. Forest Ecology and Management, 229(1-3), 247–258.

Diéguez-Aranda, U., Rojo Alboreca, A., Castedo-Dorado, F., Álvarez González, J. G., Barrio-Anta, M., Crecente-Campo, F., González González, J.M., Pérez-Cruzado, C., Rodríguez Soalleiro, R., López-Sánchez, C.A., Balboa-Murias, M.A., Gorgoso Varela, J.J., y Sánchez Rodríguez, F. (2009). Herramientas selvícolas para la gestión forestal sostenible en Galicia. Xunta de Galicia.

García, O. (1994). The state-space approach in growth modelling. Canadian Journal of Forest Research, 24(9), 1894–1903.

Luenberger D.G. (1989) Programación lineal y no lineal. Addison-Wesley Iberoamericana.

Pukkala, T., y Kellomäki, S. (2012). Anticipatory vs adaptive optimization of stand management when tree growth and timber prices are stochastic. Forestry, 85(4), 463–472.

Pukkala, T. (2009). Population-based methods in the optimization of stand management. Silva Fennica, 43(2), 261–274.

Tahvonen, O., Pihlainen, S., y Niinimäki, S. (2013). On the economics of optimal timber production in boreal Scots pine stands. Canadian Journal of Forest Research, 43(8), 719–730.

Valsta, L. (1993). Stand management optimization based on growth simulators. Research Paper 453. Finnish Forest Research Institute.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Functional prediction with Applications to the Electricity Market

Paula Raña[1], Juan Vilar[1] e Germán Aneiros[1]

[1]Departamento de Matemáticas, Universidade da Coruña

ABSTRACT

A comparative study among different techniques for prediction in the Spanish Electricity Market is presented in this paper. Specifically, it deals with the problem of daily electricity demand forecasting, together with electricity price. Different functional regression methods are applied and compared with other classical methodologies for electricity data forecasting. Exogenous variables are also considered in some of the models, including weather variables and wind power production. Also combined methods are developed in order to get better prediction errors, mixing some of the best methods provided in the comparison. Results show an improvement when the forecasting is performed treating data as functional and also adding exogenous variables.

Palabras e frases chave: Load and price; Electricity markets; Functional data; Functional time series forecasting.

1. INTRODUCTION

Forecasting has been always an important task for the agents and companies involved in the Electricity Market. The features of the electricity, that is a non-storable product, and also the rules of this competitive market create the need of good predictions, in order to anticipate decisions.

One can find several proposals for prediction, specifically addressed for electricity data, in the literature. Classical approaches are mainly based on naïve predictions or the use of time series models. In Weron (2006 and 2014), one can find a very exhaustive review about demand and price forecasting.

Over the recent years, the use of functional data has spread to a large number of areas. See the monographs Ramsay and Silverman (2005), Ferraty and Vieu (2006) or, more recently, Horváth and Kokoszka (2012), that are the main references in this topic. Also, functional data can be use to deal with electricity prediction. Some examples can be found in Antoch et al. (2010), Liebl (2013) or Cho et al. (2013).

In Spain, since the deregulation of the Electricity Market in 1998, the interest in this topic has grown. Examples of predictions for the Spanish Electricity Market can be found in Cancelo et al. (2008), Aneiros et al. (2013) and Vilar et al. (2012). Both last references also deal with functional data.

The main objective of the study is to forecast next-day electricity demand and price daily curves for the year 2012 in Spain. Different models, including some classical ones and also models for functional data are applied and their results are analysed in a comparative study. Also the proposal based on robust functional principal component analysis in Hyndman and Ullah (2007) will be considered. Models follow essentially autoregressive ideas, in the sense that the predictions for the next day will be built based on the historical recorded values for the last year. Also, some of the proposals allow to incorporate exogenous variables that improve the results.

This paper can be seen as an extension in two ways of the proposal given in Vilar et al. (2012). The first extension is related to the response in the models: Vilar et al. (2012) considered both nonparametric and partial linear models with scalar response (hourly electricity demand or price) and functional covariates. Meanwhile this paper deals, among others, with those same models but

considering functional response (daily electricity demand or price curve). The second extension takes part in the information managed by the partial linear model: while Vilar et al. (2012) did not consider exogenous covariates, this paper will take into account information from weather variables and wind power production. In the case of the demand, we consider as exogenous variables data about temperature provided by AEMET (Agencia Estatal de Meterología), which has to be transformed in order to get a linear effect (see Pardo et al. (2002)). For the price, both demand and wind power production will be used.

One of the main features of the electricity data is the presence of outliers. It is important to identify this outliers in order to prevent misleading predictions. Methods presented in Raña et al. (2015) and in Vilar et al. (Technical report) to detect outliers in functional time series, taking into account the dependence in the data, will be used for that purpose.

Extended version of this study can be found in Aneiros et al. (Technical report).

Acknowledgments
This research was partially supported by Grants MTM2014-52876-R and CN2012/130 from Spanish Ministerio de Economía y Competitividad and Xunta de Galicia, respectively. The authors also want to thank AEMET for providing the temperature data used in this study.

REFERENCES

Aneiros G, Vilar JM, Cao R and Muñoz A. (2013). Functional prediction for the residual demand in electricity spot markets. IEEE Trans. Power Syst., vol. 28, No. 4, 4201–4208.

Aneiros G, Vilar JM and Raña P. Short-term forecast of daily curves of electricity demand and price. (Technical report).

Antoch J, Prchal L, De Rosa MR, and Sarda P. (2010). Electricity consumption prediction with functional linear regression using spline estimators. Journal of Applied Statistics, vol. 37, 2027–2041.

Cancelo JR, Espasa A and Grafe R. (2008). Forecasting the electricity load from one day to one week ahead for the Spanish system operator. International Journal of Forecasting, vol. 24, 588–602.

Cho H, Goude Y, Brossat X and Yao Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach. Journal of the American Statistical Association, vol. 108, 7–21.

Ferraty F and Vieu P. (2006). Nonparametric Functional Data Analysis. Springer Series in Statistics. New York.

Horváth L and Kokoszka P. (2012). Inference for functional data with applications. Springer.

Hyndman RJ and Ullah S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. Comput. Statist. Data Anal., vol. 51, 4942–4956.

Liebl D. (2013), Modeling and forecasting electricity spot prices: a functional data perspective. Annals of Applied Statistics, vol. 7, 1562–1592.

Pardo A, Meneu V and Valor E. (2002). Temperature and seasonality influences on Spanish electricity load. Enery Econ., vol. 24, 55–70.

Ramsay JO and Silverman BW. (2005). Functional Data Analysis. New York: Springer-Verlag.

Raña P, Aneiros G and Vilar JM. (2015), Detection of outliers in functional time series. Environmetrics, vol. 26, 178–191.

Vilar JM, Cao R and Aneiros G. (2012). Forecasting next–day electricity demand and price using nonparametric functional methods. Int. J. Elect. Power Energy Syst., vol. 39, 48–55.

Vilar JM, Raña P and Aneiros G. Using robust FPCA to identify outliers in functional time series, with applications to the electricity market. (Technical report).

Weron R. (2006). Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach. Wiley.

Weron R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, vol. 30(4), 1030–1081.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

ARTICLERANK: UNA NUEVA FORMA DE ORDENAR ARTÍCULOS

Jose Ameijeiras–Alonso[1], Rosa M. Crujeiras[1] y Alberto Rodriguez–Casal[1]

[1]Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela.

RESUMEN

La evaluación del rendimiento de una publicación científica en una determinada disciplina es algo que resulta de especial interés para el campo de las ciencias de la documentación y para las administraciones públicas y privadas que dedican parte de sus recursos a generar y transferir conocimiento. En este sentido, es importante poder clasificar los resultados obtenidos por una determinada revista, artículo, investigador o institución. Con este objetivo surgen, desde mediados del siglo XX, una serie de índices que permiten realizar dicha clasificación.

Los resultados obtenidos con estos índices en el campo metodológico de la probabilidad y la estadística, al igual que en otros campos científicos, son objeto de debate entre la comunidad científica, ya que parece que ninguno de los índices empleados por las grandes webs de consulta consigue generar una clasificación satisfactoria. Con este objetivo, se ha desarrollado en este trabajo un nuevo método de clasificación, el ArticleRank, que pretende ordenar los distintos artículos en función no solo del número de citas que tenga un artículo, sino que también de la "relevancia" metodológica que estos tengan dentro de una determinada rama de estudio. A partir de este método que permitirá ordenar artículos, autores, instituciones o revistas, se presentará la clasificación de revistas en estadística y probabilidad y se comparará el ArticleRank con los índices presentados en las distintas webs de consulta: Google Scholar Metrics, Journal Citation Reports, Journal Indicators, Journal Metrics y SCImago.

Palabras y frases clave: Cadenas de Markov. Cienciometría. Medidas de influencia de artículos. Ordenación de revistas. PageRank.

1. INTRODUCCIÓN

Conocer el rendimiento de las publicaciones científicas y evaluar la actividad los distintos investigadores es una problemática que ha interesado tanto a los expertos en ciencias de la documentación, como a las distintas instituciones que dedican sus recursos a este fin. En este sentido, es importante disponer de una ordenación que permita evaluar de forma precisa cuando un determinado artículo se puede considerar "entre los mejores" de los publicados en una determinada rama de estudio. Con este objetivo nace, a mediados del siglo XX, una nueva rama que se denomina la cienciometría. Haciendo uso de los términos que aparecen en el título de una de las referencias clásicas de esta rama (véase de Solla Price, 1986), el objetivo de este trabajo será emplear las bases de datos de grandes proyectos (*big science*) para poder elaborar una modesta contribución con una nueva clasificación (*little science*) que permitirá ordenar los distintos artículos englobados dentro de la temática de probabilidad y estadística.

Con el objetivo de analizar el funcionamiento y ver las limitaciones de las clasificaciones más habituales que se emplean para esta ordenación, en la Sección 2 se describirán los distintos índices empleados por las principales webs de consulta: Thomson Reuters (2015), Google Scholar Metrics (2015) y Journal Metrics (2015), que agrupa la información de Journal Indicators (2015) y SCImago (2007). Además, se describirá un nuevo método para clasificar tanto las revistas, como los artículos de un determinado área del conocimiento. Finalmente, en la Sección 3, se empleará este nuevo método para hacer una clasificación de las revistas de probabilidad y estadística, comparando el resultado con el proporcionado por otros índices.

2. ÍNDICES PARA ORDENAR REVISTAS

En esta sección se presentará una revisión de los principales índices que sirven para ordenar revistas. Dentro de estos índices existen dos tendencias principales: aquellos que se basan únicamente en el número de citas que tienen las revistas sin importar quién hace estas referencias, que serán presentados en la primera parte de esta sección, y los que tienen en cuenta la relevancia que tiene la revista que les está citando, que serán presentados en la segunda parte. Además, para finalizar esta sección, se introducirá un nuevo método de ordenación para artículos, que será valido también para autores, instituciones o revistas.

2.1. ÍNDICES BASADOS EN EL NÚMERO DE CITAS

Para esta revisión de los distintos índices que permiten ordenar revistas, se comenzará con el indicador bibliométrico más popular y quizás controvertido de los disponibles en este momento, el factor de impacto. Empleado desde 1976 en la web de Thomson Reuters (2015) y propuesto por Garfield y Sher (1963), el factor de impacto ordena las revistas realizando el cociente entre el número de citas recibidas y el número de artículos publicados durante un período determinado, originalmente durante dos años, actualmente también durante cinco años (debido a las críticas recibidas por emplear un período tan corto; véase, por ejemplo, Seglen, 1997) y durante un año a través del índice de inmediatez, que mide cómo de rápido son citados los artículos en una revista.

El índice de Impacto Por Publicación (*IPP*) empleado por Journal Indicators (2015) se define de forma análoga. Las principales diferencias son que no tiene en cuenta todas las publicaciones, sino que solamente emplea los documentos catalogados como *peer–reviewed* (revisión por pares), y que emplea un período de tres años.

Otra de las críticas que suele recibir el factor de impacto es que no tiene en cuenta que el número de citas puede variar substancialmente de una disciplina a otra. Con el objetivo de solventar este inconveniente, además del *IPP*, en el Journal Indicators (2015) se emplea también el índice denominado como Impacto por Publicación Normalizado por Artículo (*SNIP* por sus siglas en inglés) introducido por Moed (2010). Teniendo en cuenta solo los documentos catalogados como *peer–reviewed*, este índice consiste en dividir el *IPP* de cada revista por su potencial relativo de citación. El potencial relativo se define como el potencial de citación de una revista (número de citas que reciben las revistas de la misma disciplina que la revista que se está analizando[1] entre número de artículos publicados en dicha disciplina) entre la mediana del potencial de citación de todas las revistas de la base de datos.

Otro de los índices más empleados, y que se puede encontrar en SCImago (2007) o en Google Scholar Metrics (2015), es el índice H introducido por Hirsch (2005), que se define como el máximo número h, de forma que h artículos de una determinada revista fueron citados al menos h veces, desde su publicación.

El problema que presentan el factor de impacto, el *IPP* y el índice H es que solo tienen en cuenta el número bruto de citas que se realizan a los artículos de una revista y no de quién son estas citas. En cuanto al *SNIP*, aunque también hereda este problema, sí que introduce una gran mejora con respecto al resto de índices que utilizan las grandes webs y es que relativiza según la temática de donde provienen. Con el fin de solventar el principal inconveniente que tienen los índices presentados hasta el momento, surgen los índices que se describirán a continuación y que no solo tienen en cuenta solo el número de citas, sino también quién realiza esas citas.

2.2. ÍNDICES BASADOS EN QUIÉN CITA A LOS ARTÍCULOS

El resto de índices que se analizarán, así como la nueva propuesta que se presenta, se basan en el algoritmo inicial de Google, el PageRank, introducido por Brin y Page (2012)[2]. En el caso del buscador, este algoritmo ordena las webs dándole un valor de relevancia a cada página en función del número de webs que enlacen con ella y de la relevancia que estas otras páginas tengan. Basándose en esta idea de tener en cuenta la "importancia" de quién cita a cada artículo, para

[1] Aquí, se considera de la misma disciplina a todas aquellas revistas cuyos artículos han realizado, en el año de estudio, al menos una referencia a los artículos publicados, en los últimos 8 años, por la revista que se está analizando.

[2] Este artículo es una reedición del artículo original que data de 1998.

ordenar revistas surgen los índices *eigenfactor* y el de influencia de artículo, que se pueden encontrar en Thomson Reuters (2015), y el índice *SJR*, que aparece en Journal Metrics (2015) y en SCImago (2007).

A continuación, se procederá a realizar una breve descripción de como funciona el PageRank para entender la implementación del resto de métodos de ordenación de revistas. Aunque en su artículo original, Brin y Page (2012) no describen su algoritmo desde el punto de vista de las cadenas de Markov, con el objetivo de probar que su algoritmo funciona correctamente, sí que se tomará este enfoque en este documento.

En primer lugar se va a suponer que se quieren ordenar un grupo de páginas webs, catalogadas como $\{1, \ldots, \tilde{M}\}$. Este conjunto de $\tilde{M}$ posibles valores, al que se denotará como $\tilde{S}$, constituirá lo que se denomina el espacio de estados. Si X_n representa la página web que se está visitando en el tiempo n, con $n = 0, 1, 2, \ldots$[3]; se dirá que X_n es una cadena de Markov en tiempo discreto con matriz de transición $\tilde{\mathbf{P}}$ de dimensión $\tilde{M} \times \tilde{M}$ (véase, por ejemplo, Durret, 2012; Cap. 1), si para cualquiera que sean $j, i, i_{n-1}, \ldots, i_0 \in \tilde{S}$, se verifican los siguientes puntos:

- $\mathbb{P}(X_{n+1}|X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1}|X_n = i) = \tilde{\mathbf{P}}(i, j)$,
- $\tilde{\mathbf{P}}(i, j) \geq 0$,
- $\sum_{k \in \tilde{S}} \tilde{\mathbf{P}}(i, k) = 1$.

Denominando como "navegante aleatorio" a la persona que cuando está en la página i, en el instante n, ve los enlaces que aparecen en dicha web y escoge al azar a cual de ellos ir en el instante $(n + 1)$. Se tendría que, si se denota como $\mathbf{W}(i, j)$ al elemento de la matriz $\mathbf{W}$ que toma el valor 1 si la página i tiene un enlace a la página j y 0 en otro caso, para este navegante aleatorio, la $\mathbb{P}(X_{n+1}|X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbf{W}(i, j)/\sum_{k \in \tilde{S}} \mathbf{W}(i, k) = \tilde{\mathbf{H}}(i, j)$, para cualquiera que sean las webs $j, i, i_{n-1}, \ldots, i_0 \in \tilde{S}$.

La matriz $\tilde{\mathbf{H}}$ sería una matriz de transición si no fuese porque algunas páginas webs no enlazan con ninguna otra, es decir, existen $i \in \tilde{S}$, de forma que $\sum_{k \in \tilde{S}} \tilde{\mathbf{H}}(i, k) = 0 \neq 1$. Para solucionar este inconveniente, al navegante aleatorio, cuando no encuentra páginas a las que enlazar, se le añade la posibilidad de introducir él mismo una nueva dirección web, siendo escogida una de las $\tilde{M}$ páginas al azar. La matriz $\tilde{\mathbf{H}}'$ obtenida al reemplazar las filas de 0 por el vector $(1/\tilde{M}) \cdot \mathbf{e}_{\tilde{M}}^T$, donde $\mathbf{e}_{\tilde{M}}^T$ es la traspuesta del vector $\mathbf{e}$ formado por $\tilde{M}$ unos, reflejaría el comportamiento de este navegante aleatorio.

La clasificación de las páginas webs que hace el PageRank, sería obtenida ordenando, de mayor a menor, los valores del vector fila $\boldsymbol{\pi}_{\tilde{\mathbf{H}}'}^k$, que se obtienen iterando en k, a partir de $\boldsymbol{\pi}_{\tilde{\mathbf{H}}'}^0 = (1/\tilde{M}) \cdot \mathbf{e}_{\tilde{M}}^T$, la siguiente expresión:

$$\boldsymbol{\pi}_{\tilde{\mathbf{H}}'}^k = \boldsymbol{\pi}_{\tilde{\mathbf{H}}'}^{k-1} \cdot \tilde{\mathbf{H}}', \tag{1}$$

hasta un cierto momento que vendrá dado por un criterio de parada que se alcanzará cuando $\boldsymbol{\pi}_{\tilde{\mathbf{H}}'}^k$ y $\boldsymbol{\pi}_{\tilde{\mathbf{H}}'}^{k-1}$ estén lo suficientemente "próximos". No obstante, en este caso, la convergencia de este algoritmo a un único vector que coincida con el número medio esperado de veces que se visitarán las distintas webs no está garantizada.

Para solucionar esto, al navegante aleatorio se le añade un hecho que ocurre cuando se está navegando por la red y es que no siempre se emplean los vínculos para ir a una nueva página, sino que a veces, se introduce una nueva dirección a mano. Si se le asigna una probabilidad $\alpha \in (0, 1)$ de que el navegante vaya a una nueva página a través de los enlaces y una probabilidad $(1 - \alpha)$ de que introduzca una nueva web a mano, se tendría que la siguiente matriz de transición $\tilde{\mathbf{P}}$ reflejaría el comportamiento de este navegante aleatorio:

$$\tilde{\mathbf{P}} = \alpha \tilde{\mathbf{H}}' + (1 - \alpha) \tilde{\mathbf{a}} \mathbf{e}_{\tilde{M}}^T,$$

[3]Entendiendo que se avanza del tiempo n al $(n+1)$ cada vez que se cambia de web, aunque se cambie a la misma página.

donde $\tilde{\mathbf{a}} = (1/\tilde{M}) \cdot \mathbf{e}_{\tilde{M}}$ es la probabilidad de escoger cada una de las páginas al azar.

La inclusión del factor $(1-\alpha)\tilde{\mathbf{a}}\mathbf{e}_{\tilde{M}}^T$ es fundamental para que se verifiquen las hipótesis del Teorema 1 que se enunciará a continuación, ya que la matriz $\tilde{\mathbf{H}}'$, por si sola, no tiene porqué satisfacerlas. Este teorema garantiza que el algoritmo iterativo dado en (1) converge cuando se reemplaza $\tilde{\mathbf{H}}'$ por $\tilde{\mathbf{P}}$ y tiene solución única al vector que coincide con el número medio esperado de veces que se visitarán las distintas webs. Para poder poder enunciar el Teorema 1, antes será necesario introducir el concepto de matriz de transición recurrente e irreducible.

Se dirá que una matriz de transición es recurrente si todos los estados de $\tilde{S}$ son recurrentes, es decir, $\forall i \in \tilde{S}$ se verifica que

$$\mathbb{P}(T_i < \infty | X_0 = i) = 1, \text{ donde } T_i = \text{mín}\{n \geq 1 : X_n = i\},$$

esto es, se dirá que $i \in \tilde{S}$ es un estado recurrente si se regresa de forma segura a la web i, tras un número finito de pasos, cuando se comienza en ella.

Por otro lado, se dirá que la matriz de transición es irreducible si su espacio de estados lo es, es decir, si $\forall i, j \in \tilde{S}$ se verifica que

$$\mathbb{P}(T_i < \infty | X_0 = j) > 0,$$

esto es, si es posible ir a cualquier web j desde cualquier web i en un número finito de pasos.

Cuando se tiene una matriz de transición irreducible y recurrente, como es el caso de $\tilde{\mathbf{P}}$, entonces se verifica que el algoritmo iterativo dado en (1) converge a un único vector cuando se emplea $\tilde{\mathbf{P}}$ en vez de $\tilde{\mathbf{H}}'$, y cada elemeneto de este vector coincide precisamente con el número medio esperado de veces que el navegante aleatorio visitará cada web.

Teorema 1. (Durret, 2012; Cap. 1) Si el espacio de estados $\tilde{S}$ es finito, irreducible y recurrente, entonces existe un único vector fila $\boldsymbol{\pi}_{\tilde{\mathbf{P}}}$, denominado distribución estacionaria, verificando la igualdad $\boldsymbol{\pi}_{\tilde{\mathbf{P}}}\tilde{\mathbf{P}} = \boldsymbol{\pi}_{\tilde{\mathbf{P}}}$ y para cualquier estado $i \in \tilde{S}$ se tiene que, con probabilidad uno,

$$\frac{N_n(i)}{n} \to \boldsymbol{\pi}_{\tilde{\mathbf{P}}}(i),$$

siendo $N_n(i)$ el número de visitas que se realiza al estado i después de haber realizado n pasos.

La idea que subyace el método *eigenfactor* es similar a la del PageRank (véase, por ejemplo, Bergstrom et al., 2008). Para obtener este índice, en primer lugar, será necesario construir la matriz donde se almacenará cómo conectan unas revistas con otras, para ello, se emplea una matriz similar a la del PageRank, simplemente modificando los enlaces de las páginas webs por el número total de citas que los artículos de una revista le dan a otra. Para construir esta matriz, se define como $\mathbf{Z}_5$ a la matriz cuyos elementos $\mathbf{Z}_5(i,j)$ son las citas que hacen los artículos de la revista i, en un determinado año, a los artículos publicados por la revista j en los cinco años anteriores, donde $i, j \in S = \{1, \ldots, M\}$, siendo M el número total de revistas que hay en la base de datos. A partir de estos valores, se toma como $\mathbf{H}$ a la matriz cuyos elementos son $\mathbf{H}(i,j) = \mathbf{Z}_5(i,j) / \left(\sum_{k=1}^{M} \mathbf{Z}_5(k,j) - \mathbf{Z}_5(j,j)\right)$, si $i \neq j$, y $\mathbf{H}(i,i) = 0$, con $i, j \in S$, con el objetivo de omitir las auto–citas que se hace cada revista. Una vez se obtiene esta matriz, como algunas revistas no citan a ninguna otra[4], se crea la matriz $\mathbf{H}'$ modificando las columnas de ceros de $\mathbf{H}$ (los "nodos colgantes") por el vector $\mathbf{a}$, donde $\mathbf{a}(i)$ es el número de artículos publicados por la revista i en los 5 años anteriores entre el número total de artículos publicados en ese período. Finalmente, la matriz de transición $\mathbf{P}$ que se emplea en el *eigenfactor* es:

$$\mathbf{P} = \alpha\mathbf{H}' + (1-\alpha)\mathbf{a} \cdot \mathbf{e}_M{}^T,$$

donde $\mathbf{e}_M{}^T$ es un vector fila de M unos y $\alpha \in (0,1)$. Para entender un poco mejor esta matriz, si se supone que se está ante un investigador que va visitando las revista únicamente en función de las citas que estas tienen y que se encuentra en la revista i, si esta citó dos veces a la revista j, una a la revista k y ninguna al resto de revistas en los últimos 5 años, sería de esperar que este investigador fuese a la revista j con una probabilidad de 2/3 y a la revista k con una probabilidad de 1/3. El

[4]Un ejemplo de este caso sería si solo hubiese una revista en la base de datos de una cierta temática y todos los artículos de esa revista citasen a artículos de esa disciplina que no están en la base de datos.

término α, distingue dos situaciones: con una probabilidad α el investigador sigue las referencias que hay en la revista, y con una probabilidad $(1-\alpha)$ visita cualquier artículo de cualquier revista al azar. Esta última sería la misma estrategia que seguiría el investigador cuando se encuentre ante una revista que no cite a ninguna otra.

Imitando al PageRank, una forma de ordenar las revistas se obtiene analizando cuántas veces se espera que el investigador aleatorio visite a cada revista de media. Esto es, empleando el Teorema 1, ordenando los valores de $\boldsymbol{\pi}_{\mathbf{P}}$ (la distribución estacionaria de $\mathbf{P}$) asociados a las distintas revistas. No obstante, en el caso del algoritmo de *eigenfactor*, para eliminar la aportación "artificial" realizada por los nodos colgantes, estas son ordenadas utilizando del índice $\mathbf{EF}(i)$, el cual se obtendría, a partir de la matriz original $\mathbf{H}$, como

$$\mathbf{EF}(i) = 100 \frac{(\boldsymbol{\pi}_{\mathbf{P}}\mathbf{H})(i)}{\sum_{k=1}^{M} (\boldsymbol{\pi}_{\mathbf{P}}\mathbf{H})(k)}.$$

A partir de este índice, para ver el impacto que tiene en media cada artículo de la revista $i \in S$, se define el índice de influencia de artículo de la revista i, $\mathbf{IA}(i)$, como el cociente

$$\mathbf{IA}(i) = \frac{\mathbf{EF}(i)}{100 \cdot \mathbf{a}(i)}.$$

Habitualmente, se considera como una ventaja de estos métodos (*eigenfactor* e influencia de artículo), frente al factor de impacto, el que no solo tiene en cuenta cuantas citas tiene una revista, sino que, además, considera quién la está citando, si es una revista "de prestigio" o no. No obstante, en el caso que se analiza en este documento, que es el de las revistas de probabilidad y estadística, este hecho se convierte en el principal inconveniente de este método, ya que las revistas que se sitúan en lo más alto de esta clasificación (véase, por ejemplo, Bergstrom, 2007), englobadas, en algunos caso en contextos aplicados y/o marcadamente computacionales, es mucho más probable que citen a revistas estadísticas con aplicaciones en ámbitos como la medicina o a la biología que a revistas que presenten avances metodológicos en el campo de la probabilidad y estadística, pero no tengan carácter aplicado, como se verá en la Sección 3. Otro inconveniente es que la modificación aportada por el vector $\mathbf{a}$ para la corrección de $\mathbf{H}$, no parece la más adecuada, ya que hace que aquellas revistas que publican muchos artículos aparezcan en una posición alta, lo cual entra en contradicción con la corrección que se persigue con el índice de influencia de artículo. Además, con estos métodos se está suponiendo que, cuando el investigador aleatorio busque una nueva revista, ya sea porque la revista en la que se encuentra no tiene referencias o porque está en el caso $(1-\alpha)$, el comportamiento va a ser el mismo independientemente de la rama que estuviese investigando previamente. Una posible solución, para replicar de mejor forma lo que hacen los investigadores, podría ser emplear el factor de impacto en vez de $\mathbf{a}$ ya que es más probable que un investigador visite un artículo con muchas citas que uno con pocas, ponderando estos elementos en función de la revista en la que se encontraba el investigador aleatorio.

El índice *SJR* introducido por González–Pereira et al. (2010) también se basa en el algoritmo de PageRank. En este caso, para calcular la matriz de citas, a la que se denotará como $\mathbf{G}$, toman como $\mathbf{Z}_3(i,j)$ las citas que hacen los artículos de la revista i en un determinado año a los artículos *peer–reviewed* publicados por la revista j en los tres años anteriores, con $i,j \in S$, siendo limitado el número de autocitas al 33 % de las citas totales. Esto es, los elementos de la matriz $\mathbf{G}$ serán $\mathbf{G}(i,i) = \text{mín}\left\{0{,}33; \mathbf{Z}_3(i,i)/\sum_{k=1}^{M}\mathbf{Z}_3(k,i)\right\}$ y $\mathbf{G}(i,j) = \mathbf{Z}_3(i,j)/\left(\left((1+G(j,j))\sum_{k=1}^{M}\mathbf{Z}_3(k,j)\right) - \mathbf{Z}_3(j,j)\right)$, si $i \neq j$ e $i,j \in S$. Para obtener el índice *SJR*[5], fijando un vector $\boldsymbol{\rho}_{\mathbf{1}}$, con $\boldsymbol{\rho}_{\mathbf{1}}(i) = 1/M, \forall i \in S$, se va iterando en k hasta que $\sum_{i=1}^{M} |\boldsymbol{\rho}_{\boldsymbol{k}}(i) - \boldsymbol{\rho}_{\boldsymbol{k-1}}(i)| < 10^{-5}$. Si el algoritmo se detiene en la iteración k, las revistas se ordenarán utilizando el vector de índices $\beta\boldsymbol{\rho}_{\boldsymbol{k}}(i)/\mathbf{b}(i)$, siendo β una constante y $\mathbf{b}(i)$ el número de artículos publicados por la revista i. Finalmente, la actualización de $\boldsymbol{\rho}_{\boldsymbol{k}}$ en cada iteración se obtiene como

[5]En este caso, González–Pereira et al. (2010) no tratan a su método como una cadena de Markov, ni prueban la convergencia única del SJR al vector que coincide con el número medio esperado de veces que se visitarán las distintas revistas. No obstante, en la práctica parece que si que funciona, ya que en las webs de SCImago (2007) y Journal Metrics (2015) se muestra la ordenación que obtienen.

$$\boldsymbol{\rho_k} = \frac{(1-\delta-\epsilon)\mathbf{e}_M}{M} + \epsilon \cdot \mathbf{a} + \delta \left(\varsigma \boldsymbol{\rho_{k-1}} \mathbf{G} + \mathbf{a} \sum_{l \in \mathcal{C}} \boldsymbol{\rho_{k-1}}(l) \right),$$

siendo $\delta = 0{,}9$, $\epsilon = 0{,}0999$, $\mathcal{C}$ el conjunto formado por los "nodos colgantes" y ς es la proporción de prestigio que tenían los nodos que no pertenecen a $\mathcal{C}$ entre el total disponible para esta iteración, esto es

$$\varsigma = \frac{\left(1 - \sum_{l \in \mathcal{C}} \boldsymbol{\rho_{k-1}}(l)\right)}{\sum_{i=1}^{M} \sum_{j=1}^{M} \mathbf{G}(i,j) \boldsymbol{\rho_{k-1}}(i)}.$$

En el caso del índice *SJR* los "nodos colgantes" toman más importancia que en el *eigenfactor* lo cual hace que se puedan agravar los inconvenientes asociados al vector $\mathbf{a}$ ya comentados cuando se analizó el *eigenfactor*. Además, en este índice se sigue sin tener en cuenta de qué temática proceden las referencias y, por tanto, heredará el inconveniente ya mencionado de que revistas de carácter interdisciplinar, pero que presentan pocos avances metodológicos en el campo de la estadística y la probabilidad, saldrán más altas en la clasificación.

2.3. NUEVO ÍNDICE PARA ORDENAR ARTÍCULOS

Para el diseño del nuevo índice que permitirá ordenar los distintos artículos de revistas de la misma temática, en primer lugar se recurrirá a los $\mathbf{Z}_5(i,j)$ introducidos en la Sección 2.2., que denotaban el número de citas que hacen los artículos de la revista i en un determinado año a los artículos publicados por la revista j en los cinco años anteriores, pero con una diferencia importante: en vez de tomar todas las posibles revistas $i, j \in S$, se escogerá únicamente el conjunto de las revistas de temática de probabilidad y estadística[6] al que se denotará como C, esto es, $i, j \in C \subset S$. Aunque aquí ya se ha tomado una clasificación "prefabricada", en general, se entenderá que un grupo de revistas pertenecen a la misma categoría si su matriz de transición $\mathbf{R}$, cuyos elementos son $\mathbf{R}(i,j) = \mathbf{Z}_5(i,j)/(\sum_{k=1}^{M} \mathbf{Z}_5(k,j))$, con $i, j \in C \subset S$ es irreducible y recurrente, puesto que parece adecuado catalogar que dos revistas i y j están en la misma rama del conocimiento solo cuando un investigador que empieza en la revista i puede llegar, con probabilidad 1 en un número finito de pasos, a la revista j a través de las referencias que se va encontrando.

Como la matriz de transición $\mathbf{R}$ es recurrente e irreducible, se puede emplear el Teorema 1, enunciado en la Sección 2.2., para obtener cuántas veces el investigador aleatorio visitará de media cada revista a través de la distribución estacionaria de la matriz de transición $\mathbf{R}$, $\boldsymbol{\pi}_{\mathbf{R}}$. Tomando el vector $\boldsymbol{\vartheta}$, que se obtiene realizando el cociente $\boldsymbol{\vartheta}(i) = \boldsymbol{\pi}_{\mathbf{R}}(i)/\mathbf{b}(i)$ para cada revista i, donde $\mathbf{b}(i)$ es el número de artículos publicados por la revista $i \in C$ en el año 2014, se puede obtener una aproximación de la probabilidad con que el investigador aleatorio visitará un artículo de una revista. Aunque el vector $\boldsymbol{\vartheta}$ constituye en sí una forma de ordenar las revistas que no hereda los problemas del *eigenfactor*, ni del *SJR*, las asimetrías y la fuerte presencia de atípicos que suelen presentar los índices bibliográficos (véase Franceschet, 2010), hacen que no sea aconsejable emplear esta media como medida de centralidad. No obstante, este vector $\boldsymbol{\vartheta}$ será la clave para poder crear nuestro algoritmo: el ArticleRank.

A partir del vector $\boldsymbol{\vartheta}$, que refleja de forma aproximada con qué probabilidad visitaría un investigador aleatorio que se encontrase en el año 2014 los artículos de las revistas de probabilidad y estadística de los cinco años anteriores, se puede obtener la clasificación de cuáles fueron los mejores artículos publicados en un período de 5 años (2010–2014). Para ello, denotando como Q el conjunto de artículos publicados en las revistas contenidas en C durante este período de tiempo, los elementos de la matriz $\mathbf{V}$, se tomarán $\mathbf{V}(i',j') = 1$ si el artículo i'[7] citó al artículo j' e

[6] Para ello, se tomarán las 119 revistas que tienen esta catalogación en Thomson Reuters (2015) para el 2014 a excepción de la revista *Open Systems & Information Dynamics* ya que en el año 2014 no citó a ninguna de las otras revistas de esta categoría publicadas entre 2009 y 2013, ni fue citada durante este período por ninguna de estas revistas en el 2014.

[7] Con el fin de facilitar la comprensión del trabajo, se denotará con los índices i', j' y k' a los artículos y con i, j y k a las revistas.

$i', j' \in Q$; e iguales a 0 en otro caso. A partir de estos elementos se definiría la matriz $\mathbf{L}$ como $\mathbf{L}(i', j') = \mathbf{V}(i', j')/(\sum_{k'=1}^{J} \mathbf{V}(k', j'))$, suponiendo que J es el número de artículos publicados.

El problema principal que tiene la matriz $\mathbf{L}$ es que no tiene por qué ser de transición ya que podría haber "nodos colgantes" lo cual haría que los términos de una fila sumasen 0 y no 1 como debería ocurrir para que $\mathbf{L}$ lo fuese. Otro problema es que no refleja la realidad, en el sentido de que los investigadores no siempre están buscando los artículos por las referencias que tenía el anterior, sino que, a veces, realizan una búsqueda nueva para encontrar un artículo. En esta búsqueda, parece razonable pensar que será más probable que el investigador acuda a una revista "prestigiosa" que a una que no lo sea. Por eso, si el investigador aleatorio se encuentra en el artículo i' irá con una probabilidad $\boldsymbol{\lambda}(i')$ a alguna de las referencias a las que cita este artículo y con una probabilidad $(1 - \boldsymbol{\lambda}(i'))$ a un artículo nuevo, donde la probabilidad de acudir a este artículo vendrá dada por el prestigio de la revista. Esta se obtendrá, a partir de $\boldsymbol{\upsilon}(i') = \boldsymbol{\vartheta}(j)$ si i' pertenece a la revista j, como el cociente $\boldsymbol{\tau}(i') = \boldsymbol{\upsilon}(i')/(\sum_{k'=1}^{P} \boldsymbol{\upsilon}(k'))$, para que las probabilidades de ir a otro artículo sumen uno[8].

Finalmente, hay que tener en cuenta que si un investigador solo está viendo los artículos publicados entre 2010 y 2014 es más probable que busque un nuevo artículo cuando se encuentre en uno publicado en el 2010, que cuando se encuentre en uno publicado en el 2014, ya que el publicado en 2010 tendrá pocas referencias de este periodo, en comparación, pues solo pueden ser del propio 2010. Es por este motivo que se toma como $\boldsymbol{\lambda}(i') = (1 - (2015 - \boldsymbol{\eta}(i'))/10)$, donde $\boldsymbol{\eta}(i')$ es el año en el que se publicó el artículo i'. Finalmente en los "nodos colgantes" como el artículo i' no cita a ninguno otro en este período, la única opción que tiene el investigador aleatorio es ir a un nuevo artículo, es decir $\boldsymbol{\lambda}(i') = 0$.

Realizando los cálculos anteriores, la matriz de transición $\mathbf{A}$ para el ArticleRank tendrá la forma

$$\mathbf{A} = (1/Q)\boldsymbol{\lambda} \cdot \mathbf{e}_M{}^T \cdot \mathbf{L} + (\mathbf{1} - \boldsymbol{\lambda}) \cdot \boldsymbol{\tau}^T,$$

donde $\boldsymbol{\lambda}$ es el vector que almacena con que probabilidad el investigador aleatorio seguirá buscando artículos por las referencias y dependerá del año en el que el último en el que estuvo haya sido publicado, $\mathbf{L}$ es la matriz donde aparece como están relacionados los distintos artículos y $\boldsymbol{\tau}$ es la probabilidad de acudir a un determinado artículo cuando el investigador decide no seguir las referencias del último artículo y dependerá de la "relevancia metodológica" que tengan los posibles artículos a los que puede acudir.

En este caso, la parte $(\mathbf{1} - \boldsymbol{\lambda}) \cdot \boldsymbol{\tau}^T$ garantiza que la matriz de transición $\mathbf{A}$ verifique el Teorema 1 de la Sección 2.2. Por tanto, tomando la distribución estacionaria asociada a $\mathbf{A}$ como índice se podría hacer una clasificación adecuada de los artículos ordenándolos en función de los que se espera que sean visitados más veces.

Una vez se tiene la clasificación de artículos se puede emplear como índice para ordenar la producción de los investigadores o las instituciones con publicaciones en revistas de probabilidad y estadística, la suma de las probabilidades de visitar los distintos artículos publicados por ellas. Para ordenar la "calidad estadística y probabilística" de una revista con el ArticleRank, debido a las asimetrías y gran cantidad de atípicos que presentan los índices bibliográficos, se tomará como índice la mediana de todos los artículos publicados por dicha revista.

Como última consideración para finalizar esta sección, en la descripción del ArticleRank no se han eliminado las auto–citas debido a la dificultad de realizar este método en la práctica con la base de datos que se empleará en la Sección 3. No obstante, para realizar una clasificación de artículos correcta, estas deberían ser eliminadas (o modificadas) con el fin de evitar que las revistas y autores que realizan esta práctica suban en nuestra clasificación. Para ello, bastaría con recalcular las matrices $\mathbf{R}$ y $\mathbf{L}$ modificando en la matriz $\mathbf{Z}_5$ los elementos $\mathbf{Z}_5(i, i) = 0, \forall i \in C$ e igualando en $\mathbf{V}$ los elementos de $\mathbf{V}(i', j') = 0$ si i' está publicado en la misma revista que j' o si i' y j' comparten autor, donde $i', j' \in Q$. También podría ser de interés saber qué ocurre cuando se iguala $\mathbf{V}(i', j') = 0$ si i' y j' comparten institución para ver si existe un efecto de "compañerismo" que haga que artículos de ciertas instituciones salgan más altas en la clasificación.

[8]Aunque no es el objetivo del método presentado en este documento, mencionar que aquí si se quisiese tener en cuenta, además de la "relevancia estadística y probabilística", las citas que reciben los artículos del resto de disciplinas, se podrían tomar los valores de $\boldsymbol{\upsilon}(i')$ en función del número de veces que sea citado i', lo cual correspondería a replicar qué pasaría si el investigador aleatorio escogiese un nuevo artículo en función de las citas que tuviese.

3. CLASIFICACIÓN CON LOS DISTINTOS ÍNDICES

En esta sección se procederá a mostrar la clasificación de las 118 revistas de la temática de probabilidad y estadística con los métodos analizados en Journal Metrics (2015) y en Google Scholar Metrics (2015). Los datos que se mostrarán de los índices Impacto Por Publicación (*IPP*), Impacto por Publicación Normalizado por Artículo (*SNIP*) e índice *SJR* han sido extraídos de Journal Metrics (2015) que utiliza la base de datos de Scopus (2015), donde se recopilan todas las publicaciones *peer–reviewed* publicadas desde 2012 hasta 2014. En cuanto al índice *H*, debido a la limitación de resultados mostrados por Google Scholar Metrics (2015), se ha extraído la información relativa a su clasificación de SCImago (2007) que emplea como base de datos las publicaciones *peer–reviewed* publicadas hasta la actualidad. Aunque solo se muestran los resultados de estos índices, es de esperar que los índices basados en el número de citas de Thomson Reuters (2015) (los factores de impacto de dos y cinco años y el índice de inmediatez) proporcionen resultados similares a los mostrados para el *IPP*, mientras que el índice de influencia de artículo debería presentar resultados similares a los del índice *SJR*.

Para poder aplicar nuestro índice ArticleRank será necesario obtener los datos de citación de los artículos y revistas. Teniendo en cuenta el trabajo elaborado desde Thomson Reuters (2015) para mostrar una base de datos depurada, pero que solo contiene las citas a nivel revista, se ha decidido emplear dicha base de datos para calcular vector $\boldsymbol{\vartheta}$ de las 118 revistas. En cuanto a la elaboración de la matriz $\mathbf{L}$, como Scopus (2015) tiene accesibles todos los datos necesarios de los distintos artículos de las 118 revistas, se ha decidido emplear su base de datos eliminando algunas de las referencias como, por ejemplo, aquellas que tenían el nombre del artículo incompleto, se han empleado los 38473 artículos disponibles y que fueron publicados en alguna revista de probabilidad y estadística entre 2010 y 2014. Es importante resaltar que se deben tomar con cierto cuidado los resultados que se mostrarán ya que, como se comentó en la Sección 2, no se ha tomado ninguna medida para evitar las auto–citas.

Aunque el método ArticleRank permite la clasificación de artículos, autores, instituciones o revistas, con el fin de comparar los resultados obtenidos por el ArticleRank con los otros métodos, solo se mostrará la clasificación de las distintas revistas de temática de probabilidad y estadística. Así, en las Tablas 1 y 2 se muestran las revistas que ocupan los tres primeros cuartiles de la clasificación. Además, con el objetivo de comparar esta ordenación con la dada por otros índices, también se muestra en estas tablas la clasificación con los métodos: Impacto Por Publicación (*IPP*), Impacto por Publicación Normalizado por Artículo (*SNIP*), índice *H* e índice *SJR*; incluyendo aquellos artículos que se encuentran en el primer cuartil con alguno de estos índices.

A continuación, antes de estudiar los resultados de las Tablas 1 y 2, se resumirá que es de esperar que ocurra con los distintos indicadores. Analizando los índices presentados en la Sección 2, cabría esperar que una revista "popular", pero englobada en una temática donde las revistas consiguen muchas citas y presentando referencias de revistas de poca "relevancia" en general o en el ámbito de la probabilidad y estadística, alcance su mejor posición en la clasificación con el índice *IPP*. Si cumple las premisas anteriores, pero estando encasillada en una temática donde las revistas consiguen pocas citas sería de esperar que presentase una clasificación mejor en el *SNIP*. Finalmente, si esta consigue sus citas de revistas "prestigiosas", cuando estas son de carácter aplicado se esperaría que su mejor clasificación la alcanzase en el *SJR*, mientras que cuando estas son de carácter metodológico (en probabilidad y estadística) presentarán una mejor clasificación en el ArticleRank.

Lo primero que se puede ver, observando la clasificación dada en las Tablas 1 y 2, es que revistas de carácter más metodológico (estadístico y probabilístico) como *Theory of Probability and its Applications* (ocupando el puesto 13 del ArticleRank) o *Markov Processes and Related Fields* (29) que se situaban en el cuarto cuartil con el resto de índices, es decir, revistas que no son tan citadas o que reciben pocas citas de revistas que son "importantes" cuando se consideran todas las temáticas, consiguen colocarse en el primer cuartil cuando solo se considera la "relevancia estadística y probabilística" que tienen las revistas que le citan.

Por el contrario, se puede observar en las Tablas 1 y 2 que revistas situadas en el primer cuartil de los otros índices, citadas habitualmente en otros campos por revistas de gran relevancia interdisciplinar, como son *Statistics in Medicine* (49), *Statistical Methods in Medical Research* (55), *Journal of Statistical Software* (57, ocupando el top 3 de revistas con los índices *IPP*, *SNIP* y *SJR*)

y *Multivariate Behavioral Research* bajan al segundo y al tercer cuartil, o incluso, como es el caso de *Fuzzy Sets and Systems* (107), al cuarto cuartil. Además, estos ejemplos sirven para ver que, al menos en este caso, el *SNIP* no consigue el objetivo de ponderar los índices para que pese cuáles son los más próximos a la disciplina, ya que las revistas de carácter más aplicado siguen consiguiendo colocarse en una posición similar a la obtenida con el *IPP*.

Por tanto, parece que cuando el objetivo es ordenar las revistas por las aportaciones metodológicas que presentan en el campo de la estadística y la probabilidad, el método presentado en este documento proporciona buenos resultados para realizar esta clasificación.

AR	Revista	*IPP*	*SNIP*	*H*	*SJR*
1	Journal of the Royal Statistical Society. Series B: Stat Meth	2	3	6	1
2	Biometrika	27	20	9	6
3	Econometrica	3	1	1	24
4	Annals of Statistics	9	6	5	2
5	Journal of the American Statistical Association	13	11	2	4
6	Annals of Probability	32	17	19	11
7	Probability Theory and Related Fields	24	21	24	9
8	Statistical Science	8	9	11	7
9	Technometrics	23	13	18	26
10	Annals of Applied Probability	18	12	20	17
11	Advances in Applied Probability	75	64	36	76
12	Scandinavian Journal of Statistics	65	75	37	31
13	Theory of Probability and its Applications	116	103	90	114
14	Biometrics	26	40	7	20
15	Econometric Theory	53	53	30	14
16	Journal of Computational and Graphical Statistics	34	41	17	19
17	Statistica Sinica	30	24	25	8
18	Annales de l'institut Henri Poincare (B) Probability and Statistics	39	31	57	10
19	Bernoulli	42	26	39	23
20	Stochastic Processes and their Applications	49	37	27	29
21	International Statistical Review	43	14	47	51
22	Finance and Stochastics	16	7	60	12
23	Journal of Applied Probability	76	66	40	74
24	Journal of Business and Economic Statistics	6	4	12	5
25	Scandinavian Actuarial Journal	58	46	105	79
26	Survey Methodology	89	36	108	91
27	Journal of the Royal Statistical Society. Series C: Applied Statistics	46	77	21	40
28	Canadian Journal of Statistics	86	99	50	60
29	Markov Processes and Related Fields	107	97	118	108
30	Bayesian Analysis	15	23	76	36
31	Electronic Journal of Probability	61	67	72	34
32	Journal of Theoretical Probability	88	76	66	62
33	Journal of Multivariate Analysis	55	55	34	35
34	**American Statistician**	69	93	**28**	72
35	**Biostatistics**	**11**	33	**26**	**16**
36	Annals of the Institute of Statistical Mathematics	81	79	64	52
37	**Journal of the Royal Statistical Society. Series A: Stat in Soc**	31	**27**	**22**	55
38	**Combinatorics Probability and Computing**	66	45	55	**27**
39	Infinite Dimensional Analysis, Quantum Probability and Related Topics	77	50	85	65
40	Extremes	40	63	86	43
41	Probability and Mathematical Statistics	114	105	117	107
42	**Annals of Applied Statistics**	**22**	35	61	**28**
43	Lifetime Data Analysis	97	70	62	61
44	**Statistics and Computing**	**29**	38	35	32
45	Electronic Communications in Probability	82	69	82	45
46	Journal of Time Series Analysis	71	84	48	49
47	Test	72	57	73	48

Tabla 1: Clasificación de las revistas en probabilidad y estadística según la ordenación realizada por el ArticleRank (AR) y posición que ocupan en los rankings realizados con los índices Impacto Por Publicación (IPP), Impacto por Publicación Normalizado por Artículo ($SNIP$), índice H e índice SJR. En negrita revistas colocadas en el primer cuartil de alguno de los índices estudiados y que no están en el primer cuartil de la clasificación dada por el ArticleRank.

AR	Revista	*IPP*	*SNIP*	*H*	*SJR*
48	**Electronic Journal of Statistics**	47	47	106	**21**
49	**Statistics in Medicine**	**14**	**28**	**3**	**18**
50	Journal of Official Statistics	78	78	107	96
51	Alea	92	104	115	38
52	ESAIM - Probability and Statistics	105	115	94	66
53	International Journal of Game Theory	90	59	59	69
54	**Journal of Quality Technology**	37	34	**16**	41
55	**Statistical Methods in Medical Research**	**7**	**16**	**23**	**13**
56	Stochastics	87	86	99	75
57	**Journal of Statistical Software**	**1**	**2**	**14**	**3**
58	Journal of Statistical Planning and Inference	73	58	33	47
59	Journal of Nonparametric Statistics	103	94	79	58
60	**ASTIN Bulletin**	**28**	**22**	91	39
61	Sequential Analysis	83	68	100	97
62	Biometrical Journal	51	87	41	50
63	Australian and New Zealand Journal of Statistics	80	82	74	90
64	Statistics and Probability Letters	91	89	38	68
65	**Probabilistic Engineering Mechanics**	**17**	**10**	31	**30**
66	**Econometric Reviews**	44	**15**	67	33
67	Statistica Neerlandica	95	108	77	104
68	**Computational Statistics and Data Analysis**	35	**30**	**15**	37
69	**Econometrics Journal**	36	**18**	89	**22**
70	Statistical Modelling	68	101	92	78
71	Stochastic Models	93	90	78	94
72	Stochastic Analysis and Applications	109	111	68	110
73	**Environmetrics**	**25**	56	42	53
74	Metrika	94	85	69	67
75	Statistics	98	62	87	81
76	British Journal of Mathematical and Statistical Psychology	33	48	51	42
77	Journal of Agricultural, Biological, and Environmental Statistics	50	81	44	77
78	**Journal of Chemometrics**	**20**	51	**13**	98
79	Statistical Methodology	63	49	88	70
80	Environmental and Ecological Statistics	52	96	49	73
81	**Chemometrics and Intelligent Laboratory Systems**	**5**	**19**	**8**	59
82	Methodology and Computing in Applied Probability	59	43	93	84
83	Journal of Statistical Computation and Simulation	79	65	63	86
84	Journal of Biopharmaceutical Statistics	96	110	52	87
85	Stochastics and Dynamics	106	112	97	71
86	**Oxford Bulletin of Economics and Statistics**	41	**25**	32	57
87	Communications in Statistics - Theory and Methods	113	107	53	105
88	**Multivariate Behavioral Research**	**4**	**5**	**29**	**15**
89	Applied Stochastic Models in Business and Industry	62	71	80	88
90	**Statistical Papers**	60	**29**	83	54
91	**Journal of Computational Biology**	**19**	83	**10**	44
105	**Stochastic Environmental Research and Risk Assessment**	**12**	32	46	56
107	**Fuzzy Sets and Systems**	**10**	**8**	**4**	**25**
110	**IEEE/ACM Transactions on Comput Biology and Bioinformatics**	**21**	52	43	63

Tabla 2: (Continuación de la Tabla 1) Clasificación de las mejores revistas en probabilidad y estadística según la ordenación realizada por el ArticleRank (*AR*) y posición que ocupan en los rankings realizados con los índices Impacto Por Publicación (*IPP*), Impacto por Publicación Normalizado por Artículo (*SNIP*), índice *H* e índice *SJR*. En negrita revistas colocadas en el primer cuartil de alguno de los índices estudiados y que no están en el primer cuartil de la clasificación dada por el ArticleRank.

REFERENCIAS

Bergstrom, C. (2007). Measuring the value and prestige of scholarly journals. College & Research Libraries News, 68(5), 314–316.

Bergstrom, C. T., West, J. D. y Wiseman, M. A. (2008). The Eigenfactor metrics. Journal of Neuroscience 28, 11433–11434.

Brin, S. y Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks, 56, 3825–3833.

de Solla Price, D. J. (1986). Little science, big science... and beyond. Columbia University Press.

Durrett, R. (2012). Essentials of stochastic processes. Springer Science & Business Media.

Franceschet, M. (2010). Journal influence factors. Journal of Informetrics, 4, 239–248.

Garfield, E. y Sher, H. (1963). New factors in the evaluation of scientific literature through citation indexing. American Documentation, 14, 195–201.

González–Pereira, B., Guerrero–Bote, V. P. y Moya–Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. Journal of informetrics, 4, 379–391.

Google Scholar Metrics (2015). Google Scholar Metrics. Fecha de consulta 22 de julio de 2015. Disponible en Web: https://scholar.google.com/intl/es/scholar/metrics.html

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences of the United States of America, 102, 16569–16572.

Journal Indicators (2015). CWTS Journal Indicators. Fecha de consulta 22 de julio de 2015. Disponible en Web: http://www.journalindicators.com/

Journal metrics (2015). Elseiver B. V. Journal metrics: Research analytics redefined. Fecha de consulta 22 de julio de 2015. Disponible en Web: http://www.journalmetrics.com/

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. Journal of Informetrics, 4, 265–277.

SCImago (2007). SJR — SCImago Journal & Country Rank. Fecha de consulta 22 de julio de 2015. Disponible en Web: http://www.scimagojr.com

Scopus (2015). Elseiver B. V. Scopus. Fecha de consulta 22 de julio de 2015. Disponible en Web: http://www.scopus.com/

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. British Medical Journal, 314, 498–502.

Thomson Reuters (2015). Journal Citation Reports. Fecha de consulta 22 de julio de 2015. Disponible en Web: http://thomsonreuters.com/

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Predicting using quantile regression methods

Conde-Amboage, M., González-Manteiga, W. and Sánchez-Sellero, C.

Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela

ABSTRACT

Quantile regression methods are considered for computing predictions and prediction intervals of NO_x concentration measured in the surroundings of the power plant of As Pontes (Spain). For these data, smaller prediction errors were obtained by median regression compared to mean regression. A new method to construct prediction intervals based on median regression and bootstrapping the prediction error is proposed. This new method rendered better coverage results for NO_x data, compared to classical and bootstrap prediction intervals for mean regression, and also compared to simpler prediction intervals based on quantile regression. A simulation study is given to illustrate those features that, similarly to these particular data of NO_x concentration, can be found in many other environmental datasets, and will lead to a better performance of the proposed method for obtaining prediction intervals.

Keywords: quantile regression; NO_x concentration; prediction errors; prediction intervals.

1. INTRODUCTION

The power plant of As Pontes (A Coruña, Spain) is one important productive centre property of Endesa Generación S.A. This plant is constituted by a thermal power station and a combined cycle power station. Its activity releases NO_x in quantities that should be controlled both for legal and ecological reasons. On one hand, the legislation imposes some critical levels on NO_x concentration. On the other hand, the location of the power plant near to natural enclaves of high ecological value leads to a special care of its impact on the environment. As a consequence, the power plant possesses several systems of pollution control. In particular, a Network of Vigilance of the Atmospheric Quality has seven automatic analyzers of sulfur dioxide (SO_2), oxides of nitrogen (NO_x), particles in suspension, temperature and oxygen in several positions around the power plant. A meteorological station is also providing information to help the prediction of future contamination. Prediction in 30 minutes in advance is necessary to be made, because it takes about 30 minutes for countermeasures to be arranged in the power plant, and in other contributors to the national power grid in order to compensate their effects on production.

Throughout the years, several methods have been designed to predict future pollution in surrounding areas of the power plant of As Pontes. See for example García-Jurado et al. (1995), Prada-Sanchez et al. (2000), Fernández-Castro and González-Manteiga (2008) and the references therein. It is noticeable that these previous works rely on least squares methods and prediction of pollutant mean levels.

The purpose of the present work is to provide prediction methods for NO_x concentration by means of quantile regression models. Quantile regression models were introduced by Koenker and Bassett (1978), with the purpose of estimating certain quantiles of the response variable conditional to the values of the predictors. In this way, a more complete description of the conditional distribution can be given, where the central and most known quantile is the median, but also including lower or upper quantiles. Thus, these models can describe the effect of the predictors not only on the central values of the response but also on its lower or upper range of values. Moreover, quantile regression is estimated in a more robust manner than common mean regression models, and do not require stringent assumptions, such as homoscedasticity and normality of the error distribution. A thorough description of quantile regression methods can be found in Koenker (2005).

Quantile regression was successfully applied to environmental data by several authors in recent years. Sousa et al. (2009) made use of quantile regression to predict ozone concentrations at Oporto, Northern Portugal. Salama (2005) showed that median regression analysis is more useful to detect the relationship between environmental performance and corporate financial performance than ordinary least squares regression. See also Cade and Noon (2003) for a nice review on applied quantile regression.

In this paper, quantile regression is shown to be more accurate than similar least squares methods in terms of prediction errors for NO_x concentration. A new method for computing prediction intervals is proposed based on quantile regression estimation and bootstrap approximation of the prediction error. Its good performance for our real data example is shown. Some simulations are given to illustrate those features that will make the new method recommendable.

The manuscript is organized as follows. The available data and proposed methods are described in Section 2. Results for the real data example are given in Section 3. Section 4 provides a simulation study to compare the different prediction intervals. Finally, the main conclusions are given in Section 5.

2. DATA AND METHODS

Most of the previous work on pollution around As Pontes power plant was focused on the SO_2 level, since this was the main pollutant of the power plant in its first years, when the combustion of local coal was the main power source. Lately, local coal was replaced by imported coal in the thermal power plant in order to reduce SO_2 emission. This fact, together with a new combined cycle generator, made the concentration of NO_x more relevant. For this reason, in this paper we will focus on the NO_x level.

The concentration of NO_x is measured at every minute and recorded in the automatic monitoring system. Simultaneously, the meteorological station records the temperature, wind speed and wind direction at every minute as well. Our purpose is to predict the concentration of NO_x at a time $(t+30)$ on the basis of the available information at time t (t and $(t+30)$ are measured in minutes). To this end, a regression model will be considered of the following type:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t,grad} + \beta_3 Z_{1t} + \beta_4 Z_{2t} + \beta_5 Z_{3t} + \varepsilon_t \tag{1}$$

where X_t is the NO_x concentration at time t, $X_{t,grad} = X_t - X_{t-5}$ represents the gradient of NO_x concentration in the latest five minutes, Z_{1t}, Z_{2t} and Z_{3t} are the mean values of the temperature, wind speed and wind direction in the latest six minutes (from $(t-5)$ to t), $Y_t = X_{t+30}$ is the NO_x concentration at time $(t+30)$, taken as the response variable, and ε_t represents the error. In this way, measurements in the latest six minutes are used to predict $Y_t = X_{t+30}$. Wind direction is treated as a scalar variable, since we measure the deviation angle from the actual direction to the north.

Model (1) will be fitted with the observations coming from 10 days, that will be used as training sample. The next 10 days will be used as evaluation sample to check the performance

of predictions and prediction intervals. Since this model includes observations from the latest six minutes as predictors, while the response value is scheduled 30 minutes later, data will be divided in blocks of 36 observations, so no predictors or response values are present in two blocks. This way, we prevent from possible autocorrelation. Finally, after removing some missing data, for the training sample we will have a sample size of 338 blocks with corresponding observations $(X_{t-5}, \ldots, X_t, Z_{1t}, Z_{2t}, Z_{3t}, Y_t)$. The evaluation sample will be made of another 338 blocks.

2.1. Least squares versus quantile methods

The regression model given in (1) can be interpreted as a mean regression model if we assume that the error has expectation zero, that is, $E(\varepsilon_t) = 0$. Then, this model can be estimated by minimizing the sum of squared residuals

$$\widehat{\beta}_{LS} = \arg\min_{\beta} \sum_t \left(Y_t - \beta' P_t\right)^2$$

where $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ is the vector of coefficients to be estimated, $P_t = (1, X_t, X_{t,grad}, Z_{1t}, Z_{2t}, Z_{3t})'$ is the vector of predictors, and $\widehat{\beta}_{LS}$ is the least squares estimator.

On the same expression (1), one can consider that the error has τ-quantile equal to zero, that is, $P(\varepsilon_t \leq 0) = \tau$ with $\tau \in (0, 1)$. In that situation, the regression model describes the τ-quantile of Y_t conditional to the predictors. To estimate the coefficients of this new model for each τ, one should minimize the weighted absolute values of the residuals, which can be represented by the following loss function, ρ_τ:

$$\widehat{\beta}(\tau) = \arg\min_{\beta} \sum_t \rho_\tau \left(Y_t - \beta' P_t\right)$$

where

$$\rho_\tau(z) = \begin{cases} (1-\tau) \cdot |z| & \text{if } z \leq 0 \\ \tau \cdot |z| & \text{if } z > 0 \end{cases}$$

To validate the mean regression model we will apply well-known procedures to check whether the residuals satisfy the assumptions of homoscedasticity and normality. The validation of quantile regression models is less studied in the literature. One reason is that quantile regression does not require the assumptions of homoscedasticity and normality. The linearity of the model, which is then the most critical assumption, can be checked by the test developed by Conde-Amboage et al. (2015). This test allows to check the fit of parametric quantile regression models with many predictors.

2.2. Assessment of prediction methods

The pointwise prediction for a value of Y_{t_0} at a future time t_0 using $P_{t_0} = (1, X_{t_0}, X_{t_0,grad}, Z_{1t_0}, Z_{2t_0}, Z_{3t_0})'$ as predictors, can be obtained from the mean regression model by means of $\widehat{Y}_{t_0,LS} = \widehat{\beta}'_{LS} P_{t_0}$. An alternative prediction can be obtained from the median regression model in a similar way, $\widehat{Y}_{t_0,\tau=0.5} = \widehat{\beta}(\tau = 0.5)' P_{t_0}$.

To compare the performance of these two prediction methods we will use two criteria: the mean absolute error (MAE) and the mean squared error (MSE), given by

$$MAE = n_0^{-1} \sum_{t_0} \left| Y_{t_0} - \widehat{Y}_{t_0,m} \right| \qquad MSE = n_0^{-1} \sum_{t_0} \left(Y_{t_0} - \widehat{Y}_{t_0,m} \right)^2$$

where m represents the prediction method, which will be least squares regression (LS) or median regression ($\tau = 0.5$), the times t_0 in the summation are those considered in the evaluation sample,

and n_0 is the evaluation sample size. Recall that the estimations $\widehat{\beta}_{LS}$ and $\widehat{\beta}(\tau = 0.5)$ have been calculated from the training sample.

2.3. Prediction intervals: conditional and unconditional coverage

A prediction interval for a value Y_{t_0} would be an interval that is expected to contain the true value Y_{t_0} with a (presumably) high probability $(1 - \alpha)$, usually called the confidence level. Let us denote a prediction interval as (L_{t_0}, U_{t_0}), where the endpoints L_{t_0} and U_{t_0} are obtained as functions of the training sample and the value of the predictors P_{t_0} at time t_0. Then, it would be expected that

$$P\left(Y_{t_0} \in (L_{t_0}, U_{t_0})\right) = 1 - \alpha.$$

In this expression, the probability is defined over all possible training samples and new observations. We will call this probability as *unconditional coverage*. However, since the value of the predictors at the new observation, P_{t_0}, are known, it is natural to define the above probability conditional to these predictors, that is,

$$P\left(Y_{t_0} \in (L_{t_0}, U_{t_0}) | P_{t_0} = p_{t_0}\right).$$

We call this probability the *conditional coverage*. Note that unconditional coverage can be obtained as an average of the conditional coverage with respect to the predictors distribution. A sample analogue for the unconditional coverage would be the proportion of prediction intervals that contain the new observation in a whole evaluation sample, while the conditional coverage is the same proportion but with evaluation samples taken at a certain value of the predictor P_{t_0}. See Mayr et al. (2012) for a similar explanation of these concepts.

The immediate consequence is that if the conditional coverage respects the nominal level $(1-\alpha)$, then the unconditional coverage will respect it too, while the reverse is not necessarily true. Keeping a conditional coverage in the nominal level $(1 - \alpha)$ is then a more stringent condition, that will need a more detailed use of the information gathered by the predictor P_{t_0}.

2.4. Methods for obtaining prediction intervals

We will consider four already known methods for obtaining prediction intervals: two of them based on mean regression and other two based on quantile regression. A new method is proposed at the end which is based on quantile regression estimation and a bootstrap method.

M1 A prediction interval for Y_{t_0} of level $(1 - \alpha)$ is traditionally obtained from mean regression by

$$\left(\widehat{Y}_{t_0,LS} - t_{n-6,\alpha/2}\,\widehat{\sigma}\,\sqrt{1 + P'_{t_0}(X'X)^{-1}P_{t_0}}\ ,\ \widehat{Y}_{t_0,LS} + t_{n-6,\alpha/2}\,\widehat{\sigma}\,\sqrt{1 + P'_{t_0}(X'X)^{-1}P_{t_0}}\right)$$

where $t_{n-6,\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the Student's t-distribution with $(n - 6)$ degrees of freedom, $\widehat{\sigma}^2 = (n-6)^{-1}\sum_t (Y_t - \widehat{\beta}'_{LS}P_t)^2$ is the error variance estimate based on the training sample, and X is the design matrix of the training sample. This type of interval was introduced from the very beginning of linear mean regression methods (see Seber (1977)) and it is still the most common method to obtain prediction intervals under linear regression models (see Fahrmeir et al. (2013)). The main drawback of this method is that it heavily depends on the assumptions of homoscedasticity and error normality.

M2 Stine (1985) proposed a bootstrap method that allows to skip the error normality in constructing a prediction interval based on mean regression. Homoscedasticity would still be required for their bootstrap method.

Q1 A prediction interval for Y_{t_0} of level $(1-\alpha)$ can be obtained from quantile regression models as

$$\left(\widehat{Y}_{t_0,\tau=\alpha/2}, \widehat{Y}_{t_0,\tau=1-\alpha/2}\right),$$

whose endpoints are estimations of the $\alpha/2$ and $(1-\alpha/2)$ quantiles of Y_{t_0} conditional on the values of the predictors P_{t_0}. Intervals of this kind were employed by several authors, including Meinshausen (2006) and Mayr et a. (2012), where some explanations can be found on the construction of these prediction intervals and their main advantages. Let us mention that they do not require homoscedasticity and adapt to any error distribution. Their drawback is that a parametric (commonly linear) model needs to be assumed at extreme quantiles and that the effect of estimation is not accounted, which leads to an empirical coverage smaller than the nominal one.

Q2 This method consists of a small correction over the previous one, to account for the effect of estimation. The prediction interval is defined as

$$\left(\widehat{Y}_{t_0,\tau=\alpha/2-\delta}, \widehat{Y}_{t_0,\tau=1-\alpha/2+\delta}\right)$$

where $\delta = 0.5(z_{1-\alpha/2}/n)$, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution, and n is the training sample size. The factor 0.5 comes from the expression $\sqrt{\tau(1-\tau)}$, where τ is the quantile to be estimated (here $\tau = 0.5$). This correction was proposed by Zhou and Portnoy (1996).

Q3 A new method is proposed here, where the prediction interval will be computed as

$$\left(\widehat{Y}_{t_0,\tau=0.5} + G^{\star-1}(\alpha/2), \widehat{Y}_{t_0,\tau=0.5} + G^{\star-1}(1-\alpha/2)\right)$$

where $G^{\star-1}(\alpha/2)$ and $G^{\star-1}(1-\alpha/2)$ denote $\alpha/2$ and $(1-\alpha/2)$ quantiles of the bootstrap distribution of the prediction error, respectively, obtained from the following mechanism:

Step 1 Bootstrap replicates of the training sample and new observation are drawn in this way

$$\begin{aligned} Y_t^\star &= \widehat{\beta}(\tau=0.5)'P_t + \varepsilon_t^* \qquad t \in \{1,\ldots,n\} \\ Y_{t_0}^\star &= \widehat{\beta}(\tau=0.5)'P_{t_0} + \varepsilon_{t_0}^\star \end{aligned}$$

$\widehat{\beta}(\tau=0.5)$ is an estimate of the coefficients in the median regression obtained from the training sample. The bootstrap errors are $\varepsilon_t^\star = w_t|r_t|$ where $|\cdot|$ denotes the absolute value and $r_t = Y_t - \widehat{\beta}(\tau=0.5)'P_t$ are the residuals in the original training sample. The weights w_t are drawn from the discrete distribution at the values 1 and -1 with equal probability 0.5. This distribution was proposed by Feng et al. (2011) in the context of wild bootstrap for quantile regression.

The bootstrap error for the new observation is drawn as $\varepsilon_{t_0}^\star = w_{t_0}|r_{t_0}|$, where w_{t_0} follows the same two-points distribution as w_t and the residual r_{t_0} is drawn from the following estimate of the conditional distribution the residuals given the value of the predictor P_{t_0}:

$$F(r|P_{t_0}) = \sum_{t=1}^{n} I(r_t \le r)\, W_{t,P_{t_0}}$$

where $I(r_t \le r)$ is the indicator function with value 1 if the condition $r_t \le r$ is satisfied and 0 otherwise, and

$$W_{t,P_{t_0}} = \frac{K((\widehat{\beta}(\tau=0.5)'P_t - \widehat{\beta}(\tau=0.5)'P_{t_0})/h)}{\sum_{s=1}^{n} K((\widehat{\beta}(\tau=0.5)'P_s - \widehat{\beta}(\tau=0.5)'P_{t_0})/h)}$$

are nonparametric smoothing weights. The smoothing parameter is chosen as $h = n^{-1/5}$ following the suggestion given by Cao-Abad and González-Manteiga (1993).

Step 2 From the bootstrap training sample, a bootstrap replicate of the coefficient estimate can be obtained as $\widehat{\beta}^*(\tau = 0.5)$. Then, bootstrap prediction errors are computed as

$$D^* = Y^{\star}_{t_0} - \widehat{\beta}^*(\tau = 0.5)' P_{t_0}.$$

Step 3 Steps 1 and 2 are repeated B times to obtain an empirical distribution of the differences D^*, and subsequent quantiles $G^{\star -1}(\alpha/2)$ and $G^{\star -1}(1 - \alpha/2)$.

3. RESULTS

Model (1) will be adjusted for mean regression and for regression with different quantiles, using the training sample described in Section 2. The conclusions for the mean and the median regression are similar in the sense that the most significant predictors are X_t, the current concentration of NO_x, and Z_{2t}, the mean value of wind speed. Higher values of the current concentration provide higher predictions for the 30 min future concentration, as it could be expected. Bigger wind speed is associated with lower future concentration. The explanation for this effect is that wind is supposed to carry NO_x away from the power plant surroundings, so bigger wind speed is more effective in this respect.

Quantile regression allows for a more detailed interpretation of the predictors effects on each quantile of the future concentration. Figure 1 shows the estimated coefficients as a function of the quantile order, τ, together with confidence intervals for them. The confidence intervals for the estimated parameters have been calculated by inverting a rank test as described by Koenker (1994). It can be observed that the effect of each predictor is different at each order τ. Looking at the most significant predictor, the current NO_x concentration, X_t, its coefficient is positive for all quantiles, but is larger for larger τ (representing the upper range of the future NO_x concentration). The negative effect of wind speed is less dependent on the particular quantile. Note also that all coefficients show larger confidence intervals for upper quantiles, which is related to the higher variability in the high range of NO_x concentration.

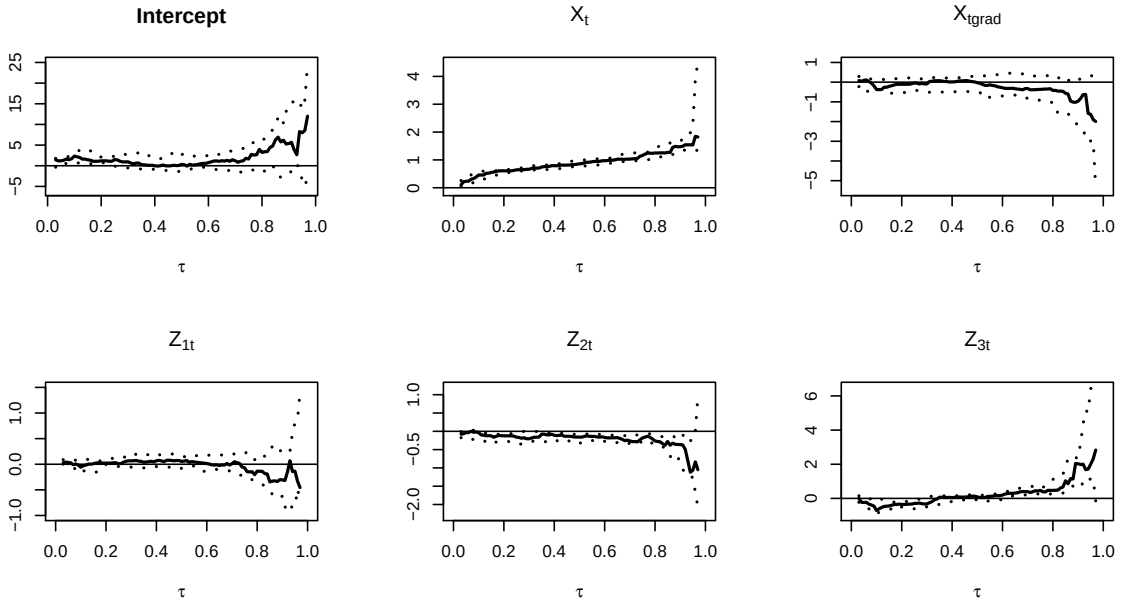


Figure 1: Coefficients associated to each predictor as a function of the order τ of the quantile. The solid line represents the coefficients, while the dotted lines represent the endpoints of confidence intervals for the coefficients.

The validity of the mean regression model was checked in order to explore whether its assumptions are satisfied. Figure 2 contains two pictures. The picture at the left shows a scatter plot of the residuals versus the fitted values by the model. Atypical observations are detected. More variability is found for higher fitted values of the response, which is showing a heteroscedastic pattern. The picture at the right is a QQ-Plot to detect deviations from the normality. There is a clear deviation due to extreme values much larger than expected from the normal distribution. Shapiro-Wilk test of normality showed a highly significant deviation from normality (p-value smaller than $2.2 \times 10^{-}6$). Since the data in the training sample is obtained in subsequent times, autocorrelation could be present. We applied Durbin-Watson test of one-lag autocorrelation and Ljung-Box test of two-lag autocorrelation and no significant autocorrelation was found by any of the two tests (p-value=0.1656 for Durbin-Watson test, p-value=0.2702 for Ljung-Box test).

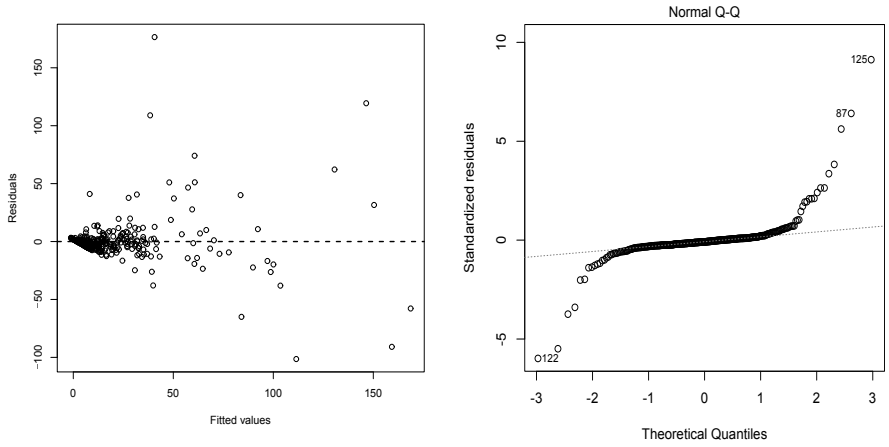


Figure 2: Validation of the mean regression model: Scatterplot of residuals versus fitted values and normal QQ-Plot of standardized residuals.

Regarding the quantile regression model, assumptions of homoscedasticity and error normality are not required. Then, the assumption to be tested would be the linearity, that is, the assumption that the predictors effect can be explained by the linear function given by (1). This assumption was checked using the test proposed by Conde-Amboage et al. (2015). Table 1 contains the p-values associated to this test when checking the linear quantile models at different quantiles. The linear model was accepted for all considered quantiles.

τ	0.025	0.05	0.10	0.5	0.9	0.95	0.975
p-value	0.6702	0.2672	0.1454	0.1332	0.3596	0.3400	0.8322

Table 1: p-values of the linearity test for different quantiles.

Now we will compare the mean regression model with the median regression model by means of the prediction errors obtained for the evaluation sample. Table 2 shows the mean absolute error and the mean squared error for the median and the mean regression models for each of the 10 days of the training sample. The last row of the table contains the mean of the 10 values in each column. We observe that the median regression model provides smaller prediction errors both in the mean absolute error and in the mean squared error.

We have computed prediction intervals by the five methods described in Section 2.4. Table 3 shows the empirical coverage of these prediction intervals, computed as the percentage of times that the real value of NO_x concentration in the evaluation sample belongs to the prediction interval.

	Mean Absolute Error		Mean Squared Error	
	Median	Mean	Median	Mean
Day 1	4.1008	5.0523	35.925	41.570
Day 2	3.1304	4.1902	19.488	33.362
Day 3	8.3023	8.7387	246.563	248.473
Day 4	8.4403	8.8304	151.346	151.759
Day 5	5.0827	6.2487	122.750	126.733
Day 6	2.3943	3.7248	8.982	19.069
Day 7	8.2810	8.5096	487.657	478.338
Day 8	2.9789	4.0173	27.353	31.836
Day 9	2.3820	4.8768	11.525	30.377
Day 10	3.3201	5.1076	42.249	64.301
Average	4.8413	5.9296	115.384	122.582

Table 2: Mean absolute error and mean squared error associated with predictions obtained by median and mean regression models.

This was done for two nominal levels: 90 % and 95 %. In order to evaluate conditional coverage, observations were partitioned in five intervals, I1 to I5, each with the same number of observations, and defined by splitting the ordered Y-values. The purpose is to capture the effect of heteroscedasticity, that seems to be related to the value of the response. The unconditional coverage is shown by means of the average percentage, which is just the mean of the acerages of the five intervals.

It can be observed that both conditional and unconditional coverages are much larger than the nominal level for methods M1, M2, Q1 and Q2, while the proposed method Q3 is quite close to the nominal level at each interval and in the average. The fact that assumptions of normality and homoscedasticity are not satisfied, as shown earlier, might be the cause for the bad behaviour of methods M1, M2, Q1 and Q2. These possible causes will be analyzed in more detail in Section 4.

	Level = 90 %					Level = 95 %				
	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
I1	100.00	97.05	88.23	88.23	91.17	100.00	100.00	94.11	97.05	97.05
I2	100.00	97.05	100.00	100.00	91.17	100.00	100.00	100.00	100.00	94.11
I3	100.00	97.05	100.00	100.00	94.11	100.00	100.00	100.00	100.00	94.11
I4	100.00	97.05	97.05	97.05	88.23	100.00	100.00	100.00	100.00	97.05
I5	100.00	88.23	94.11	94.11	85.29	100.00	100.00	97.05	100.00	91.17
Average	100.00	95.29	95.88	95.88	90.00	100.00	100.00	98.23	99.41	94.70

Table 3: Coverage (in percentages) of prediction intervals obtained by the five methods described in Section 2.4, for levels 90 % and 95 %. Intervals I1 to I5 are defined by splitting the ordered Y-values by the empirical quantiles.

4. SIMULATION STUDY

We have conducted a simulation study to illustrate those features that can be found in data that, similarly to our data example, show a deviation from the common assumptions of the classical linear

models of mean regression, and that will produce inadequate predictions and prediction intervals. In these situations, quantile regression will be a better option for prediction and the proposed method Q3 a good alternative for computing prediction intervals.

Our simulated model will be a linear model with five explanatory variables

$$Y = 1 + X_1 + X_2 + X_3 + X_4 + X_5 + \sigma(X_1, \ldots, X_5)\varepsilon$$

where $X_1, \ldots, X_5$ are independent and each follows an uniform distribution on the unit interval $(0, 1)$, $\sigma(X_1, \ldots, X_5)$ represents the effect of the predictors on the standard deviation of the response, ε is an error random variable, independent of the predictors.

Three types of conditional standard deviations will be considered:

Ho It is a homoscedastic model, where $\sigma(X_1, \ldots, X_5) = 1$.

He1 It is a heteroscedastic model, where $\sigma(X_1, \ldots, X_5) = (1 + X_1 + X_2 + X_3 + X_4 + X_5)/2$. Note that this conditional standard deviation is a linear function of the predictors.

He2 It is a heteroscedastic model, where $\sigma(X_1, \ldots, X_5) = 1 + (X_1 + X_2 + X_3 + X_4 + X_5)^4/100$. Note that this conditional standard deviation is a non-linear function of the predictors.

From this model, training samples of independent observations will be drawn of size n (different values will be considered for n) to estimate both median and mean regression models. For each training sample, an evaluation sample will be drawn of the same size to compute prediction errors obtained by the two estimated models, as well as the empirical coverage of their prediction intervals. One thousand training samples and corresponding evaluation samples will be drawn to compute mean values of the prediction errors and coverage errors. For reasons of brevity, prediction errors will be omitted and only coverages of prediction intervals will be shown.

Table 4 contains the empirical coverage obtained with the homoscedastic model, Ho, for nominal level 90 %, and for different error distributions and sample sizes. The error distributions were chosen to be standard normal, uniform on the interval $(-1, 1)$, chi-squared with two degrees of freedom and standard Cauchy. The classical prediction intervals for linear mean regression, the so-called method M1, provides very accurate results under the standard normal error distribution, while for the other three distributions provides empirical coverage higher than the nominal level. Method M2, which is based on linear regression and a bootstrap approximation of the prediction error, provides accurate coverage for all distributions (with better accuracy for larger sample size), with the only exception of the Cauchy distribution. note that Cauchy distribution does not have mean, which makes the classical estimator of the linear mean regression inconsistent. The two quantile-based methods, Q1 and Q2, show a coverage below the nominal level for all distributions, although this under-coverage goes to zero with increasing sample size. Q2 provides a small improvement with respect to Q1. The proposed method Q3 exhibits an accurate coverage for the four considered distributions, already with a small sample size.

	$n = 100$					$n = 1000$				
ε	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
N(0,1)	90.01	88.41	84.47	85.76	88.50	90.02	89.49	89.47	89.64	89.50
Uniform(-1,1)	93.13	88.17	84.26	85.65	88.76	94.90	89.26	89.44	89.61	89.65
χ^2_2	92.55	90.40	84.24	85.66	88.53	92.90	90.38	89.47	89.63	89.55
Cauchy(0,1)	95.95	92.08	83.56	85.09	88.18	98.76	92.98	89.43	89.60	89.46

Table 4: Coverage (in percentages) of prediction intervals obtained by the five methods described in Section 2.4, with homoscedastic model, Ho, for nominal level 90 %, and for different error distributions and sample sizes.

Table 5 shows the empirical coverage under the heteroscedastic model, He1, for nominal level 90 % and two sample sizes. The error distribution was chosen to be standard normal. This way, it is possible to analyze the specific effect of heteroscedasticity without mixing with the deviation from normality. Intervals I1 to I5 are defined by means of the quantiles of equal probability of the distribution of the linear function $X_1+X_2+X_3+X_4+X_5$, since this is the cause of heteroscedasticity. Coverage for each interval is in a sense an indicator of conditional coverage. Methods M1 and M2 provide inaccurate conditional coverage, where some intervals (I1 and I2) have under-coverage while others (I4 and I5) have over-coverage. This effect does not disappear with increasing sample size. Note that the observations at the first intervals have a smaller conditional standard deviation (easier to be contained in prediction intervals) while the last intervals have a larger one. It is a natural consequence of heteroscedasticity for methods M1 and M2 that assume homoscedasticity. Coverage from methods Q1 and Q2 is somewhat smaller than the nominal level for small sample size, but it converges to the nominal level for increasing sample size. This seems to be natural, since these two methods are based on estimating extreme quantiles by a linear model, which is correct under the linear heteroscedasticity of model He1. Method Q3 provides reasonably accurate coverage at any interval, with a better behavior for a larger sample size.

	$n = 100$					$n = 1000$				
	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
I1	97.13	96.17	82.82	84.26	92.62	97.27	96.95	89.30	89.43	91.28
I2	93.43	91.95	84.29	85.76	90.10	93.64	93.12	89.44	89.63	90.15
I3	90.42	88.61	84.80	86.05	88.54	90.48	89.88	89.53	89.68	89.54
I4	87.03	85.15	84.89	86.38	87.06	87.11	86.50	89.47	89.63	89.02
I5	81.79	79.52	84.86	86.43	84.55	81.75	81.01	89.51	89.67	88.25

Table 5: Coverage (in percentages) of prediction intervals obtained by the five methods described in Section 2.4, with heteroscedastic model, He1, for nominal level 90 % and two sample sizes. I1 to I5 are five intervals of ordered expected values of the response.

In Table 6, empirical coverage is given under the heteroscedastic model, He2, for nominal level 90 % and two sample sizes. The error distribution is standard normal, and the five interval I1 to I5 are constructed as described above for Table 5. The four methods M1, M2, Q1 and Q2 are unable to provide accurate coverage. In the case of methods M1 and M2, the same behavior of over-coverage for the first intervals and under-coverage for the last ones is again observed under this heteroscedastic model, as in the previous one. In the case of methods Q1 and Q2, the general under-coverage of these methods disappears with large sample size, $n = 1000$, and then the non-linearity of the heteroscedasticity gives place to over-coverage in some intervals, I2 to I4, and under-coverage in some other intervals, I1 and I5. Meanwhile, method Q3 is robust to heteroscedasticity of any form, even non-linear, and provides accurate coverage, with better results for larger sample sizes.

5. CONCLUSIONS

Quantile regression methods were considered as an alternative to mean regression for prediction and prediction intervals of NO_x concentration around the power plant of As Pontes, Spain. It was shown that for these data quantile regression provides smaller prediction errors than mean regression. Heteroscedasticity and non-normal error distribution were found in the data, which are deviations from the assumptions of the classical mean regression model, that could explain the better performance of quantile methods.

Two known methods from mean regression (classical and bootstrap-based) and two known methods from quantile regression were considered for obtaining prediction intervals. Due to the

	$n = 100$					$n = 1000$				
	M1	M2	Q1	Q2	Q3	M1	M2	Q1	Q2	Q3
I1	98.40	97.38	78.02	79.53	92.16	98.68	98.02	85.12	85.34	90.51
I2	96.75	95.20	86.95	88.08	91.44	97.22	96.26	92.38	92.52	90.73
I3	94.22	92.04	88.18	89.39	90.49	94.49	93.20	92.84	92.96	90.28
I4	88.92	86.27	87.35	88.76	87.67	89.24	87.47	91.60	91.75	89.34
I5	74.68	71.39	81.08	82.65	81.12	74.61	72.42	85.34	85.54	86.96

Table 6: Coverage (in percentages) of prediction intervals obtained by the five methods described in Section 2.4, with heteroscedastic model, He2, for nominal level 90 % and two sample sizes. I1 to I5 are five intervals of ordered expected values of the response.

special features of our data example, an additional method was proposed based on quantile regression estimation and bootstrap approximation of the prediction error. A better performance was observed in the real data example from prediction intervals obtained by this new method. A simulation study is provided to show how the deviations from the assumptions of homoscedasticity and normality affect the already-known methods for computing prediction intervals. The coverage accuracy of the new method is shown in the considered scenarios.

ACKNOWLEDGEMENTS

The authors acknowledge the support of Project MTM2013-41383P, from the Spanish Ministry of Economy and Competitiveness, including support from 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

Cade, B.S. and Noon, B.R. (2003) A gentle introduction to quantile regression for ecologists. Frontiers in Ecology and the Environment, 1, 412–420.

Cao-Abad, R. and González-Manteiga, W. (1993) Bootstrap methods in regression smoothing. Journal of Nonparametric Statistics, 2, 379–388.

Conde-Amboage M, Sánchez-Sellero C and González-Manteiga W. (2015) A lack-of-fit test for quantile regression models with high-dimensional covariates. Computational Statistics and Data Analysis, 88, 128–138.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013) Regression: Models, Methods and Applications. Springer.

Feng X, He X and Hu J. 2011. Wild bootstrap for quantile regression. Biometrika, 98, 995–999.

Fernández-Castro, B.M. and González-Manteiga. W. (2008) Boosting for real and functional samples: an application to an environmental problem. Stochastic Environmental Research and Risk Assessment, 22, 27–37.

García-Jurado, I., Gonzalez-Manteiga, W., Prada-Sánchez, J.M., Febrero-Bande, M. and Cao, R. (1995) Predicting using Box-Jenkins, nonparametric and bootstrap techniques. Technometrics, 37, 303–310.

Koenker, R. (1994). Confidence intervals for regression quantiles. Proceedings of the 5th Prague Symposium on Asymptotic Statistics, New York: Springer-Verlag, 349–359.

Koenker, R. (2005) Quantile Regression. Cambridge University Press.

Koenker, R. and Bassett, G. (1978.) Regression quantiles. Econometrica, 46, 33–50.

Mayr, A., Hothorn, T. and Fenske, N. (2012) Prediction intervals for future BMI values of individual children - a non-parametric approach by quantile boosting. BMC Medical Research Methodology, 12, 6.

Meinshausen, N. (2006) Quantile Regression Forests. Journal of Machine Learning Research, 7, 983–999.

Prada-Sánchez, J.M., Febrero-Bande, M., Cotos-Yáñez, T., González-Manteiga, W., Bermúdez-Cela, J.L. and Lucas-Domínguez, T. (2000) Prediction of SO_2 pollution incidents near a power station using partially linear models and a historical matrix of predictor-response vectors. Environmetrics, 11, 209–225.

Salama, A. (2005) A note on the impact of environmental performance on financial performance. Structural Change and Economic Dynamics, 16, 413-421.

Seber, G.A.F. (1977) Linear Regression Analysis. Wiley.

Sousa, S.I.V., Pires, J.C.M., Martins, F.G., Pereira, M.C. and Alvim-Ferraz, M.C.M. (2009) Potentialities of quantile regression to predict ozone concentrations. Environmetrics,20, 147–158.

Stine, R.A. (1985) Bootstrap prediction intervals for regression. Journal of the American Statistical Association, 80, 1026–1031.

Zhou KQ and Portnoy SL. 1996. Direct use of regression quantiles to construct confidence sets in linear models. The Annals of Statistics, 24, 287–306.

Comunicacións orais

Sesión 2 de Investigación de Operacións

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

STRATEGY-PROOFNESS AND MONOTONICITY FOR LAND RENTAL AGREEMENTS

A. Valencia-Toledo[1] and J. Vidal-Puga[2]

[1] [2]Universidade de Vigo

ABSTRACT

We consider a land rental problem where there are one single tenant and several lessors who do not necessary need all the available land. A rule should determine which lessors rent their land, in which amount and at which price. Our first result is a complete characterization of the family of rules which satisfy land reassignment and monotonicity. We define two relevant subfamilies. The first one is characterized by adding a property of weighted standard for two-person. The second one is characterized by adding consistency and continuity.

Keywords: land rental, land reassignment, land monotonicity, consistency, standard for two-person, game theory.

1 INTRODUCTION

Given the importance of land in economic environments for the vast majority of human activities, its exploitation has generated social tensions and disputes in the globe. The management of the land and natural resources is one of the most critical challenges facing developing countries today (UN-HABITAT, 2012). People and countries have tried to solve these conflicts through rules or laws. For instance, natural resource exploitation is an industrial activity that has been generating conflicts between firms and indigenous communities in many countries in South America, Africa and Asia. For instance, a recent case about activities of mining industry in Guatemala, India and Peru was reported by Sosa (2011), a particular restitution problem in Colombia where there is an indivisible good (land or property) over which two agents have rights was studied by Jaramillo et al. (2014), and mining conflict in Peru was studied by Dell (2010). In these land conflicts there exists a right over the land from both sides: On the one hand, Free Prior and Informed Consent (FPIC) is the principle that indigenous communities have the right to give or withhold its consent to proposed projects that may affect the land they customarily own, occupy or otherwise use.[1] On the other hand, the mining firm or company may has a investment or concession over those lands, as it is the case for instance in Peru. In order to solve these land conflicts there is a government or planner who needs to have all relevant information about both sides. In these cases, the identification and demarcation of the indigenous land is not clear. This situation can lead to manipulation by splitting or merging of communities or commoners, due to the fact they may have incentives to strategically misrepresent their identity in order to influence the final outcome to their own advantage. The government or planner seeks to solve this situation by assigning a price and amount of land fairly, stably and efficiently, and at the same time, to seek rules that guarantee non-manipulability.

We can also consider other situations with similar characteristics, which can be modeled in the same way, such as construction of facilities (hotels, malls, convention centers, schools, resorts, etc.), urbanisation in populated areas, and settlement of governmental bases (Valencia-Toledo and Vidal-Puga, 2015). In all of these situations there are several land commoners or owners, and a single agent who is interested in their land.

[1]Article 10 of the United Nations Declaration on the Rights of Indigenous People.

We model this situation as a land rental problem from a cooperative game perspective, where there is a single tenant who can be a mining firm, and several lessors who can be a group of commoners. Each commoner has some available amount of land c_i and they give a valuation from their land which we call as reservation price per unit of land r, that for simplicity we consider equal for all of them. The mining firm needs to obtain an optimal amount of land E, which is a completely divisible object. In this sense, the mining firm and commoners assure an optimal profit K over E.

A rule determines, for each problem, a quantity of land which is given by each community and a price that the mining firm must pay for unit of land as a way of compensation to the communities.

In order to study rules that guarantee non-manipulability, we propose a version of strategy-proofness such that we seek stability through land reassignment i.e. lessors should not find it profitable to re-assign the land among them. Also, we propose a version of land monotonicity that seeks fairness, in the sense that an increase in the quantity of land affects positively the final profits to both sides.

Our first result is a characterisation of the rules that satisfy land reassignment and land monotonicity. By adding a weighted standard for two-person property inspired in Hart and Mas-Colell (1989), we characterize a relevant family of rules. Analogously, by adding consistency (we can see in (Thomson, 2008)), and by adding continuity we characterize another relevant family of rules.

Recently, Massó et al. (2015) study efficiency and fairness properties of the equal cost sharing with maximal participation mechanism in the provision of a binary and excludable public good. There are three basic requirements such as strategy-proofness, individual rationality and no-deficit, which says that at every type profile the cost of the club is covered by agents' payments. Our point of view differs with Massó et al. (2015) about incompatibility of efficient with strategy-proofness and individual rationality. Brandt (2015) studies the strategic manipulation of set-valued social choice functions according to Kelly's preference extension (Kelly, 1977), and proposes a new variant of Maskin-monotonicity (Maskin, 1999) for set-valued Social Choice Functions called set monotonicity. Ju (2013) studies allocation rules that are robust to coalitional manipulation by transferring, merging, or splitting individual characteristics among coalition partners, as coalitional manipulation on networks (e.g. merging or splitting claims in bankruptcy problems). Moulin (2007) studies manipulation by merging, splitting, and transferring is in the rationing problem and proves that mergeproofness and splitproofness are not compatible, and that it is similarly impossible to prevent all transfers of jobs involving three or more agents. Kojima and Manea (2007) study strategy-proofness in the probabilistic serial mechanism in large random assignment problems. Ju (2003) characterizes (division) rules that are non-manipulable via (pairwise) splitting in claim problems. Sun and Yang (2003) study a general problem of n indivisible objects with a certain amount of money to n persons with a requirement that each person be assigned with one object and their main result is that the optimal fair allocation mechanism achieves simultaneously efficiency, fairness (a fair allocation is compatible with the vector of maximum compensations if it is monotone) and non-manipulability. Moulin and Shenker (2001) investigate strategyproof mechanisms that elicit individual willingness to pay, decide who is served, and then share the cost among them.

In Section 2, we present the model, in Section 3, we study and characterize a family of rules with land reassignment and monotonicity properties, in Section 4, we characterize a relevant family of rules that also satisfies weighted standard for two-person. Finally. in Section 5, we characterize a relevant family of rules that also satisfy consistency and continuity.

2 THE MODEL

Let $\mathbb{N} = \{1, 2, \dots\}$ be the set of potential land lessors and $\mathscr{N}$ be the set of all non-empty finite subset of $\mathbb{N}$ i.e $\mathscr{N} = \{N \subset \mathbb{N} : 0 < |N| < \infty\}$. Let $N = \{1, 2, \dots, n\}$ be a generic set of land lessors. Given $y \in \mathbb{R}^S$, we denote $y(S) = \sum_{i \in S} y_i$ for any finite set S, and $y_S \in \mathbb{R}^S$. Given $x, y \in \mathbb{R}^S$, we write $x \leq y$ when $x_i \leq y_i$ for all $i \in S$. 0_S denotes the vector $(0, \dots, 0) \in \mathbb{R}^S$ for any finite set S. We denote the set of nonnegative real numbers as $\mathbb{R}_+$.

A *land rental problem* is a tuple (N_0, μ, c, r) where $N_0 = \{0\} \cup N$ is the set of players with 0 the special and unique player called land tenant and N the set of land lessors, $c \in \mathbb{R}^N$ is the

vector whose coordinates $c_i > 0$ for all $i \in N$ represent the amount of available land for each lessor, $r \in \mathbb{R}_+$ is the reservation price per unit of land for lessors, and $\mu : [0, c(N)] \to \mathbb{R}_+$ is a non-decreasing function that depends on the amount of land and represents the profit of tenant. We assume $\mu(0) = 0$, and there exists a unique $E \in]0, c(N)]$ such that $\mu(E) + r(c(N) - E)$ is maximum.[2] [3] We then denote $K = \mu(E)$ as the optimal profit that the players can obtain. We assume $K \geq rE$ to have benefit of cooperation.

Under these conditions, an efficient allocation implies that the amount of land is E and the aggregate profit is K. Only E and K are relevant. Furthermore, for convenience we use N instead of N_0. Hence, we denote the land rental problem as (N, K, E, c, r) instead of (N_0, μ, c, r).

The set of feasible agreements on N is given by $A^N = \{(x, p) : (x, p) \in \mathbb{R}^N \times \mathbb{R}_+, 0_N \leq x \leq c\}$. Let $\mathscr{A} = \bigcup_{N \in \mathscr{N}} A^N$ be the set of all feasible agreements.

Given $(x, p) \in A^N$, the utility for player 0 and each player $i \in N$ are $u_0(x, p) = \mu(x(N)) - px(N)$ and $u_i(x, p) = (p - r)x_i$ for all $i \in N$ respectively. Let $\mathcal{L}$ the set of all land rental problems.

We define *a rule* as a function $\psi : \mathcal{L} \longrightarrow \mathscr{A}$ that assigns to each problem $(N, K, E, c, r) \in \mathcal{L}$ a pair $(x, p) \in A^N$ i.e. $\psi(N, K, E, c, r) \in A^N$, satisfying:

(i) $x(N) = E$.

(ii) $p(N, \alpha K, \beta E, \beta c, \frac{\alpha}{\beta} r) = \frac{\alpha}{\beta} p(N, K, E, c, r)$ and $x_i(N, \alpha K, \beta E, \beta c, \frac{\alpha}{\beta} r) = \beta x_i(N, K, E, c, r)$ for all $\alpha, \beta > 0$ and $i \in N$.

(iii) $r \leq p \leq \frac{K}{E}$.

Condition i) is *efficiency* that determines the optimal amount of land. Condition ii) is *scale invariance* that determines, that any change in the currency and land units do not change the final price and amount of land respectively. And, condition iii) is *individual rationality* that determines, that the lessors get at least zero of the benefit of cooperation when the price is greater or equal than r, and under efficiency, the tenant gets at least zero of the benefit of cooperation when the price is less or equal than $\frac{K}{E}$.

Let Ψ is a set of all rules ψ.

There exist two special classes of rules: On the one hand, a rule is *optimal for tenant* when the price is given by $p = r$. On the other hand, a rule is *optimal for lessors* when the price is given by $p = \frac{K}{E}$.

3 LAND REASSIGNMENT AND MONOTONICITY

Since the indigenous communities have customary land, it is not clear the demarcation and official registration of their land. In this sense, if they can reach an agreement of reallocating their land in order to share extra benefits so created under a rule, for the planner is not possible to see this real share extra benefits, and it may be hard to get the outcome the rule is supposed to attain. In our context manipulation implies that the lessors will benefit by merging or splitting under reallocating their land. Our aim is to fully identify rules that are free from this concern as formulated previously. We formalise this property as follows.

Land reassignment (LR) Given $(N, K, E, c, r), (N', K, E, c', r) \in \mathcal{L}$ such that $c_i = c'_i$ for all $i \in S = N \cap N'$ and $c(N \setminus S) = c'(N' \setminus S)$, then

$$\sum_{i \in N \setminus S} u_i(\psi(N, K, E, c, r)) = \sum_{i \in N' \setminus S} u_i(\psi(N', K, E, c', r)).$$

In this property, if the right-hand side of expression is larger than the left-hand side, lessors in $N \setminus S$ can gain by reallocating their land, and S is the set of lessor that remain unchanged.

The following property in general says that an increase of the available land is beneficial. In this context, when there is a lessor who increases her available land and additionally the rest of lessors remain unchanged, land monotonicity says that the tenant and this lessor will not be worse off. We formalise this property as follows.

[2] Since $rc(N)$ is constant, this condition is equivalent to says $\mu(E) - rE$ is maximum.

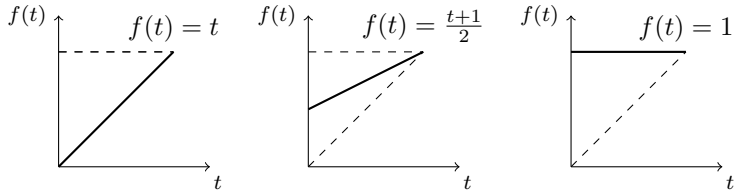
[3] μ is three times continuously differentiable and its third derivative is strictly negative or second derivative is negative.

Land-monotonicity (LM) Given $(N, K, E, c, r), (N, K, E, c', r) \in \mathcal{L}$, we have: (i) $c \leq c'$ implies $u_0(\psi(N, K, E, c, r)) \leq u_0(\psi(N, K, E, c', r))$, and (ii) for each $i \in N$, $c_i \leq c'_i$ and $c_j = c'_j$ for all $j \in N \setminus \{i\}$ imply $u_i(\psi(N, K, E, c, r)) \leq u_i(\psi(N, K, E, c', r))$.

In this property, the tenant will be better off when there are more available land, and when only one lessor has more available land and the rest of lessors remain unchanged, this lessor will be better off.

In order to characterize the family of rules that satisfy these properties, we define F as the set of functions $f : [0, 1] \to [0, 1]$ with $f(t) \geq t$ for all $t \in [0, 1]$. Now, we consider the family of rules that satisfy $p = \frac{K}{E} f(\frac{rE}{K})$ for some $f \in F$ and $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. So, we obtain different rules because it is depending on the function $f \in F$.

In Figure 1 we see three examples of functions that determine rules optimal for the tenant, lessors and an intermediate rule.



(a) (b) (c)

Figura 1: Examples of functions that determine rules such as an optimal rule for the tenant in (a), and an optimal rule for the lessors in (c), and an intermediate rule in (b).

Theorem 1. *A rule ψ satisfies LR and LM if and only if there exists a function $f \in F$ such that the price is given by $p = p(K, E, r) = \frac{K}{E} f\left(\frac{rE}{K}\right)$ and, when $p \neq r$, the amount of land is given by $x_i = x_i(c, E) = \frac{c_i E}{c(N)}$ for all $i \in N$.*

Proof. ($\Leftarrow$) Let ψ be a rule given by $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$ and, when $p \neq r$, $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. We will prove that ψ satisfies LR and LM.

In order to prove that ψ satisfies LR, let $(N, K, E, c, r), (N', K, E, c', r) \in \mathcal{L}$ and $S = N \cap N'$ given as the definition of LR. We will check that $\sum_{i \in N \setminus S} u_i(\psi(N, K, E, c, r)) = \sum_{i \in N' \setminus S} u_i(\psi(N', K, E, c', r))$. Let $t = \frac{rE}{K}$ be a value in $[0, 1]$. Since the left part of the expression $\sum_{i \in N \setminus S} u_i(\psi(N, K, E, c, r))$ by definition and hypothesis is equal to $\sum_{i \in N \setminus S} \left(\frac{K}{E} f(t) - r\right) \frac{c_i E}{c(N)}$, and this equal to $K\left(f(t) - t\right)\left(1 - \frac{c(S)}{c(N)}\right)$. Analogously from the right part of the expression $\sum_{i \in N' \setminus S} u_i(\psi(N', K, E, c', r)) = K\left(f(t) - t\right)\left(1 - \frac{c(S)}{c(N')}\right)$. Since $c(N \setminus S) = c(N' \setminus S)$ and $c_i = c'_i$ for all $i \in S$, we have that $c(N) = c(N')$. Hence the last two expressions coincide and hold the condition.

In order to prove that ψ satisfies LM, let $(N, K, E, c, r), (N, K, E, c', r) \in \mathcal{L}$ given as the definition of LM. We will check that: $c \leq c'$ implies that $u_0(\psi(N, K, E, c, r)) \leq u_0(\psi(N, K, E, c', r))$, and $c_i \leq c'_i$ and $c_j = c'_j$ for all $j \in N \setminus i$ imply that $u_i(\psi(N, K, E, c, r)) \leq u_i(\psi(N, K, E, c', r))$ for $i \in N$. If $c \leq c'$ then $u_0(\psi(N, K, E, c, r))$ and $u_0(\psi(N, K, E, c', r))$ by efficiency is equal to $K - \frac{K}{E} f\left(\frac{rE}{K}\right) E$, hence satisfy the first condition. If $c_i \leq c'_i$ and $c_j = c'_j$ for all $j \in N \setminus \{i\}$ then $u_i(\psi(N, K, E, c, r)) = (\frac{K}{E} f\left(\frac{rE}{K}\right) - r)\frac{c_i E}{c(N)} \leq (\frac{K}{E} f\left(\frac{rE}{K}\right) - r)\frac{c'_i E}{c(N)} = u_i(\psi(N, K, E, c', r))$ for $i \in N$.

($\Rightarrow$) Let ψ be a rule that satisfies LR and LM. We write (x, p) instead of $\psi(N, K, E, c, r)$, (x', p') instead of $\psi(N', K, E, c', r)$ and so on. Furthermore, we write u_i instead of $u_i(x, p)$ and u'_i instead of $u_i(x', p')$ and so on. We proceed by series of claims.

Claim 1. *If $K = E = 1$ and $N = \{1\}$, then the price does not depends on c.*

Proof. By LM if $c_1 \leq c'_1$ then $u_0 \leq u'_0$, by definition and efficiency we have $1 - p \leq 1 - p'$, hence $p \geq p'$ (a higher c_1 value, higher price p). Analogously, by LM if $c_1 \leq c'_1$ then $u_1 \leq u'_1$ from this we obtain that $p \leq p'$ (a higher c_1 value, lower price p). Therefore from last two result we have that $p = p'$. ■

We define $f(t) = p(\{1\}, 1, 1, (1), t)$, and since $0 \leq p(\{1\}, 1, 1, (1), t) \leq 1$ and $p(\{1\}, 1, 1, (1), t) \geq t$ i.e. $f(t) \geq t$, by individual rationality, $f \in F$.

Claim 2. *If $K = E = 1$, then $p = f(r)$.*

Proof. By LR $u(N)$ is equal to $u_1(\psi(\{1\}, 1, 1, (c(N)), r))$, by definition, Claim 1 and efficiency, this is equal to $p(\{1\}, 1, 1, (1), r) - r$, hence $u(N) = f(r) - r$. Furthermore by definition and efficiency $u(N) = \sum_{i \in N}(p - r)x_i = p - r$. Therefore, we have $p(N, 1, 1, c, r) = f(r)$. ■

Claim 3. $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$.

Proof. By scale invariance $p(N, K, E, c, r) = \frac{K}{E} p\left(N, 1, 1, \left(\frac{c}{E}\right), \frac{rE}{K}\right)$, and by Claim 2 we have that $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$. ■

Therefore, the price is given by Claim 3. Now we focus to study about amount of land x.

Claim 4. *If $p \neq r$ and $i, j \in N$ such that $c_i = c_j$, then $x_i = x_j$.*

Proof. We fix a lessor $\alpha \in \mathbb{N} \setminus N$. We define $(N^{i\alpha}, K, E, c^{i\alpha}, r) \in \mathcal{L}$, where $N^{i\alpha} = \{i, \alpha\}$ and $c^{i\alpha} = (c_i, c(N \setminus \{i\})$. Since $N \cap N^{i\alpha} = \{i\}$, then by LR, $u(N \setminus \{i\}) = u^{i\alpha}(N^{i\alpha} \setminus \{i\})$ and by Claim 3 and $p \neq r$ we get $x(N \setminus \{i\}) = x_\alpha^{i\alpha}$. Furthermore, by efficiency $x(N \setminus \{i\}) + x_i = E$ and $x_i^{i\alpha} + x_\alpha^{i\alpha} = E$. So, from these last three equalities we obtain that $x_i = x_i^{i\alpha}$. We define $(N^{j\alpha}, K, E, c^{j\alpha}, r) \in \mathcal{L}$, where $N^{j\alpha} = \{j, \alpha\}$, and $c_j^{j\alpha} = c_i^{i\alpha} = c_i = c_j$ and $c_\alpha^{j\alpha} = c_\alpha^{i\alpha}$. Since $N^{i\alpha} \cap N^{j\alpha} = \{\alpha\}$ and $c_i^{i\alpha} = c_j^{j\alpha}$, then by LR, $u_i(x^{i\alpha}, p^{i\alpha}) = u_j(x^{j\alpha}, p^{j\alpha})$, and by Claim 3 and $p \neq r$, we obtain $x_i^{i\alpha} = x_j^{j\alpha}$. Since $N^{j\alpha} \cap N = \{j\}$ and $c_\alpha^{j\alpha} = c(N \setminus \{j\})$, then by LR, $u_\alpha(x^{j\alpha}, p^{j\alpha}) = u(N \setminus \{j\})$, and by Claim 3 and $p \neq r$, we obtain $x_\alpha^{j\alpha} = x(N \setminus \{j\})$. Furthermore by efficiency we have $x_j^{j\alpha} + x_\alpha^{j\alpha} = E$ and $x_j + x(N \setminus \{j\}) = E$. So, from these last three equalities we obtain $x_j^{j\alpha} = x_j$. Then, from $x_i = x_i^{i\alpha}$, $x_i^{i\alpha} = x_j^{j\alpha}$ and $x_j^{j\alpha} = x_j$ we get that $x_i = x_j$. ■

Claim 5. *If $N = \{i, j\}$, $p \neq r$ and $c_i, c_j \in \mathbb{Q}$, then $x_i = \frac{c_i E}{ci + c_j}$ and $x_j = \frac{c_j E}{ci + c_j}$.*

Proof. Let $N^i, N^j \subset \mathbb{N} \setminus N$ with $N^i \cap N^j = \emptyset$, $|N^i| = a_i b_j$ and $|N^j| = a_j b_i$. Given $c_i = \frac{a_i}{b_i}$ and $c_j = \frac{a_j}{b_j}$ where a and b are non negative integers. We define $(N^{*i}, K, E, c^{*i}, r) \in \mathcal{L}$ with $N^{*i} = N^i \cup \{j\}$ and $c_k^{*i} = \frac{1}{b_i b_j}$ for all $k \in N^i$ and $c_j^{*i} = c_j$. Since $N \cap N^{*i} = \{j\}$ and $c^{*i}(N^{*i} \setminus j) = c_i$, then by LR, $u_i(x, p) = u^{*i}(N^{*i} \setminus \{j\})$, by definition, $p \neq r$ and Claim 3 it is equal to write $(p - r)x_i = (p - r)x^{*i}(N^{*i} \setminus \{j\})$, hence $x_i = x^{*i}(N^{*i} \setminus \{j\})$. We define $(N^{*ij}, K, E, c^{*ij}, r) \in \mathcal{L}$ with $N^{*ij} = N^i \cup N^j$ and $c_k^{*ij} = \frac{1}{b_i b_j}$ for all $k \in N^{*ij}$. Since $c^{*i}(N^{*i} \setminus \{j\}) = c^{*ij}(N^{*ij} \setminus N^j)$, $N^{*i} \cap N^{*ij} = N^i$ and $c^{*ij}(N^{*ij} \setminus N^i) = \sum_{l=1}^{a_j b_i} \frac{1}{b_i b_j} = c_j = c_j^{*i}$, then by LR, $u_j(x^{*i}, p^{*i}) = u^{*ij}(N^{*ij} \setminus N^i)$, by definition, $p \neq r$ and Claim 3 it is equal to write $(p - r)x_j^{*i} = (p - r)x^{*ij}(N^{*ij} \setminus N^i)$, hence $x_j^{*i} = x^{*ij}(N^{*ij} \setminus N^i) = x^{*ij}(N^j)$. On the one hand, by efficiency $x_i + x_j = E$ and $x^{*i}(N^{*i} \setminus j) + x_j^{*i} = E$, and with $x_i = x^{*i}(N^{*i} \setminus j)$ and $x_j^{*i} = x^{*ij}(N^j)$, we obtain that $x_j = x^{*ij}(N^j)$. On the other hand, by efficiency $x^{*i}(N^{*i} \setminus j) + x_j^{*i} = E$ and $x^{*ij}(N^i) + x^{*ij}(N^j) = E$, and with $x_i = x^{*i}(N^{*i} \setminus j)$ and $x_j^{*i} = x^{*ij}(N^j)$ we obtain that $x_i = x^{*ij}(N^i)$. We take $(N^{*ij}, K, E, c^{*ij}, r)$ and (N, K, E, c, r) with $N^{*ij} \cap N = \emptyset$ and $c^{*ij}(N^{*ij}) = c(N)$, then by LR, $u^{*ij}(N^{*ij}) = u(N)$, by $p \neq r$ and Claim 3 it is equal to write $x^{*ij}(N^{*ij}) = x(N)$, by Claim 4 and efficiency, we obtain that $x_k^{*ij} = \frac{E}{|N^{*ij}|}$ and it is same as $x_k^{*ij} = \frac{E}{a_i b_j + a_j b_i}$ for $k \in N^{*ij}$. Then by efficiency and $x_j = x^{*ij}(N^j)$ we have that $x_i = E - x^{*ij}(N^j)$, by Claim 4 it is equal to write that $x_i = E - a_j b_i x_k^{*ij}$. Since $x_k^{*ij} = \frac{E}{a_i b_j + a_j b_i}$, then we have that $x_i = E - \frac{a_j b_i E}{a_i b_j + a_j b_i} = \frac{c_i E}{c_i + c_j}$. Analogously $x_j = \frac{c_j E}{ci + c_j}$. ■

Claim 6. *If $N = \{i, j\}$, $p \neq r$ and $c_j \in \mathbb{Q}$, then $x_i = \frac{c_i E}{c_i + c_j}$ and $x_j = \frac{c_j E}{c_i + c_j}$.*

Proof. If we consider $c_i + c_j = E$, then $\frac{c_i E}{c_i + c_j} = c_i$ and $\frac{c_j E}{c_i + c_j} = c_j$, by efficiency $x_i = c_i$ and $x_j = c_j$ therefore $x_i = \frac{c_i E}{c_i + c_j}$ and $x_j = \frac{c_j E}{c_i + c_j}$. We now focus to prove for $c_i + c_j > E$. Let $\{c_i^s\}_{s=1}^\infty$ be a decreasing sequence of rational numbers that converges to c_i. In each $(N^s, K, E, c^s, r) \in \mathcal{L}$ with $N^s = N$ and $c^s = (c_i^s, c_j)$. By Claim 5 for x_i and x_j we have $x^s = \left(\frac{c_i^s E}{c_i^s + c_j}, \frac{c_j E}{c_i^s, c_j}\right)$. Then by LM $u_i(x, p) \leq u_i(x^s, p^s)$, by Claim 3 it is equal to write that $(p - r)x_i \leq (p - r)x_i^s$, since

$p \neq r$, this is same as $x_i \leq x_i^s$, by Claim 5 it equal to write that $x_i \leq \frac{c_i^s E}{c_i^s + c_j}$, then we have that $x_i \leq \frac{c_i E}{c_i + c_j}$. Let $\{\widehat{c}_i^s\}_{s=1}^{\infty}$ an increasing sequence of positive rational numbers that converges to c_i, this is possible since that $c_i + c_j > E$. Since $c_i > 0$ there exist each $(\widehat{N}^s, K, E, \widehat{c}^s, r) \in \mathcal{L}$ with $\widehat{N}^s = N$ and $\widehat{c}^s = (\widehat{c}_i^s, c_j)$. By Claim 5 for x_i and x_j, we have $\widehat{x}^s = \left(\frac{\widehat{c}_i^s E}{\widehat{c}_i^s + c_j}, \frac{c_j E}{\widehat{c}_i^s + c_j}\right)$. Then by LM $\widehat{c}_i^s \leq c_i \Rightarrow u_i(\widehat{x}^s, p^s) \leq u_i(x, p)$ by Claim 3 it is equal to write that $(p - r)\widehat{x}_i^s \leq (p - r)x_i$, since $p \neq r$, this is same as $\widehat{x}_i^s \leq x_i$, by Claim 5 it is equal to write that $\frac{\widehat{c}_i^s E}{\widehat{c}_i^s + c_j} \leq x_i$, then we have that $\frac{c_i E}{c_i + c_j} \leq x_i$. Since $x_i \leq \frac{c_i E}{c_i + c_j}$ and $\frac{c_i E}{c_i + c_j} \leq x_i$, we obtain $x_i = \frac{c_i E}{c_i + c_j}$. By efficiency $x_j = E - x_i$, since $x_i = \frac{c_i E}{c_i + c_j}$, then $x_j = E - \frac{c_i E}{c_i + c_j} = \frac{c_j E}{c_i + c_j}$. ■

Claim 7. *If $N = \{i, j\}$ and $p \neq r$, then $x_i = \frac{c_i E}{c_i + c_j}$ and $x_j = \frac{c_j E}{c_i + c_j}$.*

Proof. It is enough to consider that one of them as a sequence of rational numbers. Let $\{c_i^s\}$ a decreasing sequence of rational numbers that converges to $c_i \in \mathbb{R}$ $(lim_{s \to \infty}\{c_i^s\} = c_i)$. In each $(N^s, K, E, c^s, r) \in \mathcal{L}$ with $N^s = N$ and $c^s = (c_i^s, c_j)$. By Claim 6 for x_i and x_j, we have $x^s = \left(\frac{c_i^s E}{c_i^s + c_j}, x_j\right)$. Then by LM $u_i(x, p) \leq u_i(x^s, p^s)$, by Claim 3 it is equal to write that $(p - r)x_i \leq (p - r)x_i^s$, since $p \neq r$ this is same as $x_i \leq x_i^s$, by Claim 5 it is equal to write that $x_i \leq \frac{c_i^s E}{c_i^s + c_j}$, then we have that $x_i \leq \frac{c_i E}{c_i + c_j}$. Let $\{\widehat{c}_i^s\}$ an increasing sequence of rational numbers that converges to $c_i \in \mathbb{R}$, this is possible since $c_i + c_j > E$. Since $c_i > 0$ there exist each $(\widehat{N}^s, K, E, \widehat{c}^s, r) \in \mathcal{L}$ with $\widehat{N}^s = N$ and $\widehat{c}^s = (\widehat{c}_i^s, c_j)$. By Claim 6 for x_i and x_j, we have $\widehat{x}^s = \left(\frac{\widehat{c}_i^s E}{\widehat{c}_i^s + c_j}, x_j\right)$. Then by LM $u_i(\widehat{x}^s, p^s) \leq u_i(x, p)$, by Claim 3 it is equal to write that $(p - r)\widehat{x}_i^s \leq (p - r)x_i$, since $p \neq r$ this is same as $\widehat{x}_i^s \leq x_i$, by Claim 5 it is equal to write that $\frac{\widehat{c}_i^s E}{\widehat{c}_i^s + c_j} \leq x_i$, then we have thant $\frac{c_i E}{c_i + c_j} \leq x_i$. Since $(x_i \leq \frac{c_i E}{c_i + c_j})$ and $(\frac{c_i E}{c_i + c_j} \leq x_i)$, we obtain $x_i = \frac{c_i E}{c_i + c_j}$. By efficiency $x_j = E - x_i$, since $x_i = \frac{c_i E}{c_i + c_j}$, then $x_j = E - \frac{c_i E}{c_i + c_j} = \frac{c_j E}{c_i + c_j}$. ■

Claim 8. *If $p \neq r$, then $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$.*

Proof. Let $i \in N$, $j \in \mathbb{N} \setminus N$ and $(N^{ij}, K, E, c^{ij}, r) \in \mathcal{L}$, with $N^{ij} = \{i, j\}$ and $c^{ij} = (c_i, c(N \setminus \{i\}))$. By efficiency, $x_i = E - x_{N \setminus \{i\}}$. By LR and $p \neq r$ this is same as $x_i = E - x_j^{ij}$. By Claim 7 this is equal to write $x_i = E - \frac{c(N \setminus \{i\})E}{c_i + c(N \setminus \{i\})}$, then we obtain that $x_i = \frac{c_i E}{c(N)}$. ■

Therefore, the amount of land is given by Claim 8.

□

We define ψ^f as the rule corresponding to $f \in F$ that is given by $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$, and $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$.

Here, we provide some rules as examples of ψ^f.

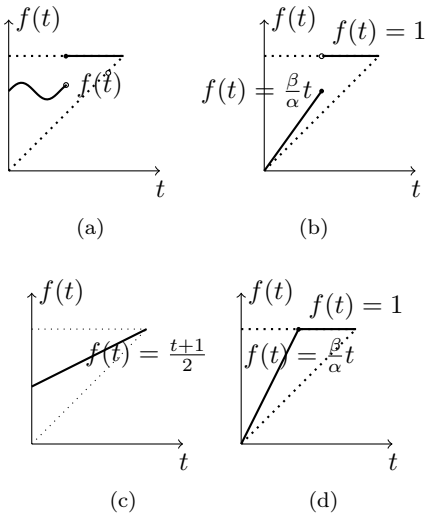


Figura 2: Examples of functions $f \in F$ in (a), (b),(c) and (d), which generate rules.

4 WEIGHTED STANDARD FOR TWO-PERSON

In order to define the subfamily of rules from ψ^f. We study the property called standard for two-person inspired on Hart and Mas-Colell (1989) which say about "divide the surplus equally" idea for two-person situations. In our context, this property said both the tenant and the lessor obtain equal benefit. Let $\mathcal{L}_2$ be a set of land rental problem with unique lessor i.e. $(N, K, E, c, r) \in \mathcal{L}_2$, where $N = \{1\}$. We formalise as follows.

Standard for two-person (S2) Given $(N, K, E, c, r) \in \mathcal{L}_2$, then $u_0\left(\psi(N, K, E, c, r)\right) = u_1\left(\psi(N, K, E, c, r)\right)$.

Theorem 2. *A rule ψ satisfies LR, LM and S2 if and only if the price is given by $p = \frac{K+rE}{2E}$ and, when $p \neq r$, the amount of land is given by $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$.*

Next, we generalize the standard for two-person concept and result to two-person property in a nonsymmetric way as follows. We called as weighted standard for two-person. In this context, Let $\omega \in [0, 1]$ be a positive weight. Since the rule satisfy efficiency, it is enough to choose a positive weight between the tenant and lessor. By this property both the tenant and the lessor obtain proportional benefit.

Weighted standard for two-person (WS2) There exists $\omega \in [0, 1]$, such that

$$(1-\omega)u_0\left(\psi(\{1\}, K, E, c, r)\right) = \omega u_1\left(\psi(\{1\}, K, E, c, r)\right) \text{ for all } (\{1\}, K, E, c, r) \in \mathcal{L}^1.$$

Theorem 3. *A rule ψ satisfies LR, LM and WS2 if and only if there exists $\omega \in [0, 1]$ such that the price is given by $p = \frac{K-\omega(K-rE)}{E}$ and, when $\omega < 1$, the quantity of land is given by $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$.*

Proof. ($\Leftarrow$) Let ψ be a rule given by $p = \frac{K+rE}{2E}$ and, when $p \neq r$, then $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. By Theorem 1 the rule satisfies LR and LM for $f(t) = \frac{1+t}{2}$ and $p = \frac{K+rE}{2E}$. So, we want to prove that $u_0(\psi(\{1\}, K, E, c, r)) = u_1(\psi(\{1\}, K, E, c, r))$. From the left side of the last equality, by definition it is equal to $\left(K - \frac{K+rE}{2E}E\right)$, and this equal to $\frac{K+rE}{2}$. Analogously, from the right side of the expression we have that $\omega u_1(\psi(\{1\}, K, E, c, r)) = \frac{K+rE)}{2E} - r)x_1$, and by efficiency we obtain $u_1(\psi(\{1\}, K, E, c, r)) = \frac{K-rE}{2}$. Therefore, the equality of interest is true.

($\Rightarrow$) Let ψ be a rule that satisfies LR and LM, as we have seen by Theorem 1 there exists $f : [0, 1] \to [0, 1]$ with $f(t) \geq t$ for all $t \in [0, 1]$ such that $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$ and, when, $p \neq r$, $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. We need to prove that $\frac{K}{E} f\left(\frac{rE}{K}\right) = \frac{K+rE}{2E}$ or equivalently $f(t) = \frac{1+t}{2}$

for $t = \frac{rE}{K} \in [0,1]$. Given $N = \{1\}$, $K = E = 1$, $c(N) = 1$ and $t = \frac{rE}{K} \in [0,1]$, we have $(\{1\}, 1, 1, (1), t) \in \mathcal{L}_2$ as base land rental problem. So, by S2 we have

$$\begin{aligned} u_0\left(\psi(\{1\}, 1, 1, (1), t)\right) &= u_1\left(\psi(\{1\}, 1, 1, (1), t)\right) \\ 1 - px_1 &= \omega(p - t)x_1 \\ 1 - p &= p - t \text{ by Eff } x_1 = 1 \\ p &= \frac{1+t}{2} \end{aligned} \tag{1}$$

Since $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$, we have $p = f(t)$. From (1) and $p = f(t)$ we obtain $f(t) = \frac{1+t}{2}$ for all $t \in [0,1]$.

□

Note that, when $\omega = 1$, it is optimal rule for the tenant, and when, $\omega = 0$, it is optimal for the lessors. Note that, when, $\omega = \frac{1}{2}$, it is the result given by Theorem 2. Given $\omega \in [0,1]$ we define ψ^ω as the rule corresponding functions $f \in F$ and $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$.

5 CONSISTENCY

Consistency in our context says that the payoff of each lessors and the tenant in a land rental problem should not be affected, if a group of lessors leave after get their payoffs. In order to explain the structure of this property, let $L = (N, K, E, c, r) \in \mathcal{L}$, $S \subset N$ be a set who leave after get their payoffs with $p \geq r$ and $x(S) < E$. Let $\mathcal{L}_c$ be a set of land rental problems that is called reduced land rental problem, and is defined as $L' = (N', K', E', c', r) \in \mathcal{L}_c$ given by $N' = N \setminus S$, new maximal profit of tenant is $K' = K - px(S)$ with $px(S)$ amount that tenant paid to lessors in S. Tenant and lessors in L' need $E' = E - x(S)$ units of land. For each lessor in N', the amount of available land is given by the vector $c' = c_{N\setminus S} \in \mathbb{R}^{N\setminus S}$ and the reservation price it is equal than the initial land rental problem r. Next, we formalise as follows.

Consistency(Cons) For all $L \in \mathcal{L}$ and $L' \in \mathcal{L}_c$ given by $N' = N \setminus S$, $S \subset N$, $x(S) < E$, $K' = K - px(S)$, $E' = E - x(S)$ and $c' = c_{N\setminus S} \in \mathbb{R}^{N\setminus S}$ such that a rule satisfies $u_i(\psi(L')) = u_i(\psi(L))$ for all $i \in N_0 \setminus S$.

Theorem 4. *A rule ψ satisfies LR, LM and Cons if and only if there exist $\alpha, \beta \in [0,1]$ with $\alpha \leq \beta$ such that the following two conditions holds:*

a) $p \in \left\{\frac{K}{E}\beta, \frac{K}{E}\right\}$ *if* $rE = \alpha K$, $p \in \left\{0, \frac{K}{E}\right\}$ *if* $r = 0$, *and otherwise*

$$p = \begin{cases} \frac{\beta}{\alpha} r & \text{if } 0 < rE < \alpha K \\ \frac{K}{E} & \text{if } rE > \alpha K \end{cases} \quad \text{and,}$$

b) *when* $p \neq r$, $x_i = \frac{c_i E}{c(N)}$ *for all* $i \in N$.

Proof. ($\Leftarrow$) Let $\alpha, \beta \in [0,1]$ with $\alpha \leq \beta$. The price and amount of land are given by a) and b) respectively. It can be written as $p = \frac{K}{E} f(\frac{rE}{K})$ with $f(\frac{rE}{K}) \geq \frac{rE}{K}$ for all $\frac{rE}{K} \in [0,1]$ which satisfies with $f(t) = \frac{\beta}{\alpha} t$ if $t < \alpha$, $f(t) = 1$ if $t > \alpha$, $f(t) \in \{\beta, 1\}$ if $t = \alpha$ and $f(t) \in \{0,1\}$ if $r = 0$. Hence, with Theorem 1 it satisfies LR and LM. Let N, N' given as in the definition of Cons. We will prove that $u_i(\psi(L')) = u_i(\psi(L))$ for all $i \in N_0 \setminus S$, where $\psi(L') = (x', p')$ and $\psi(L) = (x, p)$. Firstly, we prove that $p' = p$. We distinguish following cases:

Case 1: $r = 0$ and $p = 0$. Since these conditions, $f(0) = 0$, then $p' = 0$ and $p = 0$. Hence, it is true.

Case 2: $r = 0$ and $p = \frac{K}{E}$. Since these conditions, $f(0) = 1$, then $p' = \frac{K'}{E'}$. As $p = \frac{K}{E}$. Hence $p' = \frac{K'}{E'} = \frac{K - \frac{K}{E}x(S)}{E - x(S)} = \frac{K}{E} = p$.

Case 3: $0 < rE < \alpha K$. Hence, $p = \frac{\beta}{\alpha} r$. We know that $p' = \frac{\beta}{\alpha} r$ when $0 < rE' < \alpha K'$. Hence, it is enough to check that $0 < rE' < \alpha K'$. equivalently $0 < r\left(E - x(S)\right) < \alpha\left(K - \frac{\beta}{\alpha} r x(S)\right)$. This is equivalent to check the following inequalities. First: $0 < r\left(E - x(S)\right)$. Since $0 < rE$ and $r > 0$. Thus, it is equivalent to check that $0 < E - x(S)$, which it true by hypothesis. And, second: $rE - rx(S) < \alpha K - \alpha \frac{\beta}{\alpha} r x(S)$. Since $rE < \alpha K$, it is enough to check that $rx(S) \geq \beta r x(S)$. This is trivially true when $rx(S) = 0$. Otherwise, it is equivalent to check that $\beta \leq 1$, which are true by definition.

Case 4: $rE = \alpha K$ and $p = \frac{K}{E}\beta$. Since $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$ and $p = \frac{K}{E}\beta$, then $f\left(\frac{rE}{K}\right) = \beta$.

Since $\frac{rE'}{K'} = \frac{r(E - x(S))}{K - \frac{K}{E}\beta x(S)} = \frac{rE(E - x(S))}{K(E - \beta x(S))} \leq \frac{rE(E - x(S))}{K(E - x(S))} = \frac{rE}{K} = \alpha$.

We will show that $f\left(\frac{rE'}{K'}\right) = \frac{\beta}{\alpha}\frac{rE'}{K'}$. So, we have two subcases from ($\frac{rE'}{K'} \leq \alpha$): If $rE' < \alpha K'$, then $f\left(\frac{rE'}{K'}\right) = \frac{\beta}{\alpha}\frac{rE'}{K'}$. If $rE' = \alpha K'$, then $f\left(\frac{rE'}{K'}\right) = f(\alpha)$, by $rE = \alpha K$ it is equal to $f\left(\frac{rE'}{K'}\right) = f(\alpha)$, since $f\left(\frac{rE}{K}\right) = \beta$, it is equal to $f\left(\frac{rE'}{K'}\right) = \beta$, since $rE' = \alpha K'$, we obtain that $f\left(\frac{rE'}{K'}\right) = \frac{\beta}{\alpha}\frac{rE'}{K'}$.

Hence, $p' = \frac{K'}{E'} f\left(\frac{rE'}{K'}\right)$, since $f\left(\frac{rE'}{K'}\right) = \frac{\beta}{\alpha}\frac{rE'}{K'}$, it is euqal to $p' = \frac{K'}{E'}\frac{\beta}{\alpha}\frac{rE'}{K'} = \frac{\beta}{\alpha} r \overset{hyp}{=} \frac{K}{E}\beta = p$.

Case 5: $rE = \alpha K$ and $p = \frac{K}{E}$. Since $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$ and $p = \frac{K}{E}$, then $f\left(\frac{rE}{K}\right) = 1$. Since $\frac{rE'}{K'} = \frac{r(E - x(S))}{K - px(S)} = \frac{r(E - x(S))}{K - \frac{K}{E}x(S)} = \frac{rE(E - x(S))}{K(E - x(S))} = \frac{rE}{K} = \alpha$.

$$\frac{rE'}{K'} = \alpha. \tag{2}$$

Thus, since $rE' = \alpha K'$, $f\left(\frac{rE'}{K'}\right) \overset{(2)}{=} f(\alpha)$, since $rE = \alpha K$ and $f\left(\frac{rE}{K}\right) = 1$, we obtain $f\left(\frac{rE'}{K'}\right) = 1$. Hence, $p' = \frac{K'}{E'} f\left(\frac{rE'}{K'}\right)$, since $f\left(\frac{rE'}{K'}\right) = 1$, we have $p' = \frac{K'}{E'} = \frac{K - \frac{K}{E}x(S)}{E - x(S)} = \frac{K}{E} = p$

Case 6: $rE > \alpha K$. Hence $p = \frac{K}{E}$. We know that $p' = \frac{K'}{E'}$ when $rE' > \alpha K'$. Since $\frac{K'}{E'} = \frac{K - \frac{K}{E}x(S)}{E - x(S)} = \frac{K}{E}$. it is enough to check that $rE' > \alpha K'$. It is equivalent to check that $r(E - x(S)) > \alpha\left(K - \frac{K}{E}x(S)\right)$. Equivalently $r\left(E - x(S)\right) > \alpha K\left(\frac{E - x(S)}{E}\right)$. As $E - x(S) > 0$, it is equivalent to $rE > \alpha K$ which is true we are in this case.

Now, we check that: $u_i(\psi(L')) = u_i(\psi(L))$ for all $i \in N \setminus S$. Equivalently $(p' - r)x_i' = (p - r)x_i$ for all $i \in N \setminus S$. This is trivially true when $p = r$ and $p' = p$. Hence, we assume $p \neq r$. As $p' = p$, we need to prove $x_i' = x_i$. Since $c(N) = c(N \setminus S) + c(S)$, then $x_i' = \frac{c_i}{c(N \setminus S)}(E - x(S)) = \frac{c_i}{c(N \setminus S)}\left(E - \frac{c(S)E}{c(N)}\right) = \frac{c_i}{c(N \setminus S)} E\left(\frac{c(N) - c(S)}{c(N)}\right) = \frac{c_i}{c(N \setminus S)} E\left(\frac{c(N \setminus S)}{c(N)}\right) = \frac{c_i}{c(N)} E = x_i$. Analogously, We check for tenant $u_0(\psi(L')) = u_0(\psi(L))$. This is equivalent to prove $K' - p'E' = K - pE$. As $p' = p$ and K and $x(S)$ are greater than zero. since $K' - p'E' = K - px(S) - p(E - x(S)) = K - pE$.

($\Rightarrow$) Let ψ be a rule that satisfies LR, LM and Cons. Under LR and LM by Theorem 1 there exists $f : [0, 1] \to [0, 1]$ with $f(t) \geq t$ for all $t \in [0, 1]$, such that $p = \frac{K}{E} f\left(\frac{rE}{K}\right)$ and, when $p \neq r$, $x_i = \frac{c_i}{c(N)} E$ for all i $\in N$.

Let L' and L two land rental problems given as the definition of Cons. So, by Cons we have $u_i(\psi(L')) = u_i(\psi(L))$ for all $i \in N_0 \setminus S$, where $\psi(L') = (x', p')$ and $\psi(L) = (x, p)$. Thus, we have: $u_0(\psi(L')) = u_0(\psi(L))$ by definition it is equal to write that $K' - p'E' = K - pE$ and this equal to $K - px(S) - p'(E - x(S)) = K - pE$, from this we obtain that $p' = p$.

We will prove the existence of α and β with $\alpha \leq \beta$, which satisfy $p = \frac{\beta}{\alpha} r$ if $0 < rE < \alpha K$, $p = \frac{K}{E}$ if $rE > \alpha K$, $p \in \{\frac{\beta}{\alpha} r, \frac{K}{E}\}$ if $rE = \alpha K$ and $p \in \{0, \frac{K}{E}\}$ if $r = 0$ or equivalently $f(t) = \frac{\beta}{\alpha} t$ if $0 < t < \alpha$, $f(t) = 1$ if $t > \alpha$, $f(t) \in \{\beta, 1\}$ if $t = \alpha$ and $f(t) \in \{0, 1\}$ if $r = 0$.

We assume first that $f(t) = 1$ for all $t \in [0, 1]$, which are obtained with $\alpha = 0$ and $\beta = 1$, which satisfy our condition.

Assume now there exists $\widehat{t}$ such that $f(\widehat{t}) < 1$. Let $\alpha = Inf\{t \in [0,1] : f(t') = 1 \text{ for all } t' > t\}$ equivalently $\alpha = Sup\{t : f(t) < 1\}$ and $\beta = Sup\{f(t) : f(t) < 1\}$, clearly $\alpha, \beta \in [0,1]$ and $\alpha \leq \beta$.

Without loss of generality, it is enough to prove the result for $K = E = 1$. Let $N = \{1,2\}$, and for each $r \in [0,1]$ and $\gamma \in]0,1[$, we consider a land rental problem $(\{1,2\}, 1, 1, (\gamma, 1-\gamma), r)$. We assume that lessor 2 leaves the land rental problem. So, $K' = K - px_2 = 1 - f(r)(1-\gamma)$ and $E' = E - x_2 = \gamma$. Hence, we reduce land rental problem to $(\{1\}, 1 - f(r)(1-\gamma), \gamma, (\gamma), r)$

So, by hypothesis we have: $p = \frac{K}{E} f\left(\frac{rE}{K}\right) = f(r)$ and

$p' = \frac{K'}{E'} f\left(\frac{rE'}{K'}\right) = \frac{1-f(r)(1-\gamma)}{\gamma} f\left(\frac{r\gamma}{1-f(r)(1-\gamma)}\right)$

As we have $p' = p$ and from the last two equalities, we have:

$$f\left(\frac{r\gamma}{1-f(r)(1-\gamma)}\right) = \frac{f(r)\gamma}{1-f(r)(1-\gamma)} \text{ for all } r \in [0,1] \text{ and } \gamma \in]0,1[\tag{3}$$

In particular for $r = 0$. From (3) we have $f(0) = \frac{f(0)\gamma}{1-f(0)(1-\gamma)}$ for all γ. We have two cases: First $f(0) = 0$ it is trivially true, and Second $f(0) > 0$, then $\gamma = 1 - f(0)(1-\gamma)$ for all γ. Equivalently $\gamma(1 - f(0)) = (1 - f(0))$ for all γ. Which implies that $f(0) = 1$. Hence $f(0) \in \{0,1\}$. It satisfy one of our conditions.

In particular for $r = 1$. clearly we have $f(1) = 1$. Hence $f(1) = 1$ is true.

Now for $0 < r \leq 1$. From (3) we have $f\left(\frac{r\gamma}{1-f(r)(1-\gamma)}\right) = \frac{f(r)\gamma}{1-f(r)(1-\gamma)}$ for all γ.

We define $F^r(\delta) = \frac{r\delta}{1-f(r)(1-\delta)} \in [0,r]$ for all $\delta \in [0,1]$ as a strictly increasing and continuous function, and its inverse for some r such that $f(r) < 1$ is defined as $G^r(t) = \frac{(1-f(r))t}{r-f(r)t} \in [0,1]$ for all $t \in]0,r]$.

For all $t \in]0,r]$, Then

$$\begin{aligned} f(t) = f\left(F^r\left(G^r(t)\right)\right) &\overset{defF)}{=} f\left(\frac{rG^r(t)}{1-f(r)(1-G^r(t))}\right) \\ &\overset{(3)}{=} \frac{f(r)G^r(t)}{1-f(r)(1-G^r(t))} \\ &\overset{defG}{=} \frac{f(r)\frac{(1-f(r))t}{r-f(r)t}}{1-f(r)\left(1-\frac{(1-f(r))t}{r-f(r)t}\right)} \\ &= \frac{f(r)}{r}t \text{ for } t \in]0,r] \end{aligned} \tag{4}$$

We set r such that $f(r) < 1$ and and for all $t \in]0,r]$. $\frac{f(t)}{t} = \frac{f(r)}{r}$

We will prove that $f(t) = \pi t$ for all $t \in]0,\alpha[$, where $\pi = \frac{f(r)}{r}$ for any r such that $f(r) < 1$. For all $t \in]0,\alpha[$ exists $r' > t$ and $f(r') < 1$ such that $\frac{f(t)}{t} = \frac{f(r')}{r'}$. If $t < r$. We can take $r' = r$, thus $\frac{f(t)}{t} = \pi$. If $t \geq r$. Then $r' > r$, thus $\frac{f(r)}{r} = \frac{f(r')}{r'} = \pi$. Hence $f(t) = \pi t$ for all $t \in]0,\alpha[$. So now, we need to prove that $\pi = \frac{\beta}{\alpha}$ equivalently $r\beta = f(r)\alpha$ $\beta = Sup\{f(t) : f(r) < 1\} = Sup\{r\pi : f(r) < 1\}$. As $f(t) < 1$ then $t \leq \alpha$

Case I. If $f(\alpha) = 1$. then $f(t) < 1$, thus $t \leq \alpha$.

$$\beta = Sup\{\pi t : f(t) < 1\} = \pi Sup\{t : f(t) < 1\} = \pi\alpha$$

Case II. If $f(\alpha) < 1$. Then $f(t) = \pi t$ for all $t \in]0,\alpha]$. In particular $f(\alpha) = \pi\alpha$. $\beta = Sup\{\pi t : f(t) < 1\}$ If $t < \alpha$

$$\beta = Sup\{\pi t : f(t) < 1\} = \pi Sup\{t : f(t) < 1\} = \pi\alpha$$

If $t = \alpha$, we precede analogously and obtain $\beta = \pi\alpha$

Then α and β satisfy our condition. For $r = \alpha$. So there exists $t = r$ such that

$$f(r) = \frac{f(r)}{r} r.$$

If $\alpha = 0$ it is trivially true. If $\alpha > 0$. So we have $f(r) = f(\alpha)$. As $f(r) = r\pi = \alpha\pi$. Then $f(\alpha) = \beta$. Otherwise $f(\alpha) = 1$. Hence $f(\alpha) \in \{\beta, 1\}$ satisfy our condition. □

We are interested in rules for which small changes in the land rental problem do not cause large changes in the chosen allocation in terms of the welfare of the lessors and tenant. We consider continuity property, i.e., small changes in the valuation of the land should not cause large changes in the welfare of the lessors and tenant. We formalise this property as follow.

Continuity(Cont) A rule ψ is continuous if it is continuous function in p and x.

Corollary 1. *A rule ψ satisfies LR, LM, Cons and Cont if and only if there exists $\alpha \in [0, 1]$ such that the following two conditions holds:*

a) $p = \begin{cases} \frac{r}{\alpha} & \text{if } rE \leq \alpha K \\ \frac{K}{E} & \text{if } rE > \alpha K \end{cases}$ *and*

b) $x_i = \frac{c_i E}{c(N)}$ *for all $i \in N$.*

Proof. ($\Leftarrow$) We have $\alpha \in [0, 1]$ such that, the price and amount of land are given by a) and b) respectively. It can be written as $p = \frac{K}{E} f(\frac{rE}{K})$ with $f(\frac{rE}{K}) \geq \frac{rE}{K}$ for all $\frac{rE}{K} \in [0, 1]$ which satisfies with $f(t) = \frac{t}{\alpha}$ if $t \leq \alpha$, $f(t) = 1$ if $t > \alpha$. Hence, by Theorem 4 it satisfies LR, LM and Cons. We Will prove that satisfy Cont. It is enough to prove that $\frac{r}{\alpha} = \frac{K}{E}$. Since $p = \frac{r}{\alpha}$ and $p = \frac{K}{E} f(\frac{rE}{K}) = \frac{K}{E}$ then from last two equalities $\frac{r}{\alpha} = \frac{K}{E}$. And $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$ determines a continuous function.

($\Rightarrow$) We have LR, LM, Cons and Cont. With LR, LM and Cons by Theorem 4 there exist $\alpha, \beta \in [0, 1]$ with $\alpha \leq \beta$ such that the price is given by $p = \frac{\beta}{\alpha} r$ if $0 < rE < \alpha K$, $p = \frac{K}{E}$ if $rE > \alpha K$, $p \in \{\frac{K}{E}\beta, \frac{K}{E}\}$ if $rE = \alpha K$ and $p \in \{0, \frac{K}{E}\}$ if $r = 0$. And, when $p \neq r$, $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. And by adding Cont. we will prove that $p = \frac{r}{\alpha}$ if $rE \leq \alpha K$ and $p = \frac{K}{E}$ if $rE \geq \alpha K$. And $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. In this sense, we have discontinuous functions in $t = 0$ and $t = \alpha$, then we still need to prove the following cases:

i) If $r = 0$. We can take $p = \lim_{r \to 0} p = \lim_{r \to 0} \frac{\beta}{\alpha} r = 0$ then by Cont $p = 0$.

ii) If $rE = \alpha K$. As by Cont $\frac{\beta}{\alpha} r = \frac{K}{E}$, then $\beta = \frac{K\alpha}{rE} = 1$.

ii) We have that $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$ when $p \neq r$. we will prove that $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$ when $p = r$.

Let $L_t = (N, K, E, c, r_t)$ with $lim_{t \to \infty} r_t = 0$ and $r_t > 0$ for all t. Therefore, $x_i^t = \frac{c_i E}{c(N)}$ for all t then $x_i = \lim_{r \to 0} \frac{c_i E}{c(N)} = \frac{c_i E}{c(N)}$ □

We define $\psi^{\alpha,1}$ as the rule corresponding function $f \in F$ with $\alpha \in [0, 1]$ and $\beta = 1$. Notice that $\psi^{0,1} = \psi^0$

Corollary 2. *ψ^0 and $\psi^{1,1}$ are the unique rules that satisfy LR and LM; WS2; and Cons and Cont.*

Proof. On the one hand, under Corollary 1 we have $\alpha \in [0, 1]$ with $p = \frac{r}{\alpha}$ if $rE \leq \alpha K$ and $p = \frac{K}{E}$ if $rE > \alpha K$, and $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$, and on the other hand, by Theorem 3 we have $\omega \in [0, 1]$ with $p = \frac{K - \omega(K - rE)}{E}$ and when $p \neq r$ $x_i = \frac{c_i E}{c(N)}$ for all $i \in N$. Let $r = \frac{\alpha K}{E} \in [0, 1]$. Thus, $p = \frac{K}{E}$. This implies that $\frac{K}{E} = \frac{K - \omega(K - rE)}{E}$, then $\omega(K - rE) = 0$, then there are two possibilities: On the one hand, that $\omega = 0$ which is optimal for lessors. On the other hand, $K = rE$, and as $rE = \alpha K$ which implies $K = \alpha K$, as $K > 0$, then $\alpha = 1$ which is optimal for tenant.

Under ψ^0 and $\psi^{1,1}$ we have $p = r$ and $p = \frac{K}{E}$ respectively. If $p = r$ for all (N, K, E, c, r). It satisfies consistency and continuity for $\alpha = 1$ by Corollary 1 and it satisfies for $\omega = 1$ by Theorem 3. If $p = \frac{K}{E}$ for all (N, K, E, c, r). It satisfies weight standard for two-person for $\alpha = 0$ by Corollary 1 and it satisfies for $\omega = 0$ by Theorem 3. □

Note that ψ^0 and $\psi^{1,1}$ are rules that are optimal for tenant and lessors respectively. We can see a summary in the following table:

Table 1 Summary of properties

Property/rules	ψ^f	ψ^ω	$\psi^{\alpha,1}$	ψ^0	$\psi^{1,1}$
Land reassignment	Yes*	Yes*	Yes*	Yes	Yes
Land monotonicity	Yes*	Yes*	Yes*	Yes	Yes
Weight standard for 2-person	-	Yes*	-	Yes	Yes
Consistency	-	-	Yes*	Yes	Yes
Continuity	-	Yes	Yes*	Yes	Yes

*: It is one of the properties that characterize a rule.

References

Brandt, F. (2015). Set-monotonicity implies kelly-strategyproofness. *Social Choice and Welfare*, pages 1–12.

Dell, M. (2010). The persistent effects of Peru's mining mita. *Econometrica*, 78(6):1863–1903.

Hart, S. and Mas-Colell, A. (1989). Potential, value, and consistency. *Econometrica: Journal of the Econometric Society*, pages 589–614.

Jaramillo, P., Kayı, c., and Klijn, F. (2014). Asymmetrically fair rules for an indivisible good problem with a budget constraint. *Social Choice and Welfare*, 43(3):603–633.

Ju, B.-G. (2003). Manipulation via merging and splitting in claims problems. *Review of Economic Design*, 8(2):205–215 (2003).

Ju, B.-G. (2013). Coalitional manipulation on networks. *Journal of Economic Theory*, 148(2):627 – 662.

Kelly, J. S. (1977). Strategy-proofness and social choice functions without single-valuedness. *Econometrica*, 45(2):pp. 439–446.

Kojima, F. and Manea, M. (2007). Strategy-proofness of the probabilistic serial mechanism in large random assignment problems. Working paper, Harvard University.

Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, 66(1):23–38.

Massó, J., Nicolò, A., Sen, A., Sharma, T., and Ülkü, L. (2015). On cost sharing in the provision of a binary and excludable public good. *Journal of Economic Theory*, 155:30 – 49.

Moulin, H. (2007). On scheduling fees to prevent merging, splitting, and transferring of jobs. *Mathematics of Operations Research*, 32(2):266–283.

Moulin, H. and Shenker, S. (2001). Strategyproof sharing of submodular costs:budget balance versus efficiency. *Economic Theory*, 18(3):511–533.

Sosa, I. (2011). License to operate: Indigenous relations and free prior and informed consent in the mining industry. *Sustainalytics, Amsterdam, The Netherlands.*

Sun, N. and Yang, Z. (2003). A general strategy proof fair allocation mechanism. *Economics Letters*, 81(1):73–79.

Thomson, W. (2008). Two families of rules for the adjudication of conflicting claims. *Social Choice and Welfare*, 31(4):667–692.

UN-HABITAT (2012). Land and conflict: Toolkit and guidance for preventing and managingland and natural resources conflict. Guidance note, The United Nation.

Valencia-Toledo, A. and Vidal-Puga, J. (2015). A bilateral and sequential bargaining protocol for land rental arrangements. Working paper, Universidade de Vigo.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Coalition formation in a multi-agent inventory problem with two acquisition costs

Alejandro Saavedra-Nieves[1], Ignacio García-Jurado[2], M. Gloria Fiestras-Janeiro[3]

[1]Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela.
[2]Departamento de Matemáticas, Universidade da Coruña.
[3]Departamento de Estatística e Investigación Operativa, Universidade de Vigo.

ABSTRACT

A multi-agent inventory problem is a situation in which several agents face individual inventory problems and make an agreement to coordinate their orders with the objective of reducing costs. Fiestras-Janeiro et al. (2015) propose a model to deal with a multi-agent inventory problem in which several farmers cooperate to make joint orders of animal feed, the cost of the usual feeding ration being similar to the cost of the shortage feeding ration. However, in some situations the latter cost may be significantly smaller than the former cost since the shortage feeding ration is produced in the farm. When this happens each farm faces a continuous-review inventory problem, with a deterministic and linear demand, with no holding costs, with a limited capacity warehouse, with shortages and with two acquisition costs. In this work, we present a cooperative multi-agent inventory model to describe and analyze the system in which a group of farms cooperate by placing joint orders under these assumptions. If two acquisition costs are considered, the model and the procedure for the analysis are significantly different. The optimal inventory policies for this model are established, and a procedure to obtain a stable coalitional structure is proposed (following Elomri et al, 2012). In addition, an allocation rule for sharing the optimal costs is introduced and its performance is illustrated in a farming community in the North West of Spain.

Keywords: Multi-agent inventory problems, cooperative games, coalition formation.

REFERENCES

Elomri A., Ghaffari A., Jemai Z., Dallery Y. (2012) Coalition formation and cost allocation for joint replenishment systems. Production and Operations Management 21, 1015-1027.

Fiestras-Janeiro M.G., García-Jurado I., Meca A., Mosquera M.A. (2015) Cooperation on capacitated inventory situations with fixed holding costs. European Journal of Operational Research 241, 719-726.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

COMPUTATIONAL PROCEDURE FOR A WIDE FAMILY OF MIXED COALITIONAL VALUES

José Miguel Giménez and María Albina Puente

Department of Applied Mathematics III and Engineering School of Manresa,
Technical University of Catalonia, Spain.

ABSTRACT

We consider a family of mixed coalitional values. They apply to games with a coalition structure by combining a (induced) semivalue in the quotient game, but share within each union the payoff so obtained by applying different (induced) semivalues to a game that concerns only the players of that union. A computation procedure in terms of the multilinear extension of the original game is provided.

Keywords: cooperative game, semivalue, coalition structure, multilinear extension.

1. INTRODUCTION

Introduced in 1974 by Aumann and Drèze [6], the notion of *game with a coalition structure* gave a new impulsion to the development of value theory. These authors extended the Shapley value to this new framework in such a manner that the game really splits into subgames played by the unions isolatedly from each other, and every player receives the payoff allocated to him by the Shapley value in the subgame he is playing within his union.

In 1977, a second approach was used by Owen [19], when introducing the first *coalitional value*, called now the *Owen value*. The Owen value is the result of a *two–step procedure*: first, the unions play a *quotient game* among themselves, and each one receives a payoff which, in turn, is shared among its players in an internal game. Both payoffs, in the quotient game for unions and within each union for its players, are given by the Shapley value.

In 1982, Owen [21] applied the same procedure to the Banzhaf value and obtained the *modified Banzhaf value* or *Owen–Banzhaf value*. In this case the payoffs at both levels (unions in the quotient game and players within each union) are given by the Banzhaf value.

In 2002, Alonso and Fiestras [2] suggested to modify the two–step allocation scheme and use the Banzhaf value for sharing in the quotient game and the Shapley value within unions. This gave rise to the *symmetric coalitional Banzhaf value* or *Alonso–Fiestras value*. That same year, Amer, Carreras and Giménez [5] considered a sort of "counterpart" of the Alonso–Fiestras value where the Shapley value is used in the quotient game and the Banzhaf value within unions.

Thus, the possibilities to define a coalitional value by combining the Shapley and Banzhaf values were complete at that moment. In 2011, Carreras and Puente [11] extended the binomial semivalues to games with a coalitional structure: they used these values in the quotient game and the Shapley value within unions and obtained the *symmetric coalitional binomial semivalues*, a family depending on one parameter $q \in [0,1]$ (the Alonso–Fiestras value arises for $q = 1/2$). In 2003, Amer and Giménez [4] considered coalitional values defined by using a given semivalue in both steps of the procedure. More generally, in 2004 Albizuri and Zarzuelo [1] studied coalitional values defined by *any* pair of semivalues (*mixed modified semivalues*), one in the quotient game and the other within unions.

In 1972, Owen [17] introduced the *multilinear extension* and applied it to the calculus of the Shapley value. The computing technique based on the multilinear extension has been applied to many values: in 1975 to the Banzhaf value by Owen [18]; in 1992 to the Owen value by Owen and Winter [22]; in 1994 to the Owen–Banzhaf value by Carreras and Magaña [8]; in 1997 to

the quotient game by Carreras and Magaña [9]; in 2000 to binomial semivalues and multinomial probabilistic indices by Puente [23]; in 2003 to coalitional semivalues by Amer and Giménez [4]; in 2004 to the α–decisiveness and Banzhaf α–indices by Carreras [7]; in 2005 to the Alonso–Fiestras value by Alonso, Carreras and Fiestras [3]; in 2011 to symmetric coalitional binomial semivalues by Carreras and Puente [11]; in 2011 to semivalues by Carreras and Giménez in [10].

The present work extends the mixed modified semivalues values to games with coalition structure and presents the *generalized mixed modified semivalues*. They apply a semivalue in the quotient game that arises once the coalition structure is actually formed, but share within each union the payoff so obtained by applying different semivalues to a game that concerns only the players of that union. We will focus on giving a computation procedure for these values by means of the multilinear extension, generalizing the method obtained by Carreras and Giménez in [10] to compute semivalues in cooperative games.

2. PRELIMINARIES

Games and semivalues

Let N be a finite set of *players* and 2^N be the set of its *coalitions* (subsets of N). A *cooperative game* on N is a function $v : 2^N \to \mathbb{R}$, that assigns a real number $v(S)$ to each coalition $S \subseteq N$, with $v(\emptyset) = 0$. A game v is *monotonic* if $v(S) \leq v(T)$ whenever $S \subseteq T \subseteq N$ and *simple* if, moreover, $v(S) = 0$ or 1 for every $S \subseteq N$. A player $i \in N$ is a *dummy* in v if $v(S \cup \{i\}) = v(S) + v(\{i\})$ for all $S \subseteq N \backslash \{i\}$, and *null* in v if, moreover, $v(\{i\}) = 0$. Two players $i, j \in N$ are *symmetric* in v if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \backslash \{i, j\}$. Given a nonempty coalition $T \subseteq N$, the restriction to T of a given game v on N is the game $v_{|T}$ on T that we will call a *subgame* of v and is defined by $v_{|T}(S) = v(S)$ for all $S \subseteq T$.

Endowed with the natural operations for real–valued functions, *i.e.* $v + v'$ and λv for all $\lambda \in \mathbb{R}$, the set of all cooperative games on N is a vector space $\mathcal{G}_N$. For every nonempty coalition $T \subseteq N$, the *unanimity game* u_T is defined by $u_T(S) = 1$ if $T \subseteq S$ and $u_T(S) = 0$ otherwise, and it is easily checked that the set of all unanimity games is a basis for $\mathcal{G}_N$, so that $\dim(\mathcal{G}_N) = 2^n - 1$ if $n = |N|$. Each game $v \in \mathcal{G}_N$ can then be uniquely written as a linear combination of unanimity games, and its components are the *Harsanyi dividends* (Harsanyi [16]):

$$v = \sum_{T \subseteq N:\, T \neq \emptyset} \alpha_T u_T, \quad \text{where} \quad \alpha_T = \alpha_T(v) = \sum_{S \subseteq T} (-1)^{t-s} v(S) \tag{1}$$

and, as usual, $t = |T|$ and $s = |S|$.

By a *value* on $\mathcal{G}_N$ we will mean a map $f : \mathcal{G}_N \to \mathbb{R}^N$, that assigns to every game v a vector $f[v]$ with components $f_i[v]$ for all $i \in N$.

Following Weber's [25] axiomatic description, $\psi : \mathcal{G}_N \to \mathbb{R}^N$ is a *semivalue* iff it satisfies the following properties:

(i) *linearity*: $\psi[v + v'] = \psi[v] + \psi[v']$ (*additivity*) and $\psi[\lambda v] = \lambda \psi[v]$ for all $v, v' \in \mathcal{G}_N$ and $\lambda \in \mathbb{R}$;
(ii) *anonymity*: $\psi_{\theta i}[\theta v] = \psi_i[v]$ for all θ on N, $i \in N$, and $v \in \mathcal{G}_N$;
(iii) *positivity*: if v is monotonic, then $\psi[v] \geq 0$;
(iv) *dummy player property*: if $i \in N$ is a dummy in game v, then $\psi_i[v] = v(\{i\})$.

There is an interesting characterization of semivalues, by means of *weighting coefficients*, due to Dubey, Neyman and Weber [13]. Set $n = |N|$. Then: (a) for every *weighting vector* $\{p_k\}_{k=0}^{n-1}$ such that $\sum_{k=0}^{n-1} p_k \binom{n-1}{k} = 1$ and $p_k \geq 0$ for all k, the expression

$$\psi_i[v] = \sum_{S \subseteq N \backslash \{i\}} p_s [v(S \cup \{i\}) - v(S)] \quad \text{for all } i \in N \text{ and all } v \in \mathcal{G}_N,$$

where $s = |S|$, defines a semivalue ψ; (b) conversely, every semivalue can be obtained in this way; (c) the correspondence given by $\{p_k\}_{k=0}^{n-1} \mapsto \psi$ is bijective.

Thus, the payoff that a semivalue allocates to every player in any game is a weighted sum of his marginal contributions in the game. If p_k is interpreted as the probability that a given player i joins a coalition of size k, provided that all the coalitions of a common size have the same probability of

being joined, then $\psi_i[v]$ is the expected marginal contribution of that player to a random coalition he joins.

Well known examples of semivalues are the *Shapley value* φ (Shapley [24]), for which $p_k = 1/n\binom{n-1}{k}$, and the *Banzhaf value* β (Owen [18]), for which $p_k = 2^{1-n}$. The Shapley value φ is the only *efficient* semivalue, in the sense that $\sum_{i\in N} \varphi_i[v] = v(N)$ for every $v \in \mathcal{G}_N$.

Notice that these values are defined for each N. The same happens with the *binomial semivalues*, introduced by Puente [23] (see also Giménez [15] or Amer and Giménez [4]) as follows. Let $p \in [0,1]$ and $p_k = p^k(1-p)^{n-k-1}$ for $k = 0, 1, \ldots, n-1$. Then $\{p_k\}_{k=0}^{n-1}$ is a weighting vector and defines a semivalue that will be denoted as ψ^p and called the *p–binomial semivalue*. Using the convention that $0^0 = 1$, the definition makes sense also for $p = 0$ and $p = 1$, where we respectively get the *dictatorial index* ψ^0 and the *marginal index* ψ^1, introduced by Owen [20] and such that $\psi_i^0[v] = v(\{i\})$ and $\psi_i^1[v] = v(N) - v(N\backslash\{i\})$ for all $i \in N$ and all $v \in \mathcal{G}_N$. Of course, $p = 1/2$ gives $\psi^{1/2} = \beta$ —the Banzhaf value.

In fact, semivalues are defined on cardinalities rather than on specific player sets: this means that a weighting vector $\{p_k\}_{k=0}^{n-1}$ defines a semivalue ψ on all N such that $n = |N|$. When necessary, we shall write $\psi^{(n)}$ for a semivalue on cardinality n and p_k^n for its weighting coefficients. A semivalue $\psi^{(n)}$ induces semivalues $\psi^{(t)}$ for all cardinalities $t < n$, recurrently defined by the Pascal triangle (inverse) formula given by Dragan [12]:

$$p_k^t = p_k^{t+1} + p_{k+1}^{t+1} \quad \text{for} \quad 0 \le k < t, \tag{2}$$

A series $\psi = \{\psi^{(n)}\}_{n=1}^{\infty}$ of semivalues, one for each cardinality, is a *multisemivalue* if it satisfies Dragan's recurrence formula. Thus, the Shapley and Banzhaf values and all binomial semivalues are multisemivalues. By applying Eqs. (2) repeatedly, one gets the expression of the weighting coefficients of any induced semivalue in terms of the coefficients of the original semivalue, namely

$$p_s^t = \sum_{j=0}^{n-t} \binom{n-t}{j} p_{s+j}^n \quad \text{for} \quad 1 \le s < t < n. \tag{3}$$

The *multilinear extension* [17] of a game $v \in \mathcal{G}_N$ is the real–valued function defined on $\mathbb{R}^N$ by

$$f(X_N) = \sum_{S\subseteq N} \prod_{i\in S} x_i \prod_{j\in N\backslash S} (1 - x_j) v(S). \tag{4}$$

where X_N denotes the set of variables x_i for $i \in N$. The following properties directly derive from the definition:

(i) If $v, w \in \mathcal{G}_N$ and $\lambda, \mu \in \mathbb{R}$, then $f_{\lambda v + \mu w} = \lambda f_v + \mu f_w$.

(ii) If $\emptyset \neq T \subseteq N$, then $f_{u_T}(X_N) = \prod_{i\in T} x_i$.

(iii) In general, if $v = \sum_{T\subseteq N:\, T\neq\emptyset} \alpha_T u_T$, then $f_v(X_N) = \sum_{T\subseteq N:\, T\neq\emptyset} \alpha_T \prod_{i\in T} x_i$.

(iv) If $v \in \mathcal{G}_N$ and $\emptyset \neq T \subseteq N$, then $f_{v_{|T}}(X_T) = f_v(X_T, 0_{N\backslash T})$, where $(X_T, 0_{N\backslash T})$ denotes the set X_N with $x_i = 0$ for all $i \in N\backslash T$, and $v(T) = f_v(1_T, 0_{N\backslash T})$, where 1_T means $x_i = 1$ for all $i \in T$. Moreover,

$$\frac{\partial f_{v_{|T}}}{\partial x_i}(1_T) = v(T) - v(T\backslash\{i\}).$$

In particular, $\dfrac{\partial f_v}{\partial x_i}(1_N) = v(N) - v(N\backslash\{i\})$ for each $i \in N$.

As is well known, both the Shapley and Banzhaf values of any game v can be easily obtained from its multilinear extension. Indeed, $\varphi[v]$ can be calculated by integrating the partial derivatives of the multilinear extension of the game along the main diagonal $x_1 = x_2 = \cdots = x_n$ of the cube $[0,1]^N$ (Owen [17]), while the partial derivatives of that multilinear extension evaluated at point

$(1/2, 1/2, \ldots, 1/2)$ give $\beta[v]$ (Owen [18]). This latter procedure extends well to any p–binomial semivalue (see Puente [23], Freixas and Puente [14] or Amer and Giménez [4]) by evaluating the derivatives at point $(p, p, \ldots, p)$.

Games with coalition structure

Given $N = \{1, 2, \ldots, n\}$, we will denote by $B(N)$ the set of all partitions of N. Each $B \in B(N)$ is called a *coalition structure* in N, and a *union* each member of B. The so–called *trivial coalition structures* are $B^n = \{\{1\}, \{2\}, \ldots, \{n\}\}$ (individual coalitions) and $B^N = \{N\}$ (grand coalition). A *cooperative game with a coalition structure* is a pair $[v; B]$, where $v \in \mathcal{G}_N$ and $B \in B(N)$ for a given N. Each partition B gives a pattern of cooperation among players. We denote by $\mathcal{G}_N^{cs} = \mathcal{G}_N \times B(N)$ the set of all cooperative games with a coalition structure and player set N.

If $[v; B] \in \mathcal{G}_N^{cs}$ and $B = \{B_1, B_2, \ldots, B_m\}$, the *quotient game* v^B is the cooperative game played by the unions or, rather, by the *quotient set* $M = \{1, 2, \ldots, m\}$ of their representatives, as follows:

$$v^B(R) = v(\bigcup_{r \in R} B_r) \quad \text{for all } R \subseteq M.$$

By a *coalitional value* on $\mathcal{G}_N^{cs}$ we will mean a map $g : \mathcal{G}_N^{cs} \to \mathbb{R}^N$, which assigns to every pair $[v; B]$ a vector $g[v; B]$ with components $g_i[v; B]$ for each $i \in N$.

If f is a value on $\mathcal{G}_N$ and g is a coalitional value on $\mathcal{G}_N^{cs}$, it is said that g is a *coalitional value of* f iff $g[v; B^n] = f[v]$ for all $v \in \mathcal{G}_N$. For instance, the Owen value is a *coalitional value of the Shapley value* φ in the sense that $\Phi[v; B^n] = \varphi[v]$ for all $v \in \mathcal{G}_N$. Besides, $\Phi[v; B^N] = \varphi[v]$.

Mixed modified semivalues

Given two semivalues ψ and ϕ defined on games with n players (with possibly $\phi = \psi$), Albizuri and Zarzuelo [1] defined the concept of *mixed modified semivalue for games with coalition structure*. They also proved that every mixed modified semivalue is a coalitional semivalue and, reciprocally, every coalitional semivalue has this form.

If $[v; B] \in \mathcal{G}_N^{cs}$ and $B = \{B_1, B_2, \ldots, B_m\}$, the *modified quotient game* $v_{B_j|K}$ is defined as follows:

$$v_{B_j|K}(L) = v(\bigcup_{l \in L} B_l \setminus K') \quad \text{for all } L \subseteq M,$$

where $K' = B_j \setminus K$. This is the game played by the partition classes with the exception of B_j, that is replaced by the subset K. Given a semivalue $\psi \in \mathcal{G}_{\mathcal{N}}$, since the game $u_{B_j|K}$ is defined on a set M with m players ($1 \leq m \leq n$), we can apply the induced semivalue $\psi^{(m)}$:

$$w_j(K) = (\psi^{(m)})_j[u_{B_j|K}] \quad \forall K \subseteq B_j. \tag{5}$$

The value $w_j(K)$ shows the strategic position of the subset $K \subseteq B_j$ if this subset directly negotiate with the other classes as players in the quotient game –according to the semivalue induced by ψ– in absence of $K' = B_j \setminus K$.

Next, since the game w_j is defined on B_j, a set with $b_j = |B_j|$ players ($1 \leq b_j \leq n$), we can apply the induced semivalue $\phi^{(b_j)}$ and we define *the mixed semivalue* ψ/ϕ *modified* by the coalition structure B as

$$\big(\psi/\phi\big)_i[v; B] = (\phi^{(b_j)})_i[w_j] \quad \forall i \in B_j.$$

The mixed semivalue ψ/ϕ yields the result of a *two–step bargaining procedure* analogous to that used in [19, 21] and also in [2, 11]. Indeed, here we first apply the semivalue induced by ψ in the quotient game to get a payoff for each union; next, we use within union B_j the semivalue induced by ϕ to share the payoff by applying it to a *reduced game* played in this union.

If ψ and ϕ are respectively defined by the weighting coefficients $(q_s^n)_{s=0}^{n-1}$ and $(p_s^n)_{s=0}^{n-1}$, according to [1], the payoff given by the mixed modified ψ/ϕ semivalue to every player i in $B_j \in B$, is

$$\big(\psi/\phi\big)_i[v; B] = \sum_{R \subseteq M \setminus \{j\}} q_r^m \sum_{S \subseteq B_j \setminus \{i\}} p_s^{b_j} [v(Q \cup S \cup \{i\}) - v(Q \cup S)] \tag{6}$$

where $Q = \bigcup_{r \in R} B_r$, $b_j = |B_j|$, and $s = |S|$.

3. GENERALIZED MIXED MODIFIED SEMIVALUES

In this section we extend the notion of mixed modified semivalues values to games with a coalition structure and presents the *generalized mixed modified semivalues*, from now on GMMS. They apply a (induced) semivalue in the quotient game that arises once the coalition structure is actually formed, but share within each union the payoff so obtained by applying different (induced) semivalues to a game that concerns only the players of that union.

Definition. Let $[v;B] \in \mathcal{G}_N^{cs}$ and $B = \{B_1, B_2, \ldots, B_m\}$ a coalition structure in N. If ψ, φ_1, $\ldots$,φ_m are $m+1$ semivalues in $\mathcal{G}_N$, we define the GMMS $\psi/\varphi_1 \cdots \varphi_m$ by

$$\big(\psi/\varphi_1 \cdots \varphi_m\big)_i[v;B] = \big(\psi/\varphi_j\big)_i[v;B] \quad \forall i \in B_j \ (j = 1, \ldots, m). \tag{7}$$

From now on we will denote $\big(\psi/\varphi_1 \cdots \varphi_m\big)_i[v;B]$ by $\Phi_i[v;B]$.

Lemma. Let $[v;B] \in \mathcal{G}_N^{cs}$, $B = \{B_1, B_2, \ldots, B_m\}$ a coalition structure in N and Φ a GMMS defined by $m+1$ semivalues in $\mathcal{G}_N$. The allocations given by Φ to players belonging to a union B_j can be obtained as a linear combination of the allocations to unanimity games u_T, where $T = V \cup W$, $V \subseteq B_j$ and $W \in 2^{B \setminus B_j}$.

Notice that the number of unanimity games of this form is $(2^{b_j} - 1)2^m$ with $b_j = |B_j|$ and $m = |M|$.

Proposition. Let $B = \{B_1, B_2, \ldots, B_m\}$ be a coalition structure in N and Φ a GMMS defined by $m+1$ semivalues in $\mathcal{G}_N$, ψ, φ_1, $\ldots$,φ_m. Fixed a union B_j, the allocation to a player i belonging to B_j in a unanimity game u_T, $T = V \cup B_{i_1} \cup \cdots \cup B_{i_h}$, $V \subseteq B_j$ and $\{i_1, ..., i_h\} \subseteq M \setminus \{j\}$ is given by

$$\Phi_i[u_T;B] = \big(\psi/\varphi_j\big)_i[u_T;B] = \begin{cases} q_h^{h+1} p_{v-1}^{v} & i \in T \\ \\ 0 & i \notin T \end{cases}$$

where the weighting coefficients of the induced semivalues $(q_s^{h+1})_{s=0}^{h}$ and $(p_s^v)_{s=0}^{v-1}$ are obtained following (2) from the weighting coefficients $(q_s^n)_{s=0}^{n-1}$ and $(p_s^n)_{s=0}^{n-1}$ of the semivalues ψ and φ_j respectively.

Example. On the players set $N = \{1,2,3,4,5,6\}$, let $B = \{\{1,2,3\},\{4,5\},\{6\}\}$ be a coalition structure on N and Φ a GMMS defined by the semivalues in $\mathcal{G}_N$, ψ, φ_1, φ_2, and φ_3, where $\psi = \varphi$ is the Shapley value and φ_1 is defined by the weighting coefficients $(p_s^6)_{s=0}^5$=(1/24, 3/80, 1/30, 7/240, 1/40, 1/48).

We will obtain the allocations to players $i \in B_1$ according to Φ for the unanimity games $u_{\{1,2,4,6\}}$ and $u_{\{1,2,4,5,6\}}$. They are

$$\Phi_i[u_{\{1,2,4,6\}};B] = q_2^3\, p_1^2 = \frac{1}{3}\frac{7}{15} = \frac{7}{45}, \text{ for } i = 1,2 \text{ and}$$
$$\Phi_3[u_{\{1,2,4,6\}};B] = 0,$$

where $p_1^2 = \displaystyle\sum_{k=0}^{4} \binom{4}{k} p_{1+k}^6 = 7/15$ is the corresponding weighting coefficient of the induced semivalue.

In a similar way and according to Lemma, for $u_{\{1,2,4,5,6\}}$ we obtain

$$\Phi_i[u_{\{1,2,4,5,6\}};B] = q_2^3\, p_1^2 = \frac{1}{3}\frac{7}{15} = \frac{7}{45}, \text{ for } i = 1,2 \text{ and}$$
$$\Phi_3[u_{\{1,2,4,5,6\}};B] = 0,$$

Notice that the allocations in both games are the same because coalitions $\{1,2,4,6\}$ and $\{1,2,4,5,6\}$ intersect the same unions B_2 and B_3.

For players in B_1, their allocations does not depend on the remaining semivalues φ_2 and φ_3.

The computing technique based on the multilinear extension has been applied to many coalitional values: in 1992 to the Owen value by Owen and Winter [22]; in 1994 to the Owen–Banzhaf value by Carreras and Magaña [8]; in 1997 to the quotient game by Carreras and Magaña [9]; in 2003 to coalitional semivalues by Amer and Giménez [4]; in 2005 to the Alonso–Fiestras value

by Alonso, Carreras and Fiestras [3] or in 2011 to symmetric coalitional binomial semivalues by Carreras and Puente [11]. In next theorem we present a method to compute any GMMS by means of the multilinear extension of the game.

Theorem. Let $[v;B] \in \mathcal{G}_N^{cs}$, $B = \{B_1, B_2, \ldots, B_m\}$ a coalition structure in N and Φ a GMMS defined by $m+1$ semivalues in $\mathcal{G}_N$, ψ, φ_1, $\ldots$,φ_m with weighting coefficients $(q_s^n)_{s=0}^{n-1}$ and $(p_s^{r,n})_{s=0}^{n-1}$, $r = 1, ..., m$, respectively.

Then the following steps lead to the GMMS value of any player $i \in B_j$ in $[v;B]$.

1. Obtain the multilinear extension $f(x_1, x_2, \ldots, x_n)$ of game v.

2. For every $r \neq j$ and all $h \in B_r$, replace the variable x_h with y_r. This yields a new function of x_k for $k \in B_j$ and y_r for $r \in M\backslash\{j\}$.

3. In this new function, reduce to 1 all higher exponents, i.e. replace with y_r each y_r^q such that $q > 1$. This gives a new multilinear function denoted as $g_j((x_k)_{k\in B_j}, (y_r)_{r\in M\backslash\{j\}})$ (The modified multilinear extension of union B_j).

4. After some calculus, the obtained modified multilinear extension reduces to

$$g_j((x_k)_{k\in B_j}, (y_r)_{r\in M\backslash\{j\}}) = \sum_{V\subseteq B_j} \sum_{W\subseteq M\backslash\{j\}} \lambda_{V\cup W} \prod_{k\in V} x_k \prod_{r\in W} y_r$$

5. Multiply each product $\prod_{k\in V} x_k$ by p_{v-1}^v and each product $\prod_{r\in W} y_r$ by q_w^{w+1} obtaining a new multilinear function called $\overline{g}_j$.

6. Obtain the partial derivative of $\overline{g}_j$ with respect to x_i evaluated at point $(1, \ldots, 1)$ and

$$\Phi_i[v;B] = \frac{\partial \overline{g}_j}{\partial x_i}(1_{B_j}, 1_{M\backslash\{j\}}).$$

The following corollary shows as this procedure can be applied to compute semivalues allocations from the MLE of a cooperative game, in a similar way given by Carreras and Giménez [10].

Corollary. Let $v \in \mathcal{G}_N$ and Ψ a semivalue defined in $\mathcal{G}_N$ by the weighting coefficient p_s^n, $s = 0, ..., n-1$. Then the following steps lead to the allocation of any player $i \in N$ in game v.

1. Obtain the multilinear extension $f(x_1, x_2, \ldots, x_n)$ of game v as a sum of products

$$f(x_1, x_2, \ldots, x_n) = \sum_{S\subseteq N} \alpha_S \prod_{k\in S} x_k$$

2. Multiply each product $\prod_{k\in V} x_k$ by p_{v-1}^v obtaining a new multilinear function called $\overline{f}$.

3. Obtain the partial derivative of $\overline{f}$ with respect to x_i evaluated at point $(1, \ldots, 1)$ and

$$\Psi_i[v] = \frac{\partial \overline{f}}{\partial x_i}(1_N).$$

Example. Let $v \equiv [68; 50, 21, 20, 19, 13, 9, 3]$ be the 7–person weighted majority game and the coalition structure $B = \{\{1\}, \{2, 3, 5\}, \{4\}, \{6\}, \{7\}\}$. We will compute $\Phi[v;B]$, where the GMMS Φ is defined by 6 semivalues in $\mathcal{G}_N$, ψ, φ_1, $\ldots$,φ_5 with weighting coefficients $(q_s^7)_{s=0}^6$ and $(p_s^{r,7})_{s=0}^6$, $r = 1, ..., 5$, respectively.

The set of minimal winning coalitions of the game is

$$W^m(v) = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5, 6\}\},$$

so that players 2, 3 and 4 on one hand, and 5 and 6 on the other, are symmetric in v. Moreover, player 7 is null and the multilinear extension of v is

$$\begin{aligned} f(X_N) =& x_1x_2 + x_1x_3 + x_1x_4 - x_1x_2x_3 - x_1x_2x_4 - x_1x_3x_4 + x_1x_5x_6 + x_1x_2x_3x_4 \\ & - x_1x_2x_5x_6 - x_1x_3x_5x_6 - x_1x_4x_5x_6 + x_2x_3x_4x_5 + x_2x_3x_4x_6 - x_1x_2x_3x_4x_5 \\ & - x_1x_2x_3x_4x_6 + x_1x_2x_3x_5x_6 + x_1x_2x_4x_5x_6 + x_1x_3x_4x_5x_6 - x_2x_3x_4x_5x_6. \end{aligned}$$

The coalition structure is $B = \{\{1\}, \{2,3,5\}, \{4\}, \{6\}, \{7\}\}$ and steps 1–4 in Theorem give the modified multilinear extension of each union B_j, for $j = 1, 2, 3, 4$ (notice that player 7 is null in v and it is not necessary to compute g_5).

$$g_1(x_1, y_2, y_3, y_4, y_5) = x_1y_2 + x_1y_3 - 2x_1y_2y_3 + y_2y_3,$$

$$\begin{aligned} g_2(x_2, x_3, x_5, y_1, y_3, y_4, y_5) =& x_2y_1 + x_3y_1 + y_1y_3 - x_2x_3y_1 - x_2y_1y_3 - x_3y_1y_3 + x_5y_1y_4 + x_2x_3y_1y_3 \\ & - x_2x_5y_1y_4 - x_3x_5y_1y_4 - x_5y_1y_3y_4 + x_2x_3x_5y_3 + x_2x_3y_3y_4 - x_2x_3x_5y_1y_3 \\ & - x_2x_3y_1y_3y_4 + x_2x_3x_5y_1y_4 + x_2x_5y_1y_3y_4 + x_3x_5y_1y_3y_4 - x_2x_3x_5y_3y_4, \end{aligned}$$

$$g_3(x_4, y_1, y_2, y_4, y_5) = y_1y_2 + x_4y_1 + x_4y_2 - 2x_4y_1y_2,$$

$$g_4(x_6, y_1, y_2, y_3, y_5) = y_1y_2 + y_1y_3 + y_2y_3 - 2y_1y_2y_3.$$

Step 5 leads to $\overline{g}_j$ for each $j = 1, 2, 3, 4$.

$$\overline{g}_1(x_1, y_2, y_3, y_4, y_5) = p_0^{1,1}q_1^2x_1y_2 + p_0^{1,1}q_1^2x_1y_3 - 2p_0^{1,1}q_2^3x_1y_2y_3 + q_2^3y_2y_3,$$

$$\begin{aligned} \overline{g}_2(x_2, x_3, x_5, y_1, y_3, y_4, y_5) =& p_0^{2,1}q_1^2x_2y_1 + p_0^{2,1}q_1^2x_3y_1 + q_2^3y_1y_3 - p_1^{2,2}q_1^2x_2x_3y_1 - p_0^{2,1}q_2^3x_2y_1y_3 \\ & - p_0^{2,1}q_2^3x_3y_1y_3 + p_0^{2,1}q_2^3x_5y_1y_4 + p_1^{2,2}q_2^3x_2x_3y_1y_3 - p_1^{2,2}q_2^3x_2x_5y_1y_4 - p_1^{2,2}q_2^3x_3x_5y_1y_4 \\ & - p_0^{2,1}q_3^4x_5y_1y_3y_4 + p_2^{2,3}q_1^2x_2x_3x_5y_3 + p_1^{2,2}q_2^3x_2x_3y_3y_4 - p_2^{2,3}q_2^3x_2x_3x_5y_1y_3 - p_1^{2,2}q_3^4x_2x_3y_1y_3y_4 \\ & + p_2^{2,3}q_2^3x_2x_3x_5y_1y_4 + p_1^{2,2}q_3^4x_2x_5y_1y_3y_4 + p_1^{2,2}q_3^4x_3x_5y_1y_3y_4 - p_2^{2,3}q_2^3x_2x_3x_5y_3y_4, \end{aligned}$$

$$\overline{g}_3(x_4, y_1, y_2, y_4, y_5) = q_2^3y_1y_2 + p_0^{3,1}q_1^2x_4y_1 + p_0^{3,1}q_1^2x_4y_2 - 2p_0^{3,1}q_2^3x_4y_1y_2,$$

$$\overline{g}_4(x_6, y_1, y_2, y_3, y_5) = q_2^3y_1y_2 + q_2^3y_1y_3 + q_2^3y_2y_3 - 2q_3^4y_1y_2y_3.$$

Step 6 yields

$$\begin{aligned} \Phi_1[v; B] &= 2p_0^{1,1}q_1^2 - 2p_0^{1,1}q_2^3, \\ \Phi_i[v; B] &= p_0^{2,1}q_1^2 - p_1^{2,2}q_1^2 - p_0^{2,1}q_2^3 + p_1^{2,2}q_2^3 + p_2^{2,3}q_1^2 - p_2^{2,3}q_2^3, \qquad \text{for } i = 2, 3, \\ \Phi_4[v; B] &= 2p_0^{3,1}q_1^2 - 2p_0^{3,1}q_2^3, \\ \Phi_5[v; B] &= p_0^{2,1}q_2^3 - 2p_1^{2,2}q_2^3 - p_0^{2,1}q_3^4 + p_2^{2,3}q_1^2 - p_2^{2,3}q_2^3 + 2p_1^{2,2}q_3^4, \\ \Phi_6[v; B] &= 0, \\ \Phi_7[v; B] &= 0. \end{aligned}$$

4. CONCLUSIONS

The present work is focussed on the calculus of generalized modified mixed semivalues. More precisely, the computation of allocations to the players can be obtained from the multilinear extension by using a common procedure for all GMMS. As is well known, both the Shapley and Banzhaf values of any game v can be easily obtained from its multilinear extension. This latter procedure extends well to any p–binomial semivalue (see Puente [23], Freixas and Puente [14] or Amer and Giménez [4]).

In the context of games with a coalition structure, the multilinear extension technique has been also applied to computing the Owen value (Owen and Winter [22]), as well as the Owen–Banzhaf value (Carreras and Magaña [8]), the symmetric coalitional Banzhaf value (Alonso, Carreras and Fiestras [3]) and the symmetric coalitional binomial semivalue (Carreras and Puente [11]). In all these cases, (including the GMMS) the first three steps of the procedure are the same.

Instead, the consideration of the modified MLE g_j for the union B_j obtained from the initial one has changed the procedure: first, we weight the terms of g_j multiplying each product $\prod_{k\in V} x_k$ by p^v_{v-1} and each product $\prod_{r\in W} y_r$ by q^{w+1}_w obtaining a new multilinear function called $\overline{g}_j$. Second, we obtain players' marginal contributions by partial differentiation of $\overline{g}_j$. This new procedure has an advantage with respect to the traditional method: the allocations given by any coalitional semivalue are available since the weighting coefficients p^k_{k-1} and q^{k+1}_k can be always obtained.

Finally, the procedure presented here is a generalization of the method given by Carreras and Giménez [10] to compute semivalues allocations from the MLE of a cooperative game.

REFERENCES

[1] Albizuri, M.J. and Zarzuelo, J.M. (2004) On coalitional semivalues. Games and Economic Behavior, 49, 221–243.

[2] Alonso, J.M. and Fiestras, M.G. (2002) Modification of the Banzhaf value for games with a coalition structure. Annals of Operations Research, 109, 213–227.

[3] Alonso, J.M., Carreras, F. and Fiestras, M.G. (2005) The multilinear extension and the symmetric coalition Banzhaf value. Theory and Decision, 59, 111–126.

[4] Amer, R. and Giménez, J.M. (2003) Modification of semivalues for games with coalition structures. Theory and Decision, 54, 185–205.

[5] Amer, R., Carreras, F. and Giménez, J.M. (2002) The modified Banzhaf value for games with a coalition structure: an axiomatic characterization. Mathematical Social Sciences, 43, 45–54.

[6] Aumann, R.J. and Drèze, J. (1974) Cooperative games with coalition structures. International Journal of Game Theory, 3, 217–237.

[7] Carreras, F. (2004) α–decisiveness in simple games. Theory and Decision, 56, 77–91. Also in: Essays on Cooperative Games (G. Gambarelli, ed.), Kluwer Academic Publishers, 77–91.

[8] Carreras, F. and Magaña, A. (1994) The multilinear extension and the modified Banzhaf–Coleman index. Mathematical Social Sciences, 28, 215–222.

[9] Carreras, F. and Magaña, A. (1997) The multilinear extension of the quotient game. Games and Economic Behavior, 18, 22–31.

[10] Carreras, F. and Giménez, J.M. (2011) Power and potential maps induced by any semivalue: Some algrebraic properties and computation by multilinear extension. European Journal of Operational Research, 211, 148–159.

[11] Carreras, F. and Puente, M.A. (2011) Symmetric coalitional binomial semivalues. Group Decision and Negotiation, 21, 637–662.

[12] Dragan, I. (1997) Some recursive definitions of the Shapley value and other linear values of cooperative TU games. Working paper 328, University of Texas at Arlington, United States of America.

[13] Dubey, P., Neyman, A. and Weber, R.J. (1981) Value theory without efficiency. Mathematics of Operations Research, 6, 122–128.

[14] Freixas, J. and Puente, M.A. (2002) Reliability importance measures of the components in a system based on semivalues and probabilistic values. Annals of Operations Research, 109, 331–342.

[15] Giménez, J.M. (2001) Contribuciones al estudio de soluciones para juegos cooperativos (in Spanish). Ph.D. Thesis. Technical University of Catalonia, Spain.

[16] Harsanyi, J.C. (1959) A bargaining model for the cooperative n–person games. In: Contributions to the Theory of Games IV (A.W. Tucker and R.D. Luce, eds.), Princeton University Press, 325–356.

[17] Owen, G. (1972) Multilinear extensions of games. Management Science, 18, 64–79.

[18] Owen, G. (1975) Multilinear extensions and the Banzhaf value. Naval Research Logistics Quarterly, 22, 741–750.

[19] Owen, G. (1977) Values of games with a priori unions. In: Mathematical Economics and Game Theory (R. Henn and O. Moeschlin, eds.), Springer, 76–88.

[20] Owen, G. (1978) Characterization of the Banzhaf–Coleman index. SIAM Journal of Applied Mathematics, 35, 315–327.

[21] Owen, G. (1982) Modification of the Banzhaf-Coleman index for games with a priori unions. In: Power, Voting and Voting Power (M.J. Holler, ed.), 232–238.

[22] Owen, G. and Winter, E. (1992) Multilinear extensions and the coalitional value. Games and Economic Behavior, 4, 582–587.

[23] Puente, M.A. (2000) Aportaciones a la representabilidad de juegos simples y al cálculo de soluciones de esta clase de juegos (in Spanish). Ph.D. Thesis. Technical University of Catalonia, Spain.

[24] Shapley, L.S. (1953) A value for n-person games. In: Kuhn HW, Tucker AW (eds) Contributions to the theory of games II. Princeton University Press, Princeton, 307-317

[25] Weber, R.J. (1988) Probabilistic values for games. In: The Shapley Value: Essays in Honor of Lloyd S. Shapley (A.E. Roth, ed.), Cambridge University Press, 101–119.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

IMPLEMENTACIÓN DE FUNCIONES GENERATRICES PARA LA OBTENCIÓN DE ÍNDICES CON CONFIGURACIÓN DE COALICIONES

J. Rodríguez Veiga[1], G. I. Novoa-Flores[2] y B. Casas-Méndez[3]

[1] ITMATI, Technological Institute of Industrial Mathematics, Santiago de Compostela, Spain, jorge.rodriguez.veiga@gmail.com
[2] Department of Statistics and Operations Research, University of Santiago de Compostela, Santiago de Compostela, Spain, gnovoaflores@gmail.com
[3] Modestya Research Group, Department of Statistics and Operations Research, University of Santiago de Compostela, Santiago de Compostela, Spain, balbina.casas.mendez@usc.es

RESUMEN

En este trabajo, consideramos los índices de poder de Banzhaf-Coleman y Owen para los juegos de mayoría ponderada modificados por una configuración de coaliciones. Presentamos algoritmos de cálculo que hacen uso del método de las funciones generatrices. Por último, programamos el procedimiento en el lenguaje abierto *R* y se ilustra con un ejemplo real tomado de las ciencias sociales.

Palabras y frases clave: Juegos de mayoría ponderada, índices de poder, índice de Banzhaf-Coleman, índice de Shapley-Shubik, funciones generatrices, configuración de coaliciones.

1. INTRODUCCIÓN

En un juego cooperativo con utilidad transferible, también llamado juego TU, los agentes tienen mecanismos que les permiten tomar acuerdos vinculantes. La utilidad derivada de la cooperación de cualquier conjunto de agentes (o jugadores) puede ser transferida y dividirse de cualquier forma entre ellos. Uno de los objetivos principales de la teoría de juegos cooperativos es el estudio de los valores, es decir, las reglas de distribución que asignan a cada juego un vector, en el que cada coordenada del vector representa el pago asignado a cada uno de los jugadores. Hay numerosos valores, dos de los cuales, ampliamente utilizados, son el valor de Shapley (Shapley, 1953) y el valor de Banzhaf (Banzhaf, 1965).

Una familia importante de juegos TU está formado por los llamados juegos simples, que tienen aplicaciones interesantes, especialmente en el campo de la ciencia política. Dentro de la familia de los juegos simples, los juegos de mayoría ponderada juegan un papel relevante. Por ejemplo, un Parlamento puede ser visto como un juego de mayoría ponderada, en el supuesto de que los jugadores son los partidos políticos y las decisiones se toman por mayoría.

Cuando se trabaja con juegos simples, en lugar de hablar de valor, a menudo se utiliza la expresión índice de poder, puesto que los juegos simples se utilizan generalmente como modelos de órganos para la toma de decisiones en los que los acuerdos se toman por votación. El interés en estos juegos, por lo general, se centra en conocer el poder o influencia que tiene un jugador en el resultado final. En este contexto, el valor de Shapley y el valor de Banzhaf se denominan índice de poder de Shapley-Shubik (Shapley y Shubik, 1954) e índice de poder de Banzhaf-Coleman (Banzhaf, 1965 y Coleman, 1971), respectivamente. Numerosos trabajos en la literatura se dedican a la definición, estudio, cálculo y aplicaciones de índices de poder.

Los índices de poder se pueden obtener de una manera exacta o aproximada mediante el uso de diferentes herramientas. Uno de los métodos más utilizados para calcular los índices de manera exacta, en el caso de los juegos de mayoría ponderada, está basado en las llamadas funciones generatrices. Este método hace uso de una técnica de análisis combinatorio. En términos generales, una función generatriz es un

polinomio que permite enumerar el conjunto de posibles coaliciones, mientras que mantiene el control sobre sus respectivos pesos. Esto es muy útil, ya que permite obtener el valor exacto de los índices, incluso en los juegos con muchos jugadores, haciendo uso de algoritmos que se pueden programar con un lenguaje de ordenador. Esta técnica fue utilizada por David G. Cantor (1962) (cf. Lucas, 1983) para el índice de poder de Shapley-Shubik y Brams y Affuso (Brams y Affuso, 1976}) para calcular el índice de poder de Banzhaf-Coleman.

Un modelo más general de juegos TU añade una estructura coalicional adicional, esto es, una partición definida sobre el conjunto de jugadores, ya que éstos podrían tener algunas preferencias para unirse a los demás motivados por algunas razones como características ideológicas, o ubicación geográfica. Por esta razón, Owen (Owen, 1977) propuso un valor para juegos TU con estructura coalicional, que generaliza al valor de Shapley, mientras que el índice de poder de Banzhaf-Owen (Owen, 1982}) se propuso para juegos simples con estructura coalicional como una extensión del índice de poder de Banzhaf-Coleman. Alonso-Meijide y Bowles (2005) utilizaron funciones generatrices para calcular los valores de Owen y Banzhaf-Owen para juegos de mayoría ponderada.

No obstante, este modelo de juegos TU con estructura coalicional podría no ser apropiada en algunas situaciones en las que los jugadores podían prefieren unirse a algunos jugadores por algunas razones y unirse a otros por otras razones. Ilustremos esta situación con un ejemplo tomado de Andjiga y Courtin (2015): consideremos las relaciones diplomáticas entre los países. En la vida real, los países están organizados en coaliciones internacionales no disjuntas necesariamente. Por ejemplo, Francia y España, entre otros, pertenecen a la Unión Europea y la Organización del Tratado del Atlántico Norte (OTAN), mientras que los Estados Unidos pertenece a la OTAN y el NAFTA (una coalición con México y Canadá). Por esta razón, se introdujo el modelo más general de juegos con configuración coaliciones. En este modelo, se considera un recubrimiento del conjunto de jugadores en lugar de una partición de la misma. Albizuri y Aurrekoetxea (2006) y Albizuri, Aurrekoetxea y Zarzuelo (2006) proponen el índice de Banzhaf-Coleman generalizado y el denominado valor de configuración, para los modelos de juegos simples y juegos TU, respectivamente, con una configuración de coaliciones. Mediante estas reglas, se generalizan el índice de poder de Banzhaf-Coleman y el valor Owen, respectivamente.

En el presente trabajo, se aplica el método de las funciones generatrices para calcular el índice de Banzhaf-Coleman generalizado y el valor de configuración para la familia de juegos de mayoría ponderada con una configuración de coaliciones, de tal manera que extendemos los resultados de Lucas (1983), Brams y Affuso (1976), y Alonso-Meijide y Bowles (2005). Se ilustran los algoritmos con un pequeño ejemplo numérico y otro tomado de la vida real. También se muestra un código de *R* que implementa los nuevos algoritmos.

1. EL CASO DE LA UNIÓN EUROPEA DE LOS 28 MIEMBROS (2015)

Hemos calculado los índices generalizado de Banzhaf-Coleman y de los diferentes países que constituyen la Unión Europea (UE). Desde el 1 de julio de 2013, esta comunidad cuenta con 28 miembros.

Si en la UE identificamos cada país con un jugador cuyo peso es el número de escaños, y la cuota viene dado por la mayoría simple, podemos modelar esta situación como un juego de mayoría ponderada.

Además de esto, proponemos dos particiones de los jugadores para representar una situación más realista. Para la definición de estas particiones, seguiremos criterios inspirados en los trabajos de Albizuri y Aurrekoetxea (2006) y Courtin (2011). La primera partición, llamémosle P_1, tendrá en cuenta las posiciones culturales y geográficos de los estados de una manera que agrupará los situados en las proximidades del Mediterráneo y el Atlántico, los que están en el centro de Europa, en el Norte, y a una parte de los nuevos miembros de la UE.

La otra partición, P_2, tendrá en cuenta la situación económica de los estados, y para ello se utiliza la lista de estados soberanos en Europa según el PIB (PPA) per cápita (listas de países según el valor de todos los bienes y servicios finales producidos en cada país durante 2013, dividido por la población media para el mismo año), proporcionado por el *Banco Mundial*.

Como una configuración de coaliciones permite a los jugadores estar en más de una coalición a la vez,

definimos nuestra configuración de coaliciones como la unión de P_1 y P_2.

En cuanto a los resultados del índice generalizado de Banzhaf-Coleman, entre otros, observamos que los dos valores más altos corresponden a Francia y Alemania, que también son los estados con el mayor número de diputados. También pertenecen a la misma zona geográfica y comparten la posición económica. Esto último también ocurre con Chipre y Malta, que también pertenecen a los estados con menor número de diputados, y son los dos estados con menor valor. Dos nuevos miembros, Croacia y Lituania, que tienen el mismo poder, tienen el mismo número de diputados y posición económica similar. Grecia, Portugal, Bélgica y la República Checa tienen el mismo número de diputados y una posición económica similar en términos de PIB, pero el índice asignado es inferior en los dos primeros, lo cual está motivado porque las hemos incluido en un área geográfica diferente.

Si ahora analizamos los resultados con el índice de configuración, observamos relaciones similares, pero no idénticos, ya que en su definición, este índice tiene en cuenta el tamaño de ciertas coaliciones significativas para cada jugador. Según este índice, el país con menos poder es Finlandia, que no es el Estado que menos diputados tienen, pero lo hemos incluido en un área geográfica con sólo dos estados. Comparando los resultados de Italia y el Reino Unido, el primero tiene más poder, según el índice de configuración y menos con el índice generalizado de Banzhaf-Coleman; ambos tienen los mismos diputados y están en la misma región, pero son de diferentes coaliciones en función del nivel económico. Estonia e Irlanda también se ordenan de una u otra forma según el índice considerado; en este caso, el número de diputados, la zona geográfica y la situación económica son diferentes.

3. IMPLEMENTACIÓN EN *R* DE LOS ALGORITMOS

Hemos implementado dos funciones en el lenguaje abierto *R* que calculan los índices generalizado de Banzhaf-Coleman y de configuración, para los juegos de mayoría ponderada con configuración de coaliciones usando nuestros algoritmos creados a partir de funciones generatrices.

La razón de usar *R* es que es una herramienta de software libre, cada vez más empleada en el ámbito de la estadística, investigación de operaciones, y sus aplicaciones. *R* permite crear listas a partir de un conjunto de datos y tiene diferentes comandos que permiten su manipulación fácil y rápida. Las funciones que hemos creado proporcionan un resultado siempre que se proporciona como entrada el vector de pesos de los jugadores en el juego de mayoría ponderada, la cuota, y las coaliciones que constituyen la configuración.

4. CONCLUSIONES

Desde que Albizuri y Aurrekoetxea (2006) y Albizuri, Aurrekoetxea y Zarzuelo (2006) introdujeron el índice de Banzhaf-Coleman generalizado y el valor de configuración, respectivamente, para los juegos simples con configuración de coaliciones, como complemento de la definición y el estudio de sus propiedades hechas en estos artículos, una cuestión abierta era el cálculo eficiente de estos índices por medio de las llamadas funciones generatrices, en línea con estudios similares relativos, entre otros, a los conocidos índices de Shapley- Shubik y Banzhaf-Coleman. En este trabajo la tarea se lleva a cabo, y la construcción de algoritmos computacionales, cuya idoneidad se demuestra matemáticamente, se complementa con la programación informática usando una herramienta de software libre como es *R* y la presentación de ejemplos, tanto puramente numéricos con el fin de explicar los algoritmos, como tomados de la vida real, que muestran el alcance del modelo considerado y los algoritmos introducidos.

Se pueden solicitar a los autores los algoritmos detallados, junto con ejemplos, la implementación en *R* y el caso considerado de la Unión Europea.

Creemos que vale la pena señalar que el modelo de los juegos cooperativos con configuración de coaliciones estudiados en este trabajo, sigue siendo de interés a día de hoy para los investigadores, como Albizuri y Vidal-Puga (2015) o Andjiga y Courtin (2015), entre otros. Esto significa que los algoritmos introducidos en este artículo pueden ser objeto de revisión o adaptación en futuros trabajos.

AGRADECIMIENTOS

Esta investigación ha recibido apoyo financiero del Ministerio de Ciencia e Innovación y el Ministerio de Economía y Competitividad a través de los proyectos MTM2011-27731-C03 y MTM2014-53395-C3-2-P, respectivamente.

REFERENCIAS

Albizuri, M. J. and Aurrekoetxea, J. (2006) Coalition configurations and the Banzhaf index. Social Choice and Welfare 26, 571-596.

Albizuri, M. J., Aurrekoetxea, J., and Zarzuelo, J. M. (2006) Configuration values: Extensions of the coalitional Owen value. Games and Economic Behavior 57, 1-17.

Albizuri, M. J. and Vidal-Puga, J. (2015) Values and coalition configurations. Mathematical Methods of Operations Research 81, 3-26.

Alonso-Meijide, J. M. and Bowles, C. (2005) Generating functions for coalitional power indices: An application to the IMF. Annals of Operations Research 137, 21-44.

Andjiga, N. G. and Courtin, S. (2015) Coalition configurations and share functions. Annals of Operations Research 225, 3-25.

Banzhaf III, J. F. (1965) Weighted voting does not work: A mathematical analysis. Rutgers Law Review 19, 317-343.

Brams, S. F. and Affuso, P. J. (1976) Power and size: A new paradox. Theory and Decision 7, 29-56.

Coleman, J. Control of collectivities and the power of a collectivity to act. In: B. Lieberman (Ed.), Social Choice, Gordon and Breach, 1971, pp. 269-300.

Courtin, S. (2011) Power in the European Union: An evaluation according to a priori relations between states. Economics Bulletin 31, 534-545.

Lucas, W. F. Measuring power in weighted voting system. In: S. Brams, (Ed.), Political and Related Models, Springer-Verlag, 1983.

Owen, G. Values of games with a priori unions. In: R. Henn and O. Moeschlin (Eds.), Mathematical Economics and Game Theory, Springer, 1977, pp. 76-88.

Owen, G. Modification of the Banzhaf-Coleman index for games with a priori unions. In: M.J. Holler (Ed.), Power, Voting and Voting Power, Physica-Verlag, 1982, pp. 232-238.

Shapley, L. S. A value for n-person games. In: A.W. Tucker and H. Kuhn (Eds.), Annals of Mathematical Studies, Vol. 28, Princeton University Press, 1953, pp. 307-317.

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

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

Juegos cooperativos e índices de cohesión: El valor de Choquet-Owen

M. G. Fiestras-Janeiro[1], J. M. Gallardo[2], A. Jiménez-Losada[2] y M. A. Mosquera[1]

[1]Universidade de Vigo
[2]Universidad de Sevilla

RESUMEN

En un juego cooperativo con utilidad transferible es habitual asumir que todas las coaliciones son factibles. Sin embargo, en un juego cooperativo con una configuración de coaliciones, debido a las preferencias de los jugadores, sólo algunas coaliciones son a priori factibles. En este trabajo proponemos una generalización de los juegos con una configuración de coaliciones. En nuestro modelo la factibilidad a priori de las coaliciones viene determinada por la cohesión de sus miembros y, obviamente, esta cohesión no tiene por qué ser igual para todas las coaliciones. La cohesión de cada coalición vendrá determinada por un índice de cohesión. Se introduce la clase de juegos con un índice de cohesión y se propone una regla de asignación, la cual también se caracteriza utilizando propiedades razonables. Esta idea de cohesión, además de utilizarse en psicología, también se utiliza en otras áreas de conocimiento como el diseño del software. En dicha área, la cohesión explica la relación entre todos los elementos de un módulo de programación. La regla de asignación aquí definida puede ser aplicada a este área tal y como se muestra en este trabajo.

Palabras y frases clave: Teoría de juegos, configuración de coaliciones, cohesión de grupos, cohesión de software.

Comunicacións orais

Sesión 2 de Investigación de Operacións

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Un estudio comparativo de la estimación noparamétrica de la función de intensidad en procesos puntuales sin marcas

Borrajo M.I.[1], González-Manteiga W.[1] y Martínez-Miranda M.D.[2]

[1]Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela.
[2]Departamento de Estadística e Investigación Operativa. Universidad de Granada.

RESUMEN

Los procesos puntuales espaciales son modelos estocásticos complejos que tratan de describir la localización de eventos de interés, así como otra información adicional a través de marcas o covariables. La función de intensidad de primer orden es uno de los principales elementos caracterizadores de los procesos puntuales, y como tal, fue objeto de estudio a lo largo de los años.

En este trabajo se hace una revisión de la estimación de dicha función, centrándose especialmente en técnicas nonparamétricas, y se describe además los resultados de un estudio de simulación comparativo entre las propuestas más relevantes. Para finalizar, se muestra una aplicación de estas técnicas al análisis con datos de incendios forestales en Galicia.

Palabras y frases clave: procesos puntuales, función de intensidad, estimación noparamétrica, incendios forestales.

DESCRIPCIÓN DEL TRABAJO

Los procesos puntuales pueden definirse intuitivamente como modelos matemáticos que describen la disposición de objetos distribuidos de manera irregular o aleatoria en el espacio. Esta metodología se ha aplicado en numerosos campos de investigación como ecología (distribución de especies vegetales), biología molecular (distribución de un determinado tipo de células), astronomía (distribución de estrellas o galaxias), epidemiología (distribución de casos de enfermedades)...

Formalmente, un proceso puntual espacial, X en $\mathbb{R}^2$ es un mecanismo estocástico que genera una cantidad numerable de eventos en el plano. Un patrón puntual espacial, $(X_1, \ldots, X_N)$ es una realización de un proceso puntual en una región acotada $W \subset \mathbb{R}^2$, denominada región de observación; cada uno de los elementos de esa realización se denomina evento. Es importante tener en cuenta que el número de eventos, N, varía con cada realización, por consiguiente, tanto dicho número como las localizaciones son aleatorios. Este último dato es un factor a tener en cuenta en el estudio de los procesos puntuales y determina una diferencia fundamental con los procedimientos de muestreo clásicos.

Existen diversas condiciones de regularidad que se pueden exigir a los procesos puntuales y que, en caso de cumplirse, facilitan enormemente los desarrollos teóricos. Las principales son estacionariedad e isotropía; la primera de ellas se traduce en que las propiedades del proceso son invariantes por traslaciones; mientras que la segunda implica que dichas propiedades sean invariantes por rotaciones. Por lo tanto, las propiedades de un proceso que sea simultáneamente estacionario e isotrópico, dependerán únicamente de la longitud del vector diferencia de posiciones (un valor escalar), en lugar de las localizaciones en sí mismas.

Otro concepto importante en el ámbito de los procesos puntuales es el de CSR (*Complete Spatial Randomness*), que puede entenderse como el "ruido blanco" de un proceso puntual, ya que se caracteriza por la ausencia de estructura en los datos, es decir, es lo que ocurre cuando los eventos se distribuyen de manera totalmente aleatoria. Esta propiedad suele emplearse como hipótesis nula en contrastes de hipóstesis para determinar si existe o no estructura significativa en un determinado patrón.

Definición 1. Un proceso puntual cumple la condición de CSR si verifica los siguientes postulados:

- El número de eventos, N, en una región W, sigue una distribución $Pois(\lambda|W|)$ donde λ es la intensidad del proceso, es decir, el número de eventos por unidad de área y $|W|$ denota la medida de la región W, en el caso de proceso espacial, será el área.
- Condicionado a una realización con $N = n$ eventos, $\{X_1, \ldots, X_n\}$, es una muestra aleatoria simple de una distribución $Unif(W)$.

Atendiendo a la información adicional de la que se dispone, tendremos diferentes tipos de procesos puntuales. Si la información adicional va ligada a un evento, es decir, que tiene que tener lugar dicho evento para que podamos tener esa información, estariamos ante un proceso puntual marcado, como por ejemplo los incendios forestales en Galicia y la superficie quemada en cada uno de ellos. Si se tiene información completa sobre toda la región de observación, entonces tendríamos una covariable. Este sería el caso de la temperatura máxima registrada en un día, en el problema los incendios.

Los procesos puntuales se caracterizan fundamentalmente por dos clases de funciones, las de primer orden, entre las que destaca la función de intensidad, y las de segundo orden, que describen la estructura de dependencia de los datos.

Definición 2. Dado un proceso puntual $X \subset W$ y N un proceso de recuento tal que, $N(A)$ con $A \subset W$, indica el número de eventos que hay en A; se define la función de intensidad o intensidad de primer orden como

$$\lambda(x) = \lim_{|dx|\to 0} \frac{\mathbb{E}[N(dx)]}{|dx|}. \tag{1}$$

Definición 3. Se define la intensidad de segundo orden como

$$\lambda_2(x, y) = \lim_{|dx|,|dy|\to 0} \frac{\mathbb{E}[N(dx)N(dy)]}{|dx||dy|}. \tag{2}$$

Haciendo un paralelismo con los planteamientos más tradicionales de la estadística, en la que la hipótesis de normalidad está muy extendida y permite simplificar notablemente los desarrollos metodológicos, en procesos puntuales este papel lo juegan los procesos de Poisson.

Definición 4. Un proceso puntual X se dice Poisson homogéneo si verifica las mismas condiciones que CSR (ver Definición 1).

Definición 5. Un proceso puntual se dice Poisson no-homogéneo si cumple:

- El número de eventos en W sigue una distribución $Pois(\int_W \lambda(x)dx)$.
- Dados n eventos, estos forman una muestra independente de una distribución en W con densidad proporcional a $\lambda(x)$.

Bajo la condición de proceso de Poisson no-homogéneo se ha desarrollado la mayor parte de la metodología estadística en este campo. El estudio de la función de intensidad en el caso homogéneo es trivial, dado que dicha función sería una constante que representaría el número de eventos por unidad de área. Un estimador natural de la función de intensidad de un proceso de Poisson homogéneo sería el número de eventos observados dividido entre el área de la región de observación.

Nuestro objetivo en este trabajo es la estimación de la función de intensidad (intensidad de primer orden) de procesos puntuales no homogéneos.

La estimación paramétrica de la función de intensidad empleando técnicas de máxima verosimilitud data de los años 60 y 70, ver Illian et al (2008). Esta metodología ha sido muy utilizada debido a sus buenas propiedades asintóticas, sin embargo no siempre es sencillo, o incluso posible, describir adecuadamente la función de intensidad subyacente con un modelo paramétrico. En estos casos las técnicas noparamétricas ofrecen una aproximación más flexible sin necesidad de postular modelos que pueden conducir a conclusiones erróneas.

La propuesta inicial a este respecto fue introducida por Diggle (1985), artículo en el que se define un estimador tipo núcleo de la intensidad para procesos en la recta real; su extensión al caso bidimensional fue inmediata:

$$\hat{\lambda}_h(x) = \frac{1}{p_h(x)} \sum_{i=1}^{N} k_h(x - X_i) = \frac{1}{p_h(x)\, h^2} \sum_{i=1}^{N} k\left((x - X_i)/h\right), \tag{3}$$

donde la función núcleo $k(\cdot)$, es una densidad bivariante radialmente simétrica, $h > 0$ es la ventana o parámetro de suavizado, k_h es el núcleo reescalado y $p_h(x) = \int_W h^{-2} k((x-y)/h)dy$ es la corrección de efecto frontera, siendo como hasta ahora W la región de observación.

Este estimador recuerda al estimador tipo núcleo de la densidad, en el que se incluye esa corrección del efecto frontera necesaria en el campo de los procesos puntuales, ya que se trabaja en una región acotada W.

El principal problema del estimador de Diggle (1985) es que es inconsistente, ya que su varianza no tiende a cero al incrementar el número esperado de eventos. Recordemos que en procesos puntuales el número de eventos es aleatorio, por lo que el elemento a controlar es el número esperado de eventos, $m = \int_W \lambda(x)dx$, provocando que el ECMI (*Error Cuadrático Medio Integrado*) no converja a cero.

La inconsistencia del estimador de Diggle ha motivado dos aproximaciones noparamétricas diferentes que dan lugar a estimadores consistentes de la función de intensidad. La primera fue descrita por Cucala (2006) y está basada en la estimación de la densidad, y la segunda ha sido iniciada por Guan (2008) y seguida más recientemente por Baddeley (2012), basada en el uso de covariables.

Cucala (2006), inspirándose en el resultado de Diggle y Marron (1988) que prueba la convergencia de ECMI/m^2, definió la densidad de localización de eventos, "*density of events location*", como $f(x) = \dfrac{\lambda(x)}{\int_W \lambda(s)ds} \equiv \dfrac{\lambda(x)}{m}$, y propuso el siguiente estimador de esta función:

$$\hat{\lambda}_{0,h}(x) = \frac{1}{N} \sum_{i=1}^{N} k\left(\frac{x - X_i}{h}\right) 1_{\{N \neq 0\}}.$$

Retomando el problema inicial, el estimador de la función de intensidad $\lambda(x)$ estaría dado por $\hat{\lambda}_h(x) = N\hat{\lambda}_{0,h}(x)$. La extensión al caso espacial es análoga a la hecha para el estimador (3).

En el contexto espacial, Fuentes et al. (2015) propusieron un estimador en esta misma línea pero considerando una ventana matricial completa, que permite captar de manera más adecuada la estructura de los datos en $\mathbb{R}^2$:

$$\hat{\lambda}_{0,H}(x) = \frac{\hat{\lambda}_H(x)}{N} I[N \neq 0] = (p_H(x)N)^{-1} |H|^{-1/2} \sum_{i=1}^{N} k\left(H^{-1/2}(x - X_i)\right) I[N \neq 0], \tag{4}$$

donde la ventana H es una matriz simétrica y definida positiva.

La otra forma descrita hasta el momento para solventar el problema de la inconsistencia es el uso de covariables; sea de aquí en adelante $\mathbf{Z}$ un campo aleatorio continuo en $\mathbb{R}^2$. En la práctica la

continuidad se traduce en el conocimiento de la(s) covariable(s) en una rejilla suficientemente fina. Bajo la hipótesis de que la función de intensidad depende de ese campo de covariables a través de una función no negativa y continua, ρ, desconocida, es decir, $\lambda(x) = \rho(\mathbf{Z}(x))$, Guan (2008) propuso el siguiente estimador:

$$\hat{\lambda}(x;h) = \frac{\sum_{i=1}^{N} k\left(\frac{||\mathbf{Z}(X_i)-\mathbf{Z}(x)||}{h}\right)}{\int_W k\left(\frac{||\mathbf{Z}(u)-\mathbf{Z}(x)||}{h}\right) du}, \tag{5}$$

siendo k una densidad univariante y $||\cdot||$ la norma euclídea. En ese mismo trabajo, el autor prueba la consistencia de este estimador.

La aportación de Baddeley et al. (2012) en esta línea es más completa, presentan una serie de estimadores tipo núcleo de la función de intensidad, entre los que destacamos:

$$\hat{\rho}_R(x) = \frac{1}{g^*(x)} \sum_i k_h\left(\mathbf{Z}(x_i) - \mathbf{Z}(x)\right), \tag{6}$$

donde $G(z) = \frac{1}{|W|}\int_W 1_{\{Z(u)\leq z\}} du$ conocida como "*spatial cumulative distribution function*", g es su derivada, y $G^*(x) = |W|G(x)$ y $g^*(x) = |W|g(x)$ son las versiones sin normalizar.

Este estimador puede verse como una generalización de (5) para el caso de covariable unidimensional; basta con tener en cuenta que en la práctica la función g^* no es conocida de manera continua sobre W, sino en una rejilla, y se aproxima usando estimación tipo núcleo, por lo que el resultado en la práctica es el mismo que el de Guan (2008).

En este trabajo realizamos un estudio de simulación que permite comparar los estimadores definidos en (3), (4) y (6) en distintos escenarios, lo que nos ha permitido ver las ventajas de unos estimadores sobre otros, en función de la información adicional en forma de covariables de la que se dispone.

El comportamiento de los estimadores se ha cuantificado a través del error cuadrático medio integrado (ECMI). Los tres estimadores comparados dependen de parámetros de suavizado (ventana), que deben ser elegidos adecuadamente en la práctica. En nuestro análisis hemos considerado selectores automáticos de dichos parámetros, basados en los métodos que fueron propuestos junto con los propios estimadores de la función de intensidad por sus respectivos autores. En el caso del estimador de Diggle se trata de un procedimiento de validación cruzada. En el caso de la extensión espacial de Cucala de un procedimiento plug-in y en el caso de Baddeley la tradicional regla del pulgar de Silverman. Además hemos analizado el problema de selección de dicho parámetro para cada estimador considerando los selectores teóricos óptimos como referencia.

Como ya se comentaba al comienzo de este trabajo, la aplicación de este tipo de técnicas ha sido muy amplia, nosotros describimos una ilustración práctica con datos de incendios forestales de Galicia de agosto del 2006.

Es de todos conocida la gran repercusión medioambiental y socio-económica de los incendios forestales en Galicia, de ahí que sea un tema de interés desde el punto de vista de diversos ámbitos, también desde la perspectiva de la estadística espacial, y más concretamente en este trabajo a través de la metodología de los procesos puntuales.

Hemos analizado lo que ocurrió en un día particular, el 09/08/2006, ese verano fue especialmente famoso por la cantidad de incendios registrados así como por la superficie arbolada quemada. Durante ese día concreto se registraron en nuestra comunidad un total de 190 incendios, que serán los eventos. Hemos considerado como covariables la temperatura media y la temperatura máxima alcanzadas durante ese día. Como ya se ha mencionado anteriormente, la covariable ha de ser de naturaleza continua y en la práctica conocida en una rejilla suficientemente fina; puesto que en esa fecha solo se disponía de 44 estaciones de medición repartidas en la comunidad, hemos hecho un

suavizado con núcleo gaussiano para ambas covariables que nos permitiese tener los datos necesarios .

En la Figura 1 podemos ver los puntos de ignición en la gráfica de la izquierda, la covariable temperatura media en el centro y la covariable temperatura máxima a la derecha; estas últimas resultantes tras la suavización antes mencionada.

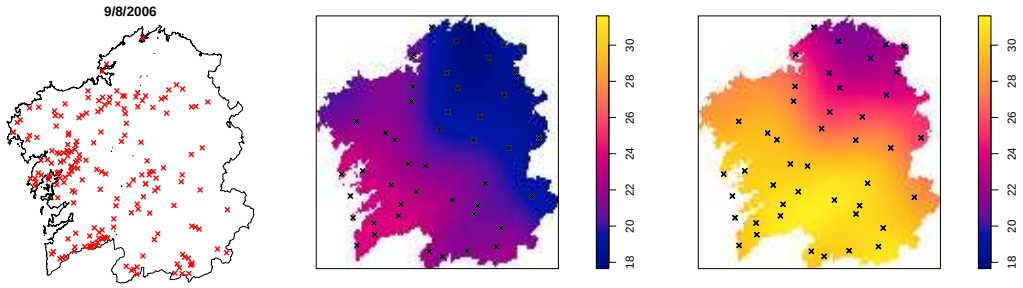


Figura 1: Puntos de ignición de los incendios el día 09/08/2006 (izquierda), gráfica de la temperatra media (centro) y temperatura máxima (derecha) durante ese mismo día.

Calculamos los tres estimadores para ese día, con los correspondientes selectores automáticos de la ventana que han sido analizados en el estudio de simulación.

En líneas generales, se observa que el estimador de Diggle y la extensión al caso espacial de Cucala, reproducen en mayor o menor medida el patrón de los puntos de ignición, aunque la escala en la que los captan pudiera ser mejorable. Sin embargo, el estimador de Baddeley no es capaz de captar esta estructura de los eventos, sino que parece dejarse influir por la covariable y tiende a reproducir el patrón de ésta. Esto permite concluir que la temperatura no es el principal factor de riesgo en los incendios en Galicia, lo que concuerda con el hecho constatado de que la mayor parte de los incendios forestales en Galicia están catalogados como “provocados”.

En esta línea sería interesante poder considerar otras covariables de tipo socio-económico en relación por ejemplo con la distribución de los núcleos urbanos o la situación de las brigadas antiincendios.

Como trabajo en curso, y basándose en la comparación de estimadores consistentes (uno empleando información de covariables y otro sin ella), se pretende definir un contraste que permita determinar si una covariable es o no un factor de riesgo para un determinado proceso.

Agradecer al Ministerio de Economía y Competitividad su apoyo económico a través del proyecto MTM2013-41383P, que incluye financiación del Fondo Europeo de Desarrollo Regional (FEDER) y a la red IAP 97/06 StUdyS del gobierno belga. El trabajo de M.I. Borrajo ha sido financiado por una beca FPU, FPU13/00473, del Ministerio de Educación, Cultura y Deporte.

REFERENCIAS

Baddeley, A., Chang, Y. M., Song, Y. y Turner, R. (2012) Nonparametric estimation of the dependence of a spatial point process on spatial covariates. Statistics and its Interface, 5(2), 221-236.

Cucala L. (2006) Espacements bidimensionnels et données entachées d’erreurs dans l’analyse des procesus ponctuels spatiaux. PhD Thesis, Université des Sciencies de Toulouse I.

Diggle P.J. (1985) A Kernel Method for Smoothing Point Process Data. Journal of the Royal Statistical Society (Series C), 34, 138-147.

Diggle P.J. y Marron J.S. (1988) Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation. Journal of the American Statistical Association, 83, 793-800.

Fuentes-Santos, I., González-Manteiga, W. y Mateu, J. (2015). Consistent smooth bootstrap kernel intensity estimation for spatial inhomogeneous poisson point processes. En revisión.

Guan, Y. (2008) On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. Journal of the American Statistical Association, 103(483), 1238-124.

Illian, J., Penttinen, A., Stoyan, H. y Stoyan, D. (2008). Statistical analysis and modelling of spatial point patterns (Vol. 70). John Wiley & Sons.

Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Inferencia sobre el variograma mediante técnicas bootstrap en procesos espaciales estacionarios y no estacionarios

S. A. Castillo Páez[1], Rubén Fernández Casal[2], P. García Soidán[3]

[1]sacastillo@espe.edu.ec, Universidad de las Fuerzas Armadas ESPE
[2]ruben.fcasal@udc.es, Universidade A Coruña
[3]pgarcia@uvigo.es, Universidad de Vigo

RESUMEN

La fiabilidad de los métodos de predicción e inferencia en procesos espaciales reside en gran medida en una apropiada caracterización de la dependencia espacial, a través de la estimación del variograma. Los estimadores empleados habitualmente con este fin presentan ciertas limitaciones, en términos del sesgo o la variabilidad, sobre todo en el caso de procesos no estacionarios. En este trabajo se propone un método bootstrap no paramétrico, que permite realizar inferencias a a partir del estimador empírico y lineal local del variograma en procesos con tendencia nula y tendencia no constante. Se ilustrará su funcionamiento mediante estudios de simulación y se compararán los resultados obtenidos por este procedimiento con los que proporcionan otros mecanismos de remuestreo utilizados en este contexto.

Palabras e frases chave: Métodos de remuestreo, estimador lineal local, variograma

1. INTRODUCIÓN

Supongamos que observamos datos de un proceso espacial $\left\{Y(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^d\right\}$ que puede ser modelado como:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{1}$$

donde $\mu(\mathbf{s})$ es la tendencia o variación a gran escala, y el término $\varepsilon(\mathbf{s})$ es el error o variación a pequeña escala. Se trabajará sobre la hipótesis de que la tendencia es determinística y que el error $\left\{\varepsilon(\mathbf{s}), \mathbf{s} \in D \subset \mathbb{R}^d\right\}$ es un proceso estacionario de segundo orden con media nula y covariograma $C(\mathbf{h}) = Cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h}))$, con $\mathbf{h} \in \mathbb{R}^d$. Usualmente, la dependencia del proceso espacial puede ser determinada a partir del variograma, definido por:

$$\gamma(\mathbf{h}) = \frac{1}{2} Var\left(Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})\right)^2, \tag{2}$$

Por simplicidad, habitualmente se suele asumir isotropía, es decir: $\gamma(\mathbf{h}) = \gamma(||\mathbf{h}||)$. Para obtener una estimación de (2), se suele partir de la información que proporciona algún estimador piloto, no paramétrico, que posteriormente se utiliza para ajustar un modelo válido de variograma. Entre las alternativas no paramétricas para la caracterización del variograma, se puede recurrir a la metodología lineal local, que en el caso isotrópico daría lugar al siguiente estimador:

$$\hat{\gamma}(u) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(u) \left(Y(\mathbf{s}_i) - Y(\mathbf{s}_j)\right)^2}{2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(u)} \tag{3}$$

donde:

$$w_{ij}(u) = K\left(\frac{\|\mathbf{s}_i - \mathbf{s}_j\| - u}{g}\right) \times$$

$$\sum_{k<l} K\left(\frac{\|\mathbf{s}_k-\mathbf{s}_l\|-u}{g}\right)(\|\mathbf{s}_k-\mathbf{s}_l\|-u)\,(\|\mathbf{s}_k-\mathbf{s}_l\|-\|\mathbf{s}_i-\mathbf{s}_j\|)$$

siendo u el salto (o distancia) entre localizaciones, K una función de densidad tipo núcleo y g el parámetro ventana. En este trabajo se propondrá la selección del parámetro ventana mediante el criterio de validación cruzada con error cuadrático relativo.

En particular, tomando $w_{ij}(u) = I_{Tol(u)}(\|\mathbf{s}_i-\mathbf{s}_j\|)$ en (3), donde $I_A(.)$ es la función indicadora del conjunto A y $Tol(u)$ representa la región de tolerancia en torno a u, se obtiene que $\hat{\gamma}(u)$ coincide con la versión isotrópica del estimador empírico del variograma.

Ambos estimadores presentan buenas propiedades para procesos estacionarios, como la insesgadez asintótica o la consistencia, que no suelen mantenerse en situaciones donde la tendencia del proceso no es constante (Ver, p.e. Cressie, 1993, Sección 3.4.3). Esto pone de manifiesto la necesidad de disponer de mecanismos que permitan permitan analizar el comportamiento de dichos estimadores en contextos más generales, para lo cual se puede recurrir a la metodología bootstrap para datos dependientes.

2. MÉTODOS BOOTSTRAP PARA DATOS DEPENDIENTES

Entre las técnicas de remuestreo para datos dependientes, cabe destacar el método bootstrap por bloques (MBB), propuesto por Davison y Hall (1993) y desarrollado en Lahiri (2003). Este método se basa en dividir la región de observación D en k bloques disjuntos de tamaño fijo b, para posteriormente construir una réplica bootstrap D^*, que se obtendría al encadenar los bloques generados mediante remuestreo entre todos los posibles bloques solapados, de similar tamaño, que se pueden construir en dicha región. El principal inconveniente de esta técnica es su dependencia de la forma de la región D, por lo que su implementación puede ser muy complicada si presenta irregularidades o si su partición en bloques no es exacta.

Como alternativa podría recurrirse al método bootstrap semiparamétrico (SPB). Esta técnica está inicialmente diseñada para su aplicación a procesos estacionarios, de modo que, partiendo de una muestra $\mathbf{y} = (Y(\mathbf{s}_1),\ldots,Y(\mathbf{s}_n))^t$, se aproximaría de forma paramétrica su matriz de varianzas covarianzas del proceso $\hat{\mathbf{\Sigma}}$ y se obtendría su factorización de Cholesky, es decir, $\hat{\mathbf{\Sigma}} = \mathbf{L}\mathbf{L}^t$. Seguidamente se podrían generar errores "incorrelados" $\mathbf{e} = \mathbf{L}^{-1}\mathbf{y}$, a partir de los cuales se construyen las muestras bootstrap independientes $\mathbf{e}^*$, y también réplicas de los datos originales, construídas como $\mathbf{y}^* = \mathbf{L}\mathbf{e}^*$. En los trabajos de Iranpanah et al (2011) y Olea y Pardo-Igúzquiza (2011) se trabaja esta metodología utilizando el estimador empírico del variograma para realizar inferencias sobre el proceso espacial. Esta técnica no depende de la geometría de D, a diferencia del MBB, si bien el SPB está expuesto a problemas de mala especificación del modelo elegido para la caracterización de la dependencia del proceso, así como al sesgo y la varianza que puede presentar el variograma para proceso con tendencia no constante.

Teniendo en cuenta lo anterior, en este trabajo proponemos la utilización de un método bootstrap no paramétrico (NPB), para su aplicación a un proceso espacial $Y(\mathbf{s})$ que responde al modelo (1) con tendencia determinística $\mu(\mathbf{s})$. Se comenzaría por seleccionar un estimador de la tendencia $\hat{\mu}(\mathbf{s})$, a partir del cual se aproximarían los residuos por:

$$\hat{\varepsilon}(\mathbf{s}) = Y(\mathbf{s}) - \hat{\mu}(\mathbf{s})$$

Para lo anterior, podríamos recurrir al estimador lineal local de la tendencia propuesto por Fan y Gijbels (1996), definido como sigue:

$$\hat{\mu}_{\mathbf{H}}(\mathbf{s}) = \mathbf{e}_1^t\left(\mathbf{X}_\mathbf{s}^t\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s}\right)^{-1}\mathbf{X}_\mathbf{s}^t\mathbf{W}_\mathbf{s}\mathbf{X}_\mathbf{s} = S_\mathbf{s}^t\mathbf{Y} \tag{4}$$

donde $\mathbf{e}_1^t = (1,0,\ldots,0)$, $\mathbf{X}_\mathbf{s}$ es una matriz cuya $i-$ésima fila es igual a $(1,(\mathbf{s}_i-\mathbf{s})^t)$, y $\mathbf{W}_\mathbf{s} = diag\{K_\mathbf{H}(\mathbf{s}_1-\mathbf{s}),\ldots,K_\mathbf{H}(\mathbf{s}_n-\mathbf{s})\}$, siendo $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1}K(\mathbf{H}^{-1}\mathbf{u})$, K una función de densidad de tipo núcleo multivariante y $\mathbf{H}$ una matriz ventana simétrica no singular de orden $d\times d$. Para la selección de $\mathbf{H}$ utilizaremos el criterio de Validación Cruzada Generalizada corregida (GCV) propuesto en el trabajo de Francisco-Fernández y Opsomer (2005).

Una vez eliminada la tendencia estimada del proceso original, se procedería a estimar el variograma $\gamma(h)$ a partir de los residuos, teniendo en cuenta que :

$$Var(\hat{\varepsilon}) = \mathbf{\Sigma} + S\mathbf{\Sigma}S^t - \mathbf{\Sigma}S^t - S\mathbf{\Sigma} = \mathbf{\Sigma}_{\hat{\varepsilon}} \tag{5}$$

donde $\mathbf{\Sigma}$ representa la matriz de varianza covarianza de los errores ε del modelo (1) y S es la matriz de suavizado.

Para corregir el efecto de este sesgo en el método NPB, proponemos un procedimiento iterativo similar al desarrollado por Fernández Casal y Francisco-Fernández (2014), y que se encuentra implementado en la función *np.svariso.corr* disponible en el paquete *npsp* del software *R*.

De este modo se obtendrían estimaciones tanto de $\mathbf{\Sigma}$ como de $\mathbf{\Sigma}_{\hat{\varepsilon}}$, que podrían utilizarse para generar las réplicas con el método NPB de acuerdo con el algoritmo que se describe a continuación:

1. Aproximar la tendencia utilizando el estimador lineal local (4) en las localizaciones observadas y los residuos $\hat{\varepsilon}(\mathbf{s})$ correspondientes.

2. Obtener $\hat{\mathbf{\Sigma}}$ y $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$ mediante el proceso iterativo explicado anteriormente y encontrar sus respectivas factorizaciones de Cholesky, tales que: $\hat{\mathbf{\Sigma}} = \mathbf{L}\mathbf{L}^t$ y $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}} = \mathbf{L}_{\hat{\varepsilon}}\mathbf{L}_{\hat{\varepsilon}}^t$

3. Calcular las variables incorreladas $\mathbf{e} = (e_1, e_2, \ldots, e_n)^t$ siendo $\mathbf{e} = \mathbf{L}_{\hat{\varepsilon}}^{-1}\hat{\varepsilon}$.

4. Recentrar estas variables, y obtener a partir de ellas muestras bootstrap independientes de tamaño n, que se denotarán por $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)^t$.

5. Generar los errores boostrap $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)^t$ como $\hat{\varepsilon}^* = \mathbf{L}\mathbf{e}^*$

6. Construir las muestras bootstrap de datos dependientes mediante: $Y^*(\mathbf{s}_i) = \hat{\mu}_{\mathbf{H}}(\mathbf{s}_i) + \hat{\varepsilon}^*(\mathbf{s}_i)$, $i = 1, \ldots, n$.

Finalmente, se podrían utilizar las réplicas $Y^*(\mathbf{s}_i)$ derivadas de este modo para estimar nuevamente la tendencia y los residuos, así como calcular $\hat{\gamma}^*(h)$ utilizando el estimador lineal local definido por (3) con las mismas ventanas.

Cabe mencionar que este método no paramétrico mantiene las ventajas indicadas para el SPB, ya que tampoco depende de la forma de la región de observación D. Por otra parte, aunque el SPB puede ser adaptado a procesos no estacionarios, éste se encuentra afectado por el sesgo debido a la estimación basada en residuos, mientras que en este caso el NPB permite obtener réplicas bootstrap corrigiendo el efecto de este sesgo, siendo éste el aspecto más relevante del método propuesto.

Incluso, suponiendo que el proceso espacial $Y(\mathbf{s})$ es estacionario, el NPB propuesto se puede adaptar fácilmente a este contexto (pues no sería necesario estimar $\mu(\mathbf{s})$ ni $\mathbf{\Sigma}_{\hat{\varepsilon}}$), manteniendo la ventaja de que no está expuesto a problemas de mala especificación del modelo, pues se utilizan estimaciones no paramétricas tanto para la tendencia como la matriz de varianzas y covarianzas. Además, desde el punto de vista computacional, la implementación del NPB es automática una vez programado, dado que no requiere de la intervención del usuario para realizar ningún ajuste de estimación.

3. COMPARACIÓN DE ESTIMADORES DEL VARIOGRAMA MEDIANTE MÉTODOS BOOTSTRAP

Utilizando los métodos bootstrap anteriores se puede obtener una aproximación de la variabilidad del estimador $\hat{\gamma}(\mathbf{h})$. Para ello, definiríamos en primer lugar el estadístico $T_n(\mathbf{h}) = \hat{\gamma}(\mathbf{h}) - \gamma(\mathbf{h})$ verificando:

$$Var(T_n(\mathbf{h})) = Var(\hat{\gamma}(\mathbf{h}) - \gamma(\mathbf{h})) = Var(\hat{\gamma}(\mathbf{h})). \tag{6}$$

Posteriormente, se generarían B réplicas de la muestra original $\{Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\}$. Luego, para cada una de estas réplicas, se obtendría $\{\hat{\gamma}_1^*(\mathbf{h}), \ldots, \hat{\gamma}_B^*(\mathbf{h})\}$ y el análogo bootstrap de $T_n(\mathbf{h})$:

$$T_{n,i}^*(\mathbf{h}) = \hat{\gamma}_i^*(\mathbf{h}) - \hat{\gamma}(\mathbf{h}), i = 1, ..., B. \tag{7}$$

Esto permitiría aproximar la varianza (6) como sigue:

$$Var^*(\hat{\gamma}^*(\mathbf{h})) = \frac{1}{B}\sum_{i=1}^{B}\left(T^*_{n,i}(\mathbf{h}) - \bar{T}^*_n(\mathbf{h})\right)^2 \tag{8}$$

Donde $\bar{T}^*_n(\mathbf{h})$ es la media muestral de $T^*_{n,i}(\mathbf{h})$ y corresponde a su vez a la versión bootstrap del sesgo del estimador $\hat{\gamma}(\mathbf{h})$. Alternativamente, cuando el proceso tiene una tendencia determinística, se puede utilizar en la expresión (7) un estimador insesgado del variograma en lugar de $\hat{\gamma}(\mathbf{h})$.

Procediendo de este modo, se compararán mediante simulación los resultados obtenidos por los tres métodos bootstrap, MBB, SPB y NPB. Para ello se generarán datos de diversos procesos isotrópicos, estacionarios y no estacionarios (con tendencia determinística), a partir de los cuales se aproximarán los errores resultantes de cada método, considerando localizaciones regularmente espaciadas en una rejilla bidimensional y diferentes configuraciones de dependencia. Además, se analizará el efecto de la mala especificación tabto del modelo de variograma como de la tendencia del proceso.

REFERENCIAS

Cressie N (1993) Statistics for Spatial Data. Wiley, New York

Davison A, Hall P (1993) On Studentizing and blocking methods for implementing the bootstrap with dependent data. Austral J Statist 35(2):215-224

Fan J, Gijbles I (1996) Local Polynomial Modelling and Its application, Taylor & Francis

Fernández-Casal R, Francisco-Fernández M (2014) Nonparametric bias-corrected variogram estimation under non-constant trend. Stoch Environ Res Risk Assess 28(5):1247-1259

Fernández-Casal R, González-Manteiga W, M Febrero-Bande (2003) Space-time dependency modeling using general classes of flexible stationary variogram models. J Geophys Res 108:8779

Francisco-Fernández M, Opsomer JD (2005) Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. Can J Stat 33:539-558

Iranpanah N, Mohammadzadeh M, Taylor C (2011) A comparison of block and semiparametric bootstrap methods for variance estimation in spatial statistics. Computational Statistics & Data Analysis 55(1):578-587

Lahiri SN (2003) Resampling methods for dependent data. Springer Science & Business Media

Olea R, Pardo-Igúzquiza E (2011) Generalized bootstrap method for assessment of uncertainty in semivariogram inference. Mathematical Geosciences 43(2):203-228

Opsomer JD, Wang Y, Yang Y (2001) Nonparametric regression with correlated errors. Stat Sci 16:134-153

Shapiro A, Botha JD (1991) Variogram fitting with a general class of conditionally nonnegative definite functions. Comput Stat Data Anal 11(1):87-96

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

Bootstrap ponderado para la bondad de ajuste del error en modelos de regresión no paramétrica

Gustavo Ignacio Rivas Martínez[1], María Dolores Jiménez Gamero[2]

[1] Dpto. de Ciencias Básicas, Universidad Nacional de Asunción, Paraguay.
[2] Dpto. Estadística e Investigación Operativa, Universidad de Sevilla, España.

RESUMEN

Se considera un test de bondad de ajuste para la distribución del error en modelos de regresión no paramétrica con posible heteroscedasticidad. El estadístico de contraste es una norma L_2 de la diferencia entre la función característica empírica de los residuos y la estimación paramétrica de la función característica de los errores bajo la hipótesis nula. La distribución nula asintótica del estadístico puede ser aproximada mediante un bootstrap paramétrico. Aunque el bootstrap paramétrico es en general fácil de implementarlo, normalmente conlleva un coste computacional mayor a medida que se incrementan el tamaño muestral y el número de parámetros. En este trabajo se investiga la estimación de la distribución nula asintótica del estadístico a través de un bootstrap ponderado. Una simulación para tamaño de muestra finito se lleva a cabo con el fin de evaluar empíricamente el desempeño del procedimiento estudiado.

Palabras y frases clave: Función característica, Bondad de ajuste, Bootstrap ponderado, Regresión no paramétrica.

1. INTRODUCCIÓN

Consideremos el siguiente modelo general de regresión no paramétrica para variables independientes e idénticamente distribuidas (iid) observables $(X_1, Y_1), ..., (X_n, Y_n)$

$$Y_j = m(X_j) + \sigma(X_j)\epsilon, \quad j = 1, 2, ..., n, \tag{1}$$

donde $m(X_j) = E(Y_j|X_j = x)$ es la función de regresión, $\sigma^2(X_j) = Var(Y_j|X_j = x)$ es la varianza condicional y ϵ es el error aleatorio, no observable, con esperanza 0 y varianza unitaria por construcción. Ambas, $m(.)$ y $\sigma(.)$ se supondrán desconocidas.

Varios tests basados en la función característica empírica (fce) han sido desarrollados para diversas hipótesis en modelos de regresión, ver por ejemplo, Jiménez-Gamero et. al. (2005), Hušková y Meintanis (2009,2010), Pardo-Fernández et. al (2015a,b), entre otros. En este trabajo consideramos el problema de contrastar la hipótesis nula

$$H_0 : F \in \mathcal{F},$$

donde F es la función de distribución acumulada (fda) de los errores y $\mathcal{F}$ es una familia paramétrica,

$$\mathcal{F} = \{F(\epsilon; \theta), \quad \theta \in \Theta\}, \quad \Theta \subseteq \mathbb{R}^p.$$

Para contrastar H_0, Hušková y Meintanis (2010) propusieron el test

$$\Psi = \begin{cases} 1, & \text{si} \quad T_{n,w}(\hat{\theta}_n) \geq t_{n,\omega,\alpha}, \\ 0, & \text{en caso contrario,} \end{cases}$$

donde, $t_{n,\omega,\alpha}$ es el percentil $1-\alpha$ de la distribución nula de $T_{n,\omega}(\hat{\theta}_n)$,

$$T_{n,\omega}(\hat{\theta}_n) = n \int |c_n(t) - c(t,\hat{\theta}_n)|^2 \omega(t) dt, \tag{2}$$

$c_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(it\hat{\epsilon}_j) = \frac{1}{n}\sum_{j=1}^{n} \cos(t\hat{\epsilon}_j) + i\frac{1}{n}\sum_{j=1}^{n} \sin(t\hat{\epsilon}_j)$, con $i = \sqrt{-1}$, es la fce de los residuos,

$$\hat{\epsilon}_j = \{Y_j - \hat{m}(X_j)\}/\hat{\sigma}(X_j), \quad j = 1, 2, ..., n,$$

$\hat{m}(\cdot)$ y $\hat{\sigma}(\cdot)$ son estimadores consistentes de $m(.)$ y $\sigma(.)$, respectivamente, $c(t;\theta) = E\{\exp(it\epsilon)\} = R(t;\theta) + iI(t;\theta)$ es la función característica (fc) de $F(\epsilon;\theta)$, con $R(t;\theta)$ e $I(t;\theta)$ parte real e imaginaria de $c(t;\theta)$, respectivamente, $\omega(t)$ es una función no negativa tal que $\int \omega(t)dt < \infty$, $\hat{\theta}_n$ es un estimador consistente de θ tal que $\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \psi(\epsilon_j;\theta) + o_p(1)$, con $E\{\psi(\epsilon_j;\theta)\} = 0$ y $E\|\psi(\epsilon_j;\theta)\|^2 < \infty$, $\|.\|$ representa la norma Euclidea.

La distribución nula exacta y la asintótica de $T_{n,\omega}(\hat{\theta}_n)$ son desconocidas. Hušková y Meintanis (2010) proponen estimar la distribución nula asintótica mediante un bootstrap paramétrico (BP). Si bien el BP es, en general, de fácil implementación, el coste computacional de su aplicación se incrementa notablemente a medida que el tamaño muestral y/o el número de parámetros aumenta. Por esta razón, el objetivo de este trabajo es investigar otras aproximaciones consistentes a la distribución nula de $T_{n,\omega}(\hat{\theta}_n)$ que sean, desde un punto de vista computacional, más económicas. Para el caso de variables aleatorias (observables) iid, Jiménez-Gamero y Kim (2015) han estudiado la aproximación que proporciona el bootstrap ponderado (WB), obteniendo buenos resultados tanto teóricos como empíricos, con importantes ganacias computacionales con respecto al PB. Por este motivo, el objetivo de este trabajo es estudiar teórica y empíricamente el uso del WB para aproximar la distribución nula asintótica del estadístico $T_{n,\omega}(\hat{\theta}_n)$.

2. RESULTADOS TEÓRICOS

Cuando H_0 es verdadera, en la demostración del Teorema 1 en Hušková y Meintanis (2010), bajo ciertas condiciones usuales no muy restrictivas, se establece que

$$T_{n,\omega}(\hat{\theta}_n) = T_{1,n,\omega}(\theta) + o_p(1), \tag{3}$$

con

$$T_{1,n,\omega}(\theta) = \int \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_1(\epsilon_j, t;\theta) \right\}^2 \omega(t) dt,$$

$$Z_1(\epsilon_j;t;\theta) = \Big\{ \cos(t\epsilon_j) - R(t;\theta) + \sin(t\epsilon_j) - I(t;\theta) - t\epsilon_j R(t;\theta) + t\epsilon_j I(t;\theta)$$

$$-t\tfrac{\epsilon_j^2-1}{2}\{R'(t;\theta) + I'(t;\theta)\} - \psi^T(\epsilon_j;\theta)\{\nabla R(t;\theta) + \nabla I(t;\theta)\} \Big\}.$$

El estadístico $\frac{1}{n}T_{1,n,\omega}(\theta)$ se puede representar como un V-estadístico degenerado de grado 2, con lo cual, de los resultados en Delhing y Mikosch (1994) (ver también Hušková y Janssen, 1993), se tiene que si $\xi_1, ..., \xi_n$ son iid con $E(\xi) = 0$ y $E(\xi^2) = 1$ independientes de $\epsilon_1, ..., \epsilon_n$, si se cumple (3), la distribución condicional, dados $(X_1, Y_1), ...(X_n, Y_n)$, de

$$T^*_{1,n,\omega}(\theta) = \int \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} Z_1(\epsilon_j, t;\theta)\xi_j \right\}^2 \omega(t) dt,$$

proporciona un estimador consistente de la distribución nula asintótica de $T_{n,\omega}(\hat{\theta}_n)$.

Sin embargo, como $T^*_{1,n,\omega}(\theta)$ depende del verdadero valor del parámetro θ (desconocido), de los errores no observables y de $\psi(\epsilon_j;\theta)$, que normalmente es desconocida, en la práctica la distribución condicional, dados $(X_1,Y_1),...(X_n,Y_n)$, de $T^*_{1,n,\omega}(\theta)$ no puede ser utilizada para aproximar la distribución nula de $T_{n,\omega}(\hat{\theta}_n)$. Para superar esta situación, se reemplaza ϵ_j por $\hat{\epsilon}_j$, θ por su estimador consistente $\hat{\theta}_n$ y $\psi(\epsilon_j;\theta)$ por $\psi_n(\hat{\epsilon}_j;\hat{\theta}_n)$, satisfaciendo

$$\frac{1}{n}\sum_{j=1}^{n}\|\psi_n(\hat{\epsilon}_j;\hat{\theta}_n)-\psi(\epsilon_j;\theta)\|^2 \xrightarrow{P} 0. \tag{4}$$

Así, para estimar la distribución nula asintótica de $T_{n,\omega}(\hat{\theta}_n)$, se considera el siguiente estadístico

$$T^*_{2,n,\omega}(\hat{\theta}_n)=\int\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}Z_2(\hat{\epsilon}_j;t;\hat{\theta}_n)\xi_j\right\}^2\omega(t)dt,$$

donde

$$\begin{aligned}Z_2(\hat{\epsilon}_j;t;\hat{\theta}_n)= \quad & \Big\{\cos(t\hat{\epsilon}_j)-R(t;\hat{\theta}_n)+\sin(t\hat{\epsilon}_j)-I(t;\hat{\theta}_n)-t\hat{\epsilon}_jR(t;\hat{\theta}_n)\\ & +t\hat{\epsilon}_jI(t;\hat{\theta}_n)-t\frac{\hat{\epsilon}_j^2-1}{2}\{R'(t;\hat{\theta}_n)+I'(t;\hat{\theta}_n)\}\\ & -\psi_n^T(\hat{\epsilon}_j;\hat{\theta}_n)\{\nabla R(t;\hat{\theta}_n)+\nabla I(t;\hat{\theta}_n)\}\Big\}.\end{aligned}$$

Se puede ver que las distribuciones condicionadas de $T^*_{2,n,\omega}(\hat{\theta})$ y $T^*_{1,n,\omega}(\hat{\theta})$, dados $(X_1,Y_1),...(X_n,Y_n)$, son asintóticamente equivalentes.
Sea

$$\Psi_*=\begin{cases}1, & \text{si} \quad T_{n,w}(\hat{\theta}_n)\geq t^*_{2,n,\omega,\alpha},\\ 0, & \text{en caso contrario,}\end{cases}$$

donde $t^*_{2,n,\omega,\alpha}$ es el percentil $1-\alpha$ de la distribución condicional de $T^*_{2,n,\omega}(\hat{\theta}_n)$, dados $(X_1,Y_1),...(X_n,Y_n)$. Cuando H_0 es verdadera, Ψ_* es asintóticamente correcto, en el sentido que el error de tipo I es asintóticamente igual al error nominal α. Cuando H_0 no es verdadera, Ψ_* es consistente, ya que es capaz de detectar asintóticamente cualquier alternativa (fija), para elecciones adecuadas de $\omega(t)$. Los resultados anteriores se cumplen tanto para los multiplicadores en bruto ξ como para los centrados $\xi-\overline{\xi}$, tal como lo sugieren en Kojadinovic y Yan (2012) (ver también Jiménez-Gamero y Kim, 2015).

3. IMPLEMENTACIÓN PRÁCTICA

Dada una muestra $(X_1,Y_1),...,(X_n,Y_n)$ iid, de (1) para calcular el p-valor WB, se siguen los siguientes pasos:

1. Calcular $\hat{\epsilon}_1,...,\hat{\epsilon}_j$.
2. Calcular $\hat{\theta}_n$ y el valor observado del estadístico $T_{n,\omega}(\hat{\theta}_n)_{obs}$.
3. Cálcular $m_{jk}=\int Z_2(\hat{\epsilon}_j;t;\hat{\theta}_n)Z_2(\hat{\epsilon}_k;t;\hat{\theta}_n)\omega(t)dt, \quad 1\leq j\leq k\leq n.$
4. Para un entero suficientemente grande B, repetir los pasos 4(a)-4(b), para todo $b\in\{1,2,...,B\}$.

(a) Generar una muestra de n variables aleatorias $\xi_1,...,\xi_n$ iid con esperanza 0 y varianza 1.

(b) Calcular $T_1^{*b}=\frac{1}{n}\sum_{j,k}^{n}\xi_j\xi_km_{jk}$ ó $T_2^{*b}=\frac{1}{n}\sum_{j,k}^{n}(\xi_j-\overline{\xi})(\xi_k-\overline{\xi})m_{jk}$.

5. Estimar el p-valor bootstrap ponderado, mediante $\hat{p}_{wb1}=\frac{1}{B}\sum_{b=1}^{B}I\{T_1^{*b}>T_{n,\omega}(\hat{\theta}_n)_{obs}\}$ y $\hat{p}_{wb2}=\frac{1}{B}\sum_{b=1}^{B}I\{T_2^{*b}>T_{n,\omega}(\hat{\theta}_n)_{obs}\}$.

4. RESULTADOS EMPÍRICOS

Para investigar el comportamiento para tamaños de muestra finito de la aproximación que prporciona el WB se ha realizado un estudio de simulación. Los programas utilizados fueron realizados empleando el lenguaje de programación R (http://www.cran.r-proyect.org). Consideramos el constraste $H_0 : \epsilon \sim Normal(0,1)$. Para la estimación de los residuos se utilizaron los estimadores Nadaraya-Watson con el núcleo el de Epanechnikov, $K(u) = 0{,}75 \times (1-u^2)$. Siguiendo a Pardo-Fernández et. al. (2015a), para el parámetro de ventana (h) se utiliza la expresión $c \times n^{-0{,}375}$. La función de peso utilizada es $\omega(t) = e^{-\lambda t^2}$. Se realiza una simulación previa para la elección adecuada de c y λ, en el sentido de proporcionar un error tipo I cercano al nominal. Los resultados en las tablas de abajo son para $c = 1{,}2$ y $\lambda = 0{,}04$, aunque no son los únicos que proporcionan buenos resultados. El modelo que genera los datos es $Y_j = X_j + X_j^2 + (X_j + 0{,}5) \times \epsilon$, $j = 1, 2, ..., n$, donde X_j sigue una distribución uniforme $(0, 1)$. Las distribuciones del error son: Normal (N), Laplace (LP), Chi-cuadrado con 3 grados de libertad (χ^2_3), t-student con 5 grados de libertad (T_5). Se realizan 1000 repeticiones del experimento. Se presentan los resultados para el PB siguiendo los pasos descritos en Hušková y Meintanis (2010) y para el WB con multiplicadores brutos y centrados, denotados como PB, WB1 y WB2, respectivamente. En la Tabla 1 se muestran los resultados para n=50, 100 y 200 para niveles de significación $0,05$ y $0,10$.

	N=50						N=100						N=200					
	$\alpha = 0,05$			$\alpha = 0,10$			$\alpha = 0,05$			$\alpha = 0,10$			$\alpha = 0,05$			$\alpha = 0,10$		
	PB	WB1	WB2	PB	WB1	WB2	PB	WB1	WB2	PB	WB1	WB2	PB	WB1	WB2	PB	WB1	WB2
N	5.7	5.2	6.6	10.1	9.8	10.3	5.8	5.5	6.0	10.7	11.1	11.8	4.8	4.6	4.9	9.8	11.0	11.4
LP	66.2	95.1	95.5	77.4	98.1	99.6	79.2	98.4	98.6	84.3	99.2	99.3	89.1	100	100	95.4	100	100
χ^2_3	66.2	95.3	94.4	77.6	98.1	100	79.2	98.1	98	84.4	99.1	99.3	89.4	100	100	95.5	100	100
T_5	18.6	41.8	45.2	29.8	57.2	59.6	30	84.8	85.8	49.6	91.6	92.2	57.8	99.8	99.8	79.2	99.8	99.8

Tabla 1: Porcentaje de rechazo para Hipótesis nula de Normalidad para tamaño de muestra $n = 50, 100$ y 200, para 5 % y 10 % de nivel de significación.

En la Tabla 1 se puede apreciar que en los tamaños de muestra ensayados, los 3 métodos arrojan estimaciones del error tipo I muy similares y cercanas a los nominales 5 % y 10 %, especialmente para n=200. Para las h y λ escogidas, el WB1 y WB2 son más potentes en las alternativas ensayadas. En cuanto al consumo de tiempo, mostrados en la Tabla 2, para n=50 el consumo de tiempo es de casi 15 veces menos que el PB, para n=100 aproximadamente 31 veces menos y para n=200 de casi 55 veces menos. Por lo tanto, desde un punto de vista computacional, el WB es más eficiente que el PB.

n	PB	WB1	WB2
50	5.36	0.34	0.37
100	12.97	0.40	0.43
200	35.34	0.62	0.66

Tabla 2: Tiempo empleado en segundos para obtener un p-valor con 1000 repeticiones.

REFERENCIAS

Delhing, H., Mikosch, T. (1994) Random quadratic forms and the bootstrap for U-statistics. J. Multivariate Anal. 51, 392-413.

Hušková, M., Janssen, P. (1993) Consistency of the generalized bootstrap for degenerate U-statistics. Test. 19, 92-112.

Hušková, M., Meintanis, S.G. (2009) Goodness-of-fit tests for parametric regression models based on empirical characteristics functions. Kibernetika. 45, 960-971.

Hušková, M., Meintanis, S.G. (2010) Test for the error distribution in nonparametric possiby heteroscedatic regression models. Test. 19, 92-112.

Jiménez-Gamero, M.D., Kim, H-M. (2015) Fast goodness-of-fit test based on the characteristic function. Comput. Statist. Data Anal. 89, 172-191.

Jiménez-Gamero, M.D., Muñoz-García, J., Pino-Mejías, R. (2005) Testing goodness of fit for the distribution of errors in multivariate linear models. J. Multivariate Anal. 95, 301-322.

Kojadinovic, I., Yan, J., (2012) Goodness-of-fit testing based on a weighted bootstrap: A fast sample alternative to the parametric bootstrap. Can. J. Statist. 40, 480-500.

Pardo-Fernández, J.C., Jiménez-Gamero, M.D., El Ghouch, A., (2015a) A nonparametric ANOVA-type test for regression curves based on characteristic functions. Scand. J. Stat. 42, 197-213.

Pardo-Fernández, J.C., Jiménez-Gamero, M.D., El Ghouch, A., (2015b) Tests for the equality of conditional variance functions in nonparametric regression. Electron. J. Statist. 9, 1826-1851.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

TESTS FOR THE EQUALITY OF CONDITIONAL VARIANCE FUNCTIONS IN NONPARAMETRIC REGRESSION

Juan Carlos Pardo-Fernández[1], María Dolores Jiménez-Gamero[2] and Anouar El Ghouch[3]

[1]Departamento de Estatística e Investigación Operativa, Universidade de Vigo
[2]Departamento de Estadística e Investigación Operativa, Universidad de Sevilla
[3]ISBA, Université catholique de Louvain

ABSTRACT

In this piece of research we are interested in checking whether the conditional variances are equal in two or more location-scale regression models. Our procedure is fully nonparametric and is based on the comparison of the error distributions under the null hypothesis of equality of variances and without making use of this null hypothesis. We propose four test statistic based on empirical distribution functions (Kolmogorov-Smirnov and Cramér-von Mises type test statistics) and two test statistics based on empirical characteristic functions. The limiting distributions of these six test statistics are established under the null hypothesis and under local alternatives. We show how to approximate the critical values using either an estimated version of the asymptotic null distribution or a bootstrap procedure. Simulation studies are conducted to assess the finite sample performance of the proposed tests. We also apply our tests to data on household expenditures.

Keywords: empirical characteristic function; empirical distribution function; kernel smoothing; regression residuals.

1. INTRODUCTION

When comparing $k \geq 2$ populations it is interesting not only comparing the means, but also other characteristics like the variances. For example, in quality control, it is important to check the uniformity and the stability of the production process under different experimental and practical conditions. In biomedical research, detecting variation in gene expression levels is important for many reasons, for example, to identify experimental and environmental factors that affect a biological process. Equality of variances, when satisfied, can also be used to develop more powerful and simple ANOVA-type test statistics. Without controlling for the effect of covariates, there are a substantial number of tests available in the literature for the equality of (unconditional) variances from two or more populations. The standard procedures include the classical F-test and Levene's test (Levene, 1960) which is known to be more robust to the violation of normality; see Gastwirth *et al.* (2009) for a recent review and some interesting examples and applications. In this paper, we are interested in the comparison of conditional variances.

We assume that in each population, along with the variable of interest or response variable, Y, it is also observed another variable, X, the covariate, so that the mean and the variance of the response variable depend on the values of X. More specifically, let (X_j, Y_j), $1 \leq j \leq k$, be k independent random vectors satisfying general nonparametric regression models

$$Y_j = m_j(X_j) + \sigma_j(X_j)\varepsilon_j, \tag{1}$$

where $m_j(x) = E(Y_j \mid X_j = x)$ is the regression function, $\sigma_j^2(x) = Var(Y_j \mid X_j = x)$ is the conditional variance function and ε_j is the regression error, which is assumed to be independent of X_j. Note that, by construction, $E(\varepsilon_j) = 0$ and $Var(\varepsilon_j)$=1. The covariate X_j is continuous with density function f_j. Since the objective is to compare the variance functions, it is reasonable to assume

that the covariates have common support, say R. The regression functions, the variance functions, the distribution of the errors and the distribution of the covariates are completely unknown and no parametric models are assumed for them. Thus, our approach is completely nonparametric. In this conditional setting, the hypothesis of equality of variances is stated in terms of the conditional variance functions: $H_0 : \sigma_1^2(x) = \sigma_2^2(x) = \cdots = \sigma_k^2(x)$, for all $x \in R$. Or equivalently,

$$H_0 : \sigma_j(x)/\sigma_0(x) = 1, \quad \text{for } 1 \leq j \leq k,$$

where $\sigma_0^2(x)$ is the common variance that can be expressed as $\sigma_0^2(x) = \sum_{j=1}^k \pi_j(x)\sigma_j^2(x)$, for some positive functions $\pi_1, \ldots, \pi_k$ satisfying $\sum_{j=1}^k \pi_j(x) = 1$. The alternative hypothesis is

$$H_1 : \sigma_j(x)/\sigma_0(x) \neq 1, \quad \text{for some } j \in \{1 \ldots, k\}.$$

We will develop several test statistics and study their distribution under H_0 and under local alternatives converging to the null hypothesis at the rate $n^{-1/2}$, n being the sample size. Specifically, we consider the following local alternative hypothesis

$$H_{1,n} : \sigma_j(x)/\sigma_0(x) = 1 + n^{-1/2}\delta_j(x), \quad \text{for } 1 \leq j \leq k,$$

for some functions δ_j. To be more precise, in the previous expression we should have written $\sigma_{n,j}(x)/\sigma_{n,0}(x)$ instead of $\sigma_j(x)/\sigma_0(x)$, as this function depends on n. However, to short the notation we suppress this explicit dependence on n. Observe that as n increases $H_{1,n}$ becomes closer and closer to H_0. Also, when $\delta_j(x) = 0$, for $1 \leq j \leq k$, $H_{1,n}$ reduces to H_0.

Statistical literature concerning the problem of testing for common features in several regression models has mainly focused on testing for common regression curves or testing for common error distributions. The problem of testing for the equality of regression curves in nonparametric settings has been extensively treated; see for example Delgado (1993), Kulasekera (1995), Neumeyer and Dette (2003), Pardo-Fernández *et al.* (2007, 2015a), Srihera and Stute (2010). González-Manteiga and Crujeiras (2013) give a review on this topic. On the other hand, testing for the equality of error distributions has been addressed in Pardo-Fernández (2007). To the best of our knowledge, the comparison of conditional variance functions has not been studied before. Most papers dealing with testing on the conditional variance function focus on studying the hypothesis of homoscedasticity (see for example Liero, 2003, or Dette and Marchlewski, 2010, and the references therein), or more in general, if the conditional variance function follows some fixed parametric form (see for example Dette *et al.*, 2007, or Koul and Song, 2010, and the references therein).

In order to construct a test for testing H_0, several approaches are possible. Here we follow the ideas in Pardo-Fernández *et al.* (2007, 2015a) for testing the equality of the regression functions, which consist of comparing the distribution of the errors of the regression models. Specifically, let

$$\varepsilon_j = \frac{Y_j - m_j(X_j)}{\sigma_j(X_j)}, \tag{2}$$

be the regression error in population j, $1 \leq j \leq k$. Define

$$\varepsilon_{0j} = \frac{Y_j - m_j(X_j)}{\sigma_0(X_j)} = \varepsilon_j \frac{\sigma_j(X_j)}{\sigma_0(X_j)} \tag{3}$$

to be the error under the null hypothesis, $1 \leq j \leq k$. Let $F_{\varepsilon_j}(t) = P(\varepsilon_j \leq t)$ and $F_{\varepsilon_{0j}}(t) = P(\varepsilon_{0j} \leq t)$ be the cumulative distribution function (CDF) of ε_j and ε_{0j}, respectively. The following theorem shows that H_0 is true if and only if the distributions of ε_j and ε_{0j} coincide.

Theorem 1. *Assume that σ_j is a continuous function in R and $0 < E(\varepsilon_j^4) < \infty$, $1 \leq j \leq k$.*

(a) *H_0 is true if and only if the random variables ε_j and ε_{0j} have the same distribution for all $1 \leq j \leq k$.*

(b) *Let $p_1, \ldots, p_k$ be such that $p_j > 0$, $1 \leq j \leq k$, and $\sum_{j=1}^k p_k = 1$. Let $F_\varepsilon(t) = \sum_{j=1}^k p_j F_{\varepsilon_j}(t)$ and $F_{\varepsilon_0}(t) = \sum_{j=1}^k p_j F_{\varepsilon_{0j}}(t)$. Assume also that $E(\varepsilon_1^4) = \cdots = E(\varepsilon_k^4)$. Then H_0 is true if and only if $F_\varepsilon(t) = F_{\varepsilon_0}(t)$, for all t.*

The assertions in the previous result can be interpreted in terms of the CDF or in terms of any other function characterizing a probability law, such as the characteristic function (CF). In this paper we will consider both cases, that is, to test H_0 we will compare consistent estimators of the CDFs and CFs of the random variables ε_j and ε_{0j}, $1 \leq j \leq k$.

2. THE TEST STATISTICS

Let (X_j, Y_j), $1 \leq j \leq k$, be k independent random vectors satisfying general nonparametric regression models (1). For $1 \leq j \leq k$, let ε_j and ε_{0j} be as defined in (2) and (3), respectively. As justified in Theorem 1, to test for H_0 we will compare consistent estimators of the CDFs and CFs of the random variables ε_j and ε_{0j}, $1 \leq j \leq k$, and also consistent estimators of the CDFs F and F_0 and of their associated CFs. Since neither ε_j nor ε_{0j} are observable, the inference must be based on the estimated residuals. Next we construct them.

Let (X_{jl}, Y_{jl}), $1 \leq l \leq n_j$, be independent and identically distributed (iid) observations from (X_j, Y_j), $1 \leq j \leq k$ and let $n = \sum_{j=1}^{k} n_j$. Along the paper it will be assumed that $n_j/n \to p_j > 0$, $1 \leq j \leq k$. In order to estimate the errors, we first need to estimate the regression functions, $m_j(x) = E(Y_j|X_j = x)$, the variance functions, $\sigma_j^2(x) = E[\{Y_j - m_j(x)\}^2|X_j = x]$, and the common variance function under H_0, $\sigma_0^2(x)$. With this aim we use nonparametric estimators based on kernel smoothing techniques. We use the following estimators for the functions m_j, σ_j^2 and σ_0^2:

$$\hat{m}_j(x) = \sum_{l=1}^{n_j} w_{jl}(x) Y_{jl}, \quad \hat{\sigma}_j^2(x) = \sum_{l=1}^{n_j} w_{jl}(x) Y_{jl}^2 - \hat{m}_j^2(x), \quad \hat{\sigma}_0^2(x) = \sum_{j=1}^{k} \pi_j(x) \hat{\sigma}_j^2(x),$$

where the quantities w_{jl} are, either the local-linear weights or the Nadaraya-Watson weights (for details, see for example Fan and Gijbels, 1996). Based on these estimators, for each population j, $1 \leq j \leq k$, we construct two samples of residuals,

$$\hat{\varepsilon}_{jl} = \frac{Y_{jl} - \hat{m}_j(X_{jl})}{\hat{\sigma}_j(X_{jl})} \qquad \text{and} \qquad \hat{\varepsilon}_{0jl} = \frac{Y_{jl} - \hat{m}_j(X_{jl})}{\hat{\sigma}_0(X_{jl})},$$

$1 \leq l \leq n_j$. Then we can construct the corresponding empirical CDFs (ECDFs),

$$\hat{F}_{\varepsilon_j}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{jl} \leq t) \qquad \text{and} \qquad \hat{F}_{\varepsilon_{0j}}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{0jl} \leq t),$$

and empirical CFs (ECFs),

$$\hat{\varphi}_{\varepsilon_j}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{jl}) \qquad \text{and} \qquad \hat{\varphi}_{\varepsilon_{0j}}(t) = \frac{1}{n_j} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{0jl}),$$

respectively. These ECDFs are consistent kernel based nonparametric estimators of the population CDFs $F_{\varepsilon_j}(t)$ and $F_{\varepsilon_{0j}}(t)$, respectively. Analogously, the above ECFs are consistent kernel based nonparametric estimators of the population CFs $\varphi_{\varepsilon_j}(t) = E\{\exp(\mathrm{i}t\varepsilon_j)\}$ and $\varphi_{\varepsilon_{0j}}(t) = E\{\exp(\mathrm{i}t\varepsilon_{0j})\}$, respectively. We can also consider the following ECDFs

$$\hat{F}_{\varepsilon}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{jl} \leq t) \qquad \text{and} \qquad \hat{F}_{\varepsilon_0}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} I(\hat{\varepsilon}_{0jl} \leq t),$$

and ECFs,

$$\hat{\varphi}_{\varepsilon}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{jl}) \qquad \text{and} \qquad \hat{\varphi}_{\varepsilon_0}(t) = \frac{1}{n} \sum_{j=1}^{k} \sum_{l=1}^{n_j} \exp(\mathrm{i}t\hat{\varepsilon}_{0jl}),$$

which estimate $F_{\varepsilon}(t) = \sum_{j=1}^{k} p_j F_{\varepsilon_j}(t)$, $F_{\varepsilon_0}(t) = \sum_{j=1}^{k} p_j F_{\varepsilon_{0j}}(t)$, $\varphi_{\varepsilon}(t) = \sum_{j=1}^{k} p_j \varphi_{\varepsilon_j}(t)$ and $\varphi_{\varepsilon_0}(t) = \sum_{j=1}^{k} p_j \varphi_{\varepsilon_{0j}}(t)$, respectively.

To test for H_0, we will construct Kolmogorov-Smirnov type statistics and Cramér-von Mises type statistics to compare the ECDFs, and weighted L_2-distances to compare the ECFs. More precisely, the considered statistics are

$$T_{KS}^1 = \sum_{j=1}^{k} \sqrt{n_j} \sup_t |\hat{F}_{\varepsilon_j}(t) - \hat{F}_{\varepsilon_{0j}}(t)|, \quad T_{CM}^1 = \sum_{j=1}^{k} n_j \int \{\hat{F}_{\varepsilon_j}(t) - \hat{F}_{\varepsilon_{0j}}(t)\}^2 d\hat{F}_{\varepsilon_{0j}}(t)$$

$$T_{KS}^2 = \sqrt{n} \sup_t |\hat{F}_{\varepsilon}(t) - \hat{F}_{\varepsilon_0}(t)|, \quad T_{CM}^2 = n \int \{\hat{F}_{\varepsilon}(t) - \hat{F}_{\varepsilon_0}(t)\}^2 d\hat{F}_{\varepsilon_0}(t),$$

$$T_1 = \sum_{j=1}^{k} n_j \int \left|\hat{\varphi}_{\varepsilon_j}(t) - \hat{\varphi}_{\varepsilon_{0j}}(t)\right|^2 w(t)dt, \quad T_2 = n \int |\hat{\varphi}_{\varepsilon}(t) - \hat{\varphi}_{\varepsilon_0}(t)|^2 w(t)dt,$$

where w is a positive weight function that is needed to guarantee consistency. Note that in the case of T_1 and T_2, $|\cdot|$ represents the modulus of a complex number.

3. SUMMARY OF ASYMPTOTIC RESULTS AND CONCLUSIONS

Due to the lack of space, we will not explain here in detail all the results obtained in this piece of research. The complete work can be found in Pardo-Fernández *et al.* (2015b). Below is a brief description of the results that will be explained in the presentation:

- Under certain regularity conditions, we have done a detailed study of the asymptotic distribution of the test statistics under the null hypothesis H_0 and under the local alternative $H_{1,n}$. The tests are consistent against any fixed alternative and are able to detect contiguous alternatives converging to the null at a rate $n^{-1/2}$. The assumptions needed to derive these properties are weaker for the ECF-based tests (specifically, no requirement is imposed on the distributions of the errors).

- We have proposed and studied approximations of the asymptotic null distribution the test statistics based on the ECF in order to obtain critical values for the test.

- We have studied the practical behaviour of the proposed test statistics by means of simulations. The simulation study includes the analysis of the tests based on critical values obtained from the asymptotic null distribution of the test statistics and from a bootstrap approximation.

- We have applied the proposed tests to a set of econometric data concerning households expenditures.

REFERENCES

Delgado, M.A. (1993) Testing the equality of nonparametric regression curves. *Statistics and Probability Letters*, 17, 199–204.

Dette, H. and Marchlewski, M. (2010) A robust test for homoscedasticity in nonparametric regression. *Journal of Nonparametric Statistics*, 22, 723–736.

Dette, H., Neumeyer, N. and Van Keilegom, I. (2007) A new test for the parametric form of the variance function in non-parametric regression. *Journal of the Royal Statistical Society, Series B*, 69, 903–917.

Fan, J. and Gijbels. I. (1996) *Local Polynomial Modelling and Its Applications*. Chapman & Hall.

Gastwirth, J.L., Gel, Y. and Miao, W. (2009) The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, 24, 343–360.

González-Manteiga, W. and Crujeiras, R.M. (2013) An updated review of Goodness-of-Fit tests for regression models. *Test*, 22, 361–411.

Koul, H.L. and Song, W. (2010) Conditional variance model checking. *Journal of Statistical Planning and Inference*, 140, 1056–1072.

Kulasekera, K.B. (1995) Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association*, 90, 1085–1093.

Levene, H. (1960) Robust tests for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow and H.B. Mann, Eds.), 278–292. Stanford University Press, Stanford.

Liero, H. (2003) Testing homoscedasticity in nonparametric regression. *Journal of Nonparametric Statistics*, 15, 31–51.

Neumeyer, N. and Dette, H. (2003) Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31, 880–920.

Pardo-Fernández, J.C. (2007) Comparison of error distributions in nonparametric regression. *Statistics and Probability Letters*, 77, 350–356.

Pardo-Fernández, J.C., Jiménez-Gamero, M.D. and El Ghouch, A. (2015a) A nonparametric ANOVA-type test for regression curves based of characteristic functions. *Scandinavian Journal of Statistics*, 42, 197–213.

Pardo-Fernández, J.C., Jiménez-Gamero, M.D. and El Ghouch, A. (2015b) Tests for the equality of conditional variance functions in nonparametric regression. *Electronic Journal of Statistics*, 9, 1826-1851.

Pardo-Fernández, J.C., Van Keilegom, I. and González-Manteiga, W. (2007) Testing for the equality of k regression curves. *Statistica Sinica*, 17, 1115-1137.

Srihera, R. and Stute, W. (2010) Nonparametric comparison of regression functions. *Journal of Multivariate Analysis*, 101, 2039–2059.

Comunicacións orais

Sesión 3 de Estatística

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

TTS: PROPUESTA DE UNA NUEVA LIBRERÍA EN R PARA LA ESTIMACIÓN DE LA CURVA MAESTRA MEDIANTE SUPERPOSICIÓN TIEMPO-TEMPERATURA

Antonio Meneses[1], Salvador Naya[2] y Javier Tarrío-Saavedra[2]

[1]Universidad Nacional de Chimborazo. Riobamba, Ecuador

[2]Escola Politécnica Superior. Departamento de Matemáticas. Grupo MODES. Universidade da Coruña

RESUMEN

En este trabajo se presenta una nueva librería desarrollada por los autores, TTS, que estima la curva maestra mediante Superposición Tiempo-Temperatura. El objetivo que se persigue es dotar de una herramienta informática para un problema de fiabilidad, como es la predicción de las propiedades viscoelásticas de materiales (polímeros) a partir de datos obtenidos mediante el análisis mecánico dinámico (DMA). Este paquete incorpora, dos métodos clásicos paramétricos Arrhenius y Williams-Landel-Ferry, y un nuevo no paramétrico basado en el cálculo y desplazamiento de las funciones derivadas. Las curvas maestras son obtenidas para cada temperatura mediante los métodos mencionados y suavizadas mediante la aplicación de B-Splines con el fin de predecir sus propiedades fuera del rango experimental. Con el fin de presentar convenientemente los resultados, se han incorporado gráficos de curvas maestras y factores de traslaciones horizontales y verticales.

Palabras clave: Superposición tiempo-temperatura, Software R, WLF, Arrhenius, Derivadas, Análisis Mecánico Dinámico.

1. INTRODUCCIÓN

Dentro del Software disponible para el tratamiento de datos en Análisis Mecánico Dinámico (DMA) de materiales (polímeros), referente a la construcción de curvas maestras mediante Superposición Tiempo-Temperatura (TTS) existen en el mercado varias alternativas, siendo el R una de las que ofrecen mayor interés por la comunidad estadística, dado el carácter de *lingua franca* que ofrece este lenguaje.

Hasta el momento no existen paquetes para la construcción de curvas maestras mediante modelos de TTS con el software R. Este nuevo paquete TTS calcula la curva maestra por Superposición Tiempo-Temperatura, con respecto a una temperatura de referencia. En este trabajo se presentan tres métodos diferentes para determinar los factores de desplazamientos de las curvas (módulo vs frecuencia): El método nuevo basado en las Derivadas, el de Williams-Landel-Ferry y el de Arrhenius. Como herramientas estadísticas se usan par la estimación de la curva maestra suavizadores de B-Splines. Como salidas del programa se dan estimaciones y gráficas con desplazamientos horizontales y verticales, datos TTS, TTS ajustados con B-Splines e intervalos de confianza bootstrap.

La superposición tiempo/temperatura (TTS) es una de las técnicas de extrapolación más útiles, con una amplia gama de aplicaciones. La TTS es válida para polímeros termo-reológicos simples, donde el tiempo y la temperatura son generalmente equivalentes en su efecto sobre el módulo de almacenamiento (E') de polímeros. La superposición de tiempo-temperatura se implementa mediante factores de desplazamiento horizontales y verticales, y siempre con respecto a una temperatura de referencia.

La obtención de una curva maestra, definida a una temperatura dada, es una función entre una de las propiedades viscoelásticas de un polímero y la frecuencia-tiempo. La curva maestra se construye a partir

de las curvas de módulo elástico (E') experimentales obtenidas a diferentes temperaturas, estimándose E' para cada una de las zonas de comportamiento mecánico de un polímero. Permite estimar las propiedades del polímero fuera del rango experimental. Por tanto será una herramienta muy útil para predecir tiempos de vida de materiales que son analizados a diferentes temperaturas.

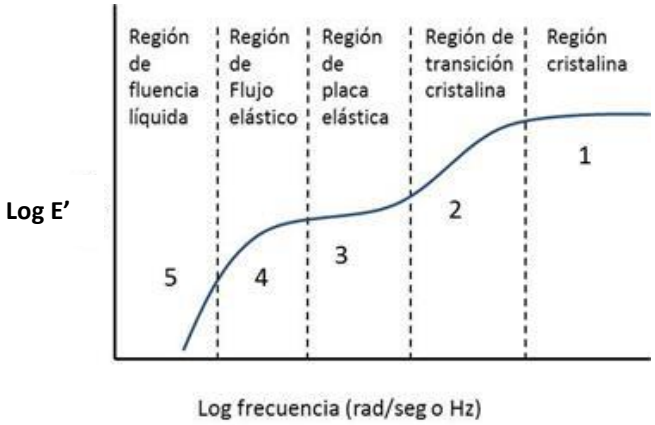


Figura 1: Curva maestra con zonas de comportamiento mecánico de un polímero.

En este trabajo se presentan las aplicaciones de la librería TTS para los datos de módulo de almacenamiento correspondientes al policarbonato (PC) y obtenidos mediante Análisis Mecánico Dinámico (DMA).

2. PROPUESTA DEL PAQUETE TTS Y SUS APLICACIONES

La librería tiene fundamentalmente dos funciones y una base de datos, en la que se ha aplicado los distintos métodos. La base de datos del policarbonato (PC), conteniendo los resultados del ensayo dinámico mecánico (DMA), con las variables temperatura, tiempo y módulo elástico o de almacenamiento. La función principal es la que permite la aplicación de los diferentes métodos, tiene el siguiente formato y argumentos:

TTS(x, reference.temperature=150, n=100, nB=100, method = c("derived", "WLF", "Arrhenius"))

El paquete TTS dispone de varias funciones para presentar los resultados gráficamente:

PLOT.TTS.data()	Función genérica que grafica los puntos experimentales, desplazados horizontalmente y verticalmente con respecto a una temperatura de referencia.
PLOT.TTS.gam()	Función genérica que grafica la estimación gam de la Curva Maestra e intervalos de confianza bootstrap al 95%.
PLOT.res()	Función genérica que gráficos residuales de la estimación gam de la Curva Maestra.

A continuación se muestran los gráficos característicos de un análisis TTS proporcionados por la función PLOT.TTS de la librería propuesta (Figuras 2-4). Ejemplo de obtención de la curva maestra para el E' del policarbonato (PC) a 150ºC

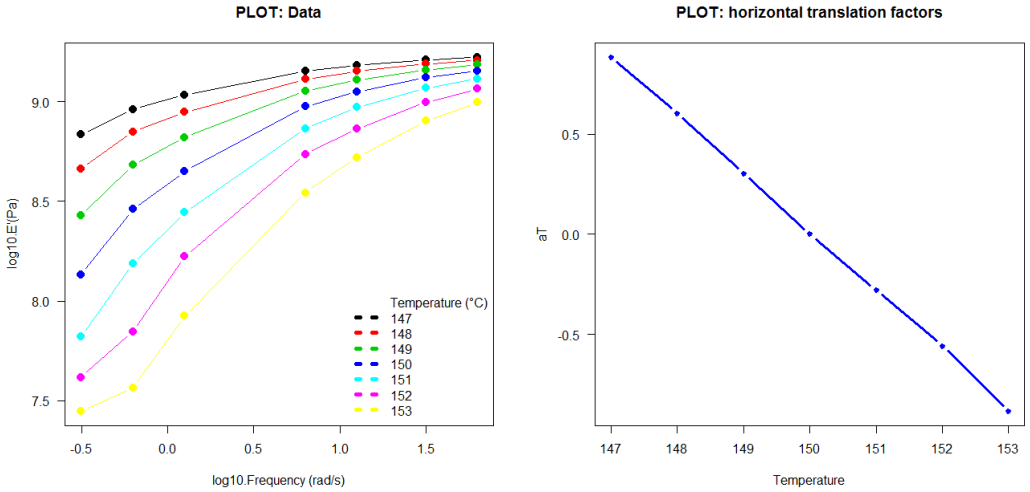


Figura 2: A la izquierda se muestran los datos experimentales. A la derecha se observan los desplazamientos horizontales correspondientes a una curva maestra del módulo del PC, estimada a 150ºC. El método que se aplica por defecto es el basado en el cálculo y desplazamiento de las primeras derivadas de las curvas.

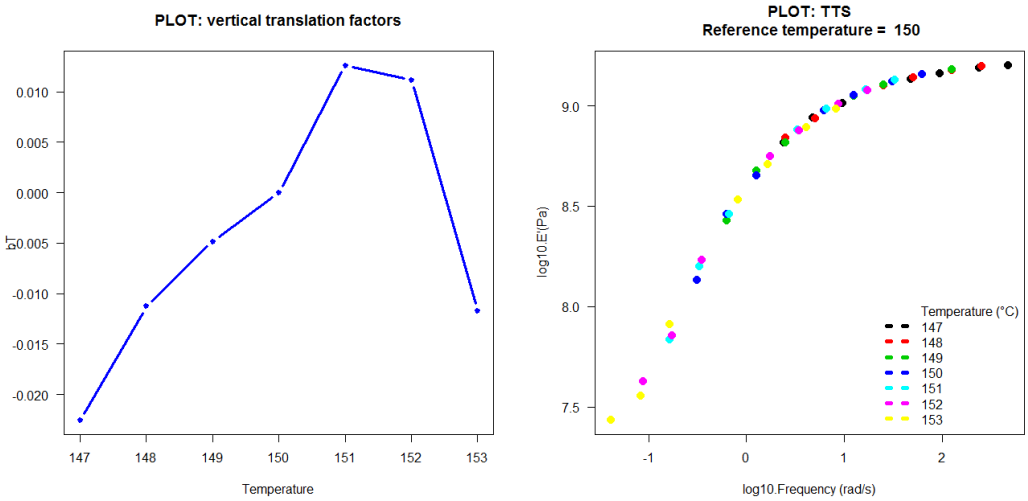


Figura 3: Desplazamientos verticales (izquierda) y curva maestra sin suavizar (derecha) obtenida por el método de desplazamiento de derivadas.

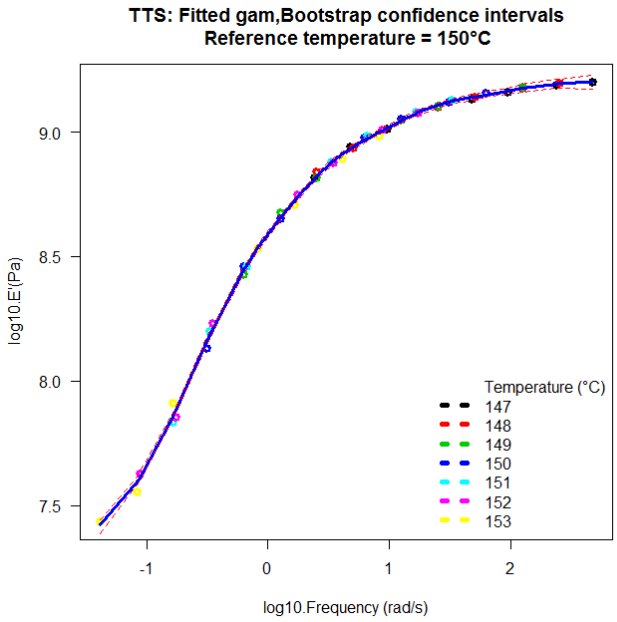


Figura 4: Curva maestra suavizada del módulo de almacenamiento del PC a 150ºC. Se obtienen estimaciones de E', con intervalos de confianza bootstrap, para un rango de frecuencias y tiempos diferente al experimental (compárese con la Figura 3 (derecha)).

4. CONCLUSIONES

En este trabajo, se presenta una nueva librería para estimar el valor de propiedades viscoelásticas de materiales fuera del rango experimental mediante modelos basados en el principio TTS. Se muestra una aplicación al estudio del E' del PC, obtenido mediante DMA. Una de las facilidades de este paquete es que permite seleccionar de entre tres modelos TTS el que mejor se ajuste a los datos, con la finalidad de obtener de manera precisa y confiable predicciones de las propiedades viscoelásticas. Además, la librería TTS permite realizar un análisis exploratorio e inferencial mediante gráficas de todos los componentes del proceso para la construcción y estimación de la curva maestra.

Agradecimientos

Los autores agradecen la ayuda de la SENESCYT del Gobierno de Ecuador, del proyecto MTM2013-41383P (fondos FEDER incluidos) y TA Instruments.

REFERENCIAS

Naya, S., Meneses A., Tarrío-Saavedra, J., Artiaga R., López-Beceiro, J. and Gracia-Fernández C. (2013) New method for estimating shift factors in time–temperatura superposition models. *Journal of Thermal Analysis and Calorimetry*. ISSN 1388-6150. DOI 10.1007/s10973-013-3193-1.

Williams, ML., Landel, RF. and Ferry, JD. (1955) The temperature dependence of relaxation mechanisms in amorphous polymers and other glass-forming liquids. *Journal of American Chemical Society*. 77:3701–7.

Ferry, JD. (1950) Mechanical proporties of substances of high molecular weight. VI. Dispersion in concentrated polymer solutions and its dependence on temperaures and concentration. *Journal of American Chemical Society*. 72:3746–52.

Artiaga, R., García, A. (2005) *Fundamentals of DMA*. In: Artiaga R editor. Thermal analysis. Fundamentals and applications to material characterization. A Coruña, Spain: Publicaciones de la Universidade da Coruña, pp. 183–06.

Wood, SN. (2006) *Generalized additive models: an introduction with R*. New York: Chapman and Hall/CRC Press.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

QCR: PROPUESTA DE UNA NUEVA LIBRERÍA EN R PARA EL CONTROL DE CALIDAD

Salvador Naya[1], Miguel Flores[2], Rubén Fernández Casal[3] y Javier Tarrío-Saavedra[1]

[1]Escola Politécnica Superior. Departamento de Matemáticas. Grupo MODES. Universidade da Coruña
[2]Facultad de Ciencias. Departamento de Matemáticas. Escuela Politécnica Nacional. Quito. Ecuador.
[3]Facultad de Informática. Departamento de Matemáticas. Grupo MODES. Universidade da Coruña.

RESUMEN

En este trabajo se presenta una nueva librería desarrollada por los autores, qcr, que implementa la mayoría de las herramientas estadísticas del control de calidad. El objetivo es ofrecer a la comunidad científica y usuarios en general del control de calidad una aplicación informática que permita un tratamiento sencillo y eficaz de las herramientas estadísticas del control de calidad, como son los gráficos de control por variables y atributos; el cálculo de índices de capacidad, las herramientas básicas bajo la metodología Seis Sigma al tiempo que incorpora también nuevas aplicaciones como los gráficos de control no paramétricos o los índices de capacidad de tercera y cuarta generación.

Palabras clave: Quality control; Software R; control chart; Seis, Sigma, data dept.

1. INTRODUCCIÓN

Dentro del Software disponible para el control estadístico de la calidad existen en el mercado varias alternativas, siendo el R una de las que ofrecen mayor interés por la comunidad estadística, dado el carácter de *lingua franca* que ofrece este lenguaje.

Hasta el momento ya existen diferentes paquetes para el control estadístico de calidad con R, entre los que destacan las librerías: qcc, IQCC; SixSigma y QualityTools. Estas librerías tienen ciertas restricciones para el tratamiento de los datos y no incorporan nuevas herramientas como gráficos de control estadístico e índices de capacidad más novedosos, como es el caso de los gráficos de control no paramétricos desarrollado por Regina Liu, que utilizan el concepto de profundidad de datos y los índices de capacidad no paramétricos propuestos Pearn y Chen, y en especial en los trabajos de Kerstin Vännman (véanse referencias).

Este nuevo paquete qcr permite elaborar y generar además de los gráficos clásicos tipo Shewhart nuevos procedimientos. Entre sus aplicaciones proporciona la funcionalidad básica para el análisis de control de calidad univariable y multivariable, incluyendo las siguientes herramientas: Xbar, Xbar-uno, S, R, n, np, c, g, ewna, cusum, mewna, mcusum y gráficos T2 en el contexto paramétrico; y para la parte no paramétrica incluye los gráficos de control de calidad: r, Q y S, así también los índices de capacidad de tercera y cuarta generación.

En este trabajo se presentarán algunas aplicaciones de la librería qcr. Comenzamos en la Sección 2, tratando de analizar las librerías existentes. En la sección 3 se discute la aplicación en el control de calidad del paquete qcr propuesto. En una última sección se exponen las conclusiones finales.

2. LIBRERÍAS DE R PARA EL CONTROL ESTADÍSTICO DE LA CALIDAD

En esta sección se presenta una visión general sobre el uso del R para el control de calidad, que es el motor de desarrollo de la aplicación computacional implementada en este trabajo y se menciona los

paquetes para el control de calidad que hasta la fecha (última actualización julio 2015), se encuentran alojados en la página web ttp://www.r-project.org/.

Como es conocido, R es un lenguaje y entorno de programación para análisis estadístico. Se trata de un proyecto de software libre cuyos primeros pasos se deben a Ross Ihaka y Robert Gentleman del Departamento de Estadística de la Universidad de Auckland, aunque las sucesivas versiones son controladas y desarrolladas por el R Development Core Team del que forman parte muchos colaboradores de todo el mundo. R se distribuye gratuitamente bajo los términos de la licencia pública general (GNU) y compila y se ejecuta en una diversidad de plataformas: UNIX, Windows y MacOS.

El lenguaje de programación incluye condicionales, lazos, y funciones. Con respecto a los argumentos de las funciones que están implementadas o las que el usuario crea pueden ser objetos de R, datos, fórmulas, expresiones, etc. En el entorno de R se utilizan paquetes que son un conjunto de funciones sobre una misma temática que proporcionan las facilidades para manejar conjuntos de funciones o datos.

En cuanto al control de calidad y herramientas de mejora. Los principales paquetes en relación con esos temas son los siguientes:

- *qcc*, para gráficos de control de calidad .
- *IQCC*, para la mejora de los gráficos de control de calidad
- *AcceptanceSampling*, para algunas técnicas de muestreo de aceptación
- *qualityTools*, un paquete para la enseñanza de los métodos estadísticos en el campo de Ciencias de la Calidad.

El paquete qcc ha sido desarrollado por el profesor Luca Scrucca del departamento de Economía, Finanzas y Estadística de la Universidad de Peruga. Permite las implementaciones de los gráficos Shewhart de control de calidad por variable y atributos; así también los gráficos Cusum y Ewma para variables continuas con dependencia; y para el caso de los Gráficos de control multivariable tiene implementado T2 Hotelling. Adicionalmente tiene implementado funciones para evaluar y calcular las características de funcionamiento de las curvas de operación OC, análisis de capacidad de proceso. Además se pueden realizar los diagramas de Pareto y de causa y efecto.

Por otro parte, el paquete IQCC es mantenido por el profesor Emanuel P. Barbosa del Instituto de Matemáticas de la Universidad Estatal de Campinas y tiene implementados un número menor de gráficos de control que el qcc aunque incorpora gráficos multivariantes .

Los paquetes MCUSUM (Multivariate Cumulative Sum Control Chart), MEWMA(Multivariate ExponentiallyWeighted Moving Average Control Chart) y MSQC (Multivariate Statistical Quality Control) han sido desarrollados por Edgar Santos Fernández de la empresa de Telecomunicaciones de Cuba.

El paquete MSQC, es un conjunto de herramientas para el proceso de control multivariante. Contiene las principales alternativas de gráficos de control multivariante tales como: Hotelling (T2), Chi cuadrado, MEWMA, MCUSUM y gráfico de control de la Varianza Generalizada. Además, incluye algunas herramientas para evaluar la normalidad multivariada.

En el caso en que no se pueda garantizar la normalidad de los datos, se podría utilizar gráficos de tipo no paramétrico que no requieren esta hipótesis, se tiene el paquete mnspc (Multivariate Nonparametric Statistical Process Control) desarrollado por Martín Bezener y Peihua Qiu de la Universidad de Minnesota. En este paquete se tiene una versión del procedimiento CUSUM.

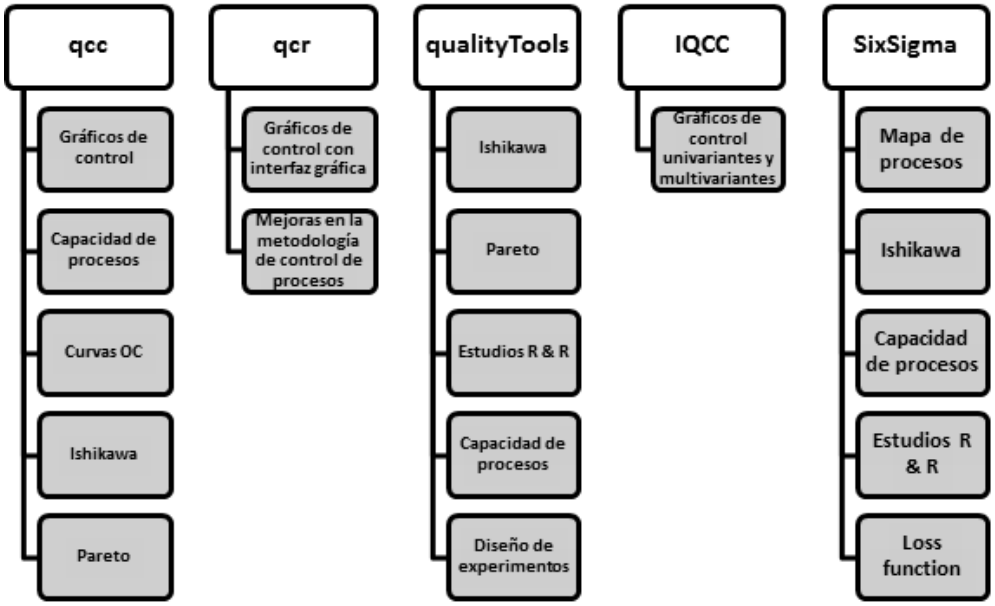


Figura 1: Funciones disponibles en los distintos paquetes de Control de Calidad con R.

3. PROPUESTA DEL PAQUETE QCR Y SUS APLICACIONES

En el paquete que se presenta, el qcr, además de incluir todas las alternativas de los paquetes qcc, qualityTools y Sixsigma, se implementan también gráficos no paramétricos que utiliza el concepto de profundidad como es el caso de los gráficos de control de calidad (r, Q, D y S) propuestos por Regina Liu (1990-).

Por tanto las principales razones de desarrollar el paquete qcr en este trabajo son:

- Tener en un sólo paquete integrado los gráficos de control de calidad más utilizados para el caso paramétrico y noparamétrico.
- Mejorar las salidas gráficas de los paquetes de control de calidad antes mencionados.
- Diseñar las funciones de tal forma que la interacción con el desarrollo de la interfaz gráfica sea más efícaz ya que la filosofía de desarrollo está en base al paquete Shiny.
- Facilitar la inclusión de nuevas funciones para la creación de gráficos de control ya que el diseño es modular, es decir, no depende de las anteriores implementaciones.
- No redundar en código ya que se utiliza las funciones implementadas en los paquetes qcc y MSQC.
- Permitir utilizar las funciones de otros paquetes en el caso de que ya estén implementadas para el cálculo de las medidas estadística y así sólo se preocupa de las salidas gráficas.
- Implementar gráficos de control de calidad multivariante noparamétricas que no se encuentran actualmente desarrolladas en ningún paquete.

Bajo estos objetivos se ha creado este nuevo paquete con las novedades siguientes:

1. Integrar todos los gráficos en un paquete.
2. Gráficos de control no paramétricos.
3. Gráficos de control para datos dependientes.
4. Funciones para facilitar la estimación en la fase I y el monitoreo en la Fase II.
5. Mejora en la presentación de los gráficos.

Desde la misma forma para índices de capacidad tanto paramétricos como no paramétricos, además de tipo univariantes y multivariantes.

1.Integrar todos los índices de capacidad en un mismo paquete.
2. Posibilidad de usar índices de capacidad no parametricos univariantes.
3. Posibilidad de usar índices de capacidad multivariantes de tipo parametrico y no paramétrico.

4. Posibilidad de usar índices gráficos para índices de capacidad como los propuestos por Vännman, (2001).

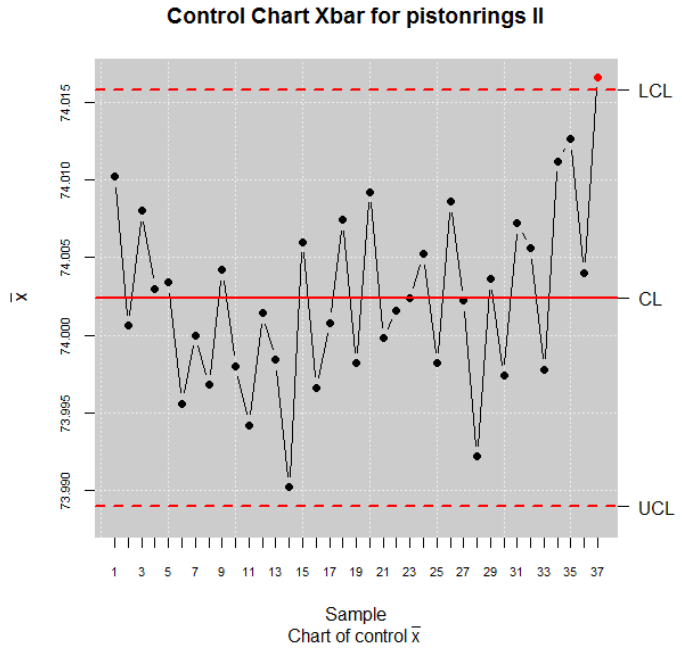


Figura 2: Gráfico de control para el diámetro medio de los pistones del paquete qcr.

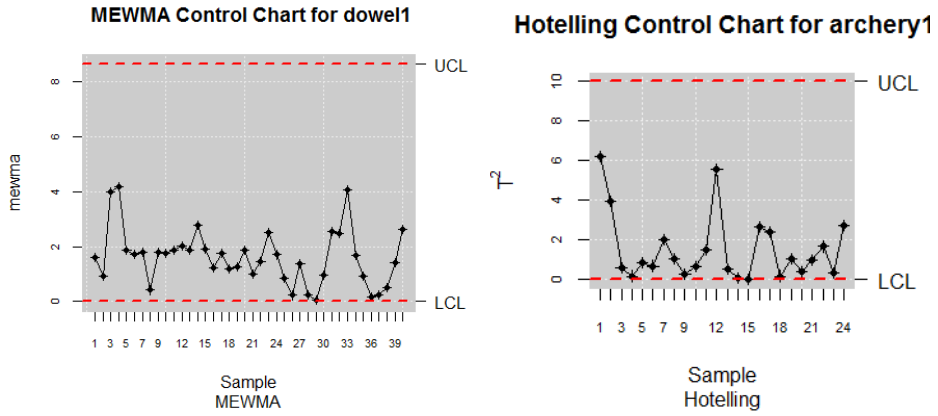


Figura 3: Gráfico izquierda (MEWMA) y derecha (Hotelling) para dos conjuntos de datos del paquete qcr.

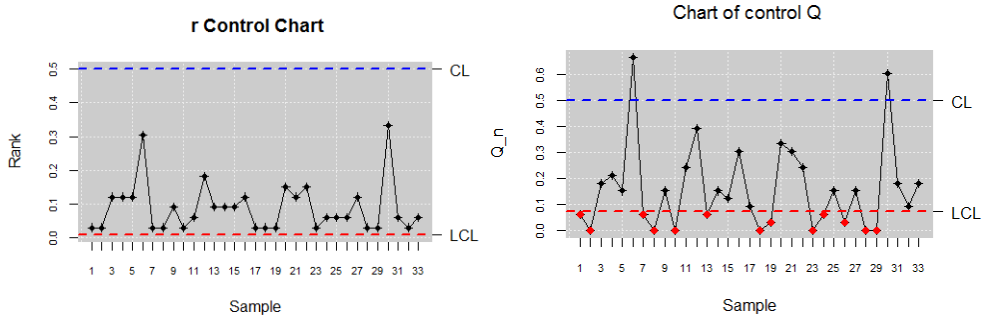


Figura 4: Gráficos no paramétricos tipo r y tipo Q del paquete qcr.

4. CONCLUSIONES

En este trabajo, se presenta una nueva librería para el control estadístico de la calidad. Entre las ventajas de este herramienta se encuentran: Tener en un sólo paquete integrados los gráficos de control de calidad más utilizados para el caso paramétrico y no paramétrico. Mejorar las salidas gráficas de los paquetes de control de calidad precedentes. Facilitar la inclusión de nuevas funciones para la creación de gráficos de control ya que el diseño es modular, es decir, no depende de las anteriores implementaciones.

Como novedad está la implementar gráficos de control de calidad multivariante no paramétricas que no se encuentran actualmente desarrolladas en ningún otro paquete de r y también el estudio de la capacidad de los procesos mediante índices robustos.

Agradecimientos

Los autores agradecen la ayuda de los Proyecto del ministerio MTM2011-22393 y MTM2013-41383P (fondos FEDER incluidos).

REFERENCIAS

Albing M. (2006). Process Capability Analysis with Focus on Indices for One-sided Specification Limits. Licentiate Thesis.

Cascos I, López A., and Romo J. Data Depth in Multivariate Statistics. Boletín de Estadística e Investigación Operativa. Vol. 27, No. 3, Octubre 2011, pp. 151-174.

Cheng, A., Liu, R., Luxho J. Monitoring multivariate aviation safety data by data depth: control charts and threshold systems, in IIE, Transactions (2000) 32, pp. 861-872.

Flores, M., Naya S., Tarrío-Saavedra, J. and Fernández-Casal R. (2014). qcr (quality, control and reliability. URL. http://www.R-project.org.

Liu, R. On a notion of simplicial depth. Proc. Natl. Acad. Sci. USA. Vol. 85, pp. 1732-1734, March 1988 Statistics.

Liu, R. On a notion of data depth based on random simplices. The Annals of Statistics. 1990, Vol. 18, No. 1, 405-414.

Liu, R. Control Charts for Multivariate Process. Journal of the American Statistical Asociation, Vol. 90. No. 432 (Dec., 1995), pp. 1380-1387.

Liu, R., Parelius J. and Singh K. Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference. The Annals of Statistics. 1999, Vol 27, No 3, 783-858.

Li, J. and Liu R. New Nonparametric Tests of Multivariate Locations and Scales Using Data Depth. Statistical Science 2004, Vol. 19, No 4,686-696.

Montgomery D. C. (2012). Statistical Quality Control, 7a ed., Wiley, Nueva York.

Tong, L.I. and Chen, J.P. (1998), Lower confidence limits of process capability indices for nonnormal process distributions. International Journal of Quality & Reliability Management, Vol. 15 No. 8/9, pp. 907-919.

Vännman, K. (1998). Families of Capability Indices for One-sided Specification Limits. Statistics, 31, 43-66.

Vännman, K. (2001). A Graphical Method to Control Process Capability. Frontiers in Statistical Quality Control, No 6, Editors: H-J Lenz and P-THWilrich. Physica-Verlag, Heidelberg, 290-311.

Vännman, K. and Albing, M. (2007). Process Capability Plots for One-Sided Specification Limits. Quality Technology & Quantitative Management Vol. 4, No. 4, pp.569-590.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

PROPUESTA DE UN MÉTODO PARA EL CONTROL DE CALIDAD DE LA EFICIENCIA ENERGÉTICA BASADO EN EL ESTUDIO DE DATOS FUNCIONALES

Manuel Oviedo de la Fuente[1], Javier Tarrío-Saavedra[2], Sonia Zaragoza[3], Manuel Febrero-Bande[1] y Salvador Naya[2]

[1]Facultade de Matemáticas. Departamento de Estatística e Investigación Operativa. Grupo MODESTYA. Universidade de Santiago de Compostela.

[2]Escola Politécnica Superior. Departamento de Matemáticas. Grupo MODES. Universidade da Coruña

[3]Escola Politécnica Superior. Departamento de Enxeñería Industrial II. Grupo PROTERM. Universidade da Coruña.

RESUMEN

Los datos funcionales (FDA) son un tipo de datos que pueden considerase dentro del tipo de Big Data pueden ser usado para describir datos generados a partir de una variedad de sensores, como los usado para medir el confort térmico. El empleo de técnicas de FDA supone una nueva oportunidad para los investigadores en el control estadístico de la calidad y de soluciones innovadoras en la industria. En este trabajo, se propone la aplicación de este tipo de herramientas para resolver problemas de eficiencia energética.

Palabras clave: Quality control; complex data; energy efficiency; data mining.

1. INTRODUCIÓN

La calidad del aire, confort térmico y la eficiencia energética son términos que miden la calidad de confort de los edificios. En este contexto suelen usarse sensores que permiten tomar medidas de calefacción, ventilación y del aire acondicionado (HVAC, por las siglas del inglés de estos términos). Los problemas de esta índole tienen el reto de controlar al mismo tiempo distintas variables críticas de calidad: temperatura, humedad, concentración de CO2).

Los datos producidos a partir de la medición del confort térmico de los edificios representan un interesante ejemplo de lo que podría llamarse datos funcionales. Estos datos son tomados con sensores térmicos que permiten tener datos cada 5 minutos de varias variables como la temperatura, CO2 o la humedad relativa (HR). Por tanto, los gráficos de control multivariantes como la T2 de Hotelling pueden ser una alternativa estadística para el control de calidad. También la disminución de variables usando componente principales (PA) o mínimos cuadrados parciales (PLS) son alternativa para usarse para resolver este problema. Pero sin duda, la propuesta de nuevos gráficos de control en el marco de análisis de datos funcionales, pueden y deben proponerse para controlar estas variables funcionales.

En este trabajo se propone una aplicación a grandes volúmenes de datos en el control estadístico de calidad puede ayudar a dar sentido a los datos complejos o grandes. Comenzamos en la Sección 2, tratando de responder a la pregunta sobre la relación entre datos funcionales y control de calidad. En la sección 3 se discute la aplicación en el control de calidad de la eficiencia energética. En una última sección se exponen las conclusiones finales.

2. DATOS FUNCIONALES Y CONTROL DE CALIDAD

En esta época es frecuente que se obtengan datos procedentes de sensores que en sí representan un problema para su depuración y tratamiento estadístico. Concretamente se tiene un exceso de datos y la dificultad para separar lo que podría llamarse “la señal del ruido”. Por tanto este tipo de datos de sensores térmicos pueden también considerarse como un ejemplo de Big Data, pues este tipo de datos se tienen

cuando son difícil de almacenar o analizar usando herramientas tradicionales de bases de datos o de Software estadístico. Aunque dado su carácter funcional los estudiaremos con herramientas de FDA.

Los datos funcionales se originaron dentro del campo de la quimiometría en la década de 1960. Desde entonces el tema ha recibido una gran atención por parte de la comunidad estadística desde entonces, que abarca una amplia gama de temas estadísticos importantes (Ferraty y Romain, 2011). Se remite al lector a Ramsay y Silverman (2002, 2005) para un estudio más profundo de los temas.

En estos conjuntos de datos de sensores, los métodos de control de calidad que se utilizan como los gráficos de control multivariable pueden llegar a ser muy costoso computacionalmente. Debido a esto varios autores han considerado métodos de selección de variables combinadas con gráficos de control para detectar rápidamente cambios en el proceso en una variedad de escenarios prácticos, incluyendo la detección de fallos, procesos multietapa. También se han propuesto métodos de aprendizaje máquni como el LASSO propuesto por Tibshirani (1996) combinado con un multivariante ponderado exponencialmente media móvil (EWMA). Por otra parte, Capizzi y Masarotto (2011) desarrollaron un gráfico de control multivariable (EWMA) utilizando el algoritmo de mínimo ángulo de regresión (LAR) para detectar cambios tanto en la media y como en la varianza para procesos de dimensional alta.

3. APLICACIÓN A DATOS DE EFICIENCIA ENERGÉTICA

Se aplicarán diferentes métodos de control estadístico de la calidad a datos reales obtenido por sensores situado en un edificio de oficinas. Concretamente se la temperatura (º C), HR (%), y la concentración de CO2 (ppm) en el aire interior se obtuvieron cada 5 minutos y cada hora. Además, también se midió el consumo de energía consumida por la calefacción, la ventilación y el aire acondicionado (HVAC). En total, hay 10 variables críticas a las instalaciones de climatización. La Figura 1 muestra la oficina estudiado y las ubicaciones de los sensores.



Figura 1: Plano de la oficina en la que se obtuvieron los datos y las ubicaciones de los sensores.

Se miden, 9 sensores en general. Cada variable se mide por un sensor. En total hay 3 sensores situados en la oficina de fondo, entrada y en el medio de la oficina. Además, se mide también el consumo de energía HVAC (kW).

Puesto que hay más de una variable para el control de la calidad, se propone la aplicación de gráficos de control multivariantes como T2 de Hotelling, MEWMA o MCUSUM (Montgomery, 2013). En el caso de conjuntos de datos complejos, donde cada individuo se define por un número relativamente elevado de variables (como el caso que nos ocupa), el rendimiento promedio de estos gráficos para detectar un cambio en la media de estas variables pierde importancia en el espacio p-dimensional de las variables de proceso (Montgomery, 2013). Por lo tanto, se propone la aplicación de gráficos de control multivariante T2 a variables latentes obtenidos por mínimos cuadrados parciales (PLS) para resolver este problema .Teniendo en cuenta que las variables latentes se ordenan en orden descendente de explica la variabilidad de los datos, podemos reducir el número de variables estudiadas de 10 a 2, sin perder información relevante.

En este caso, para medir la calidad de la eficiencia energética, lo importante es medir el consumo de energía y los parámetros del proceso en función de las temperaturas y el CO2. Esto hace que las técnicas de reducción tipo PLS serán adecuadas, véase Montgomery (2013) y Frank y Friedman (1993) para más detalles.

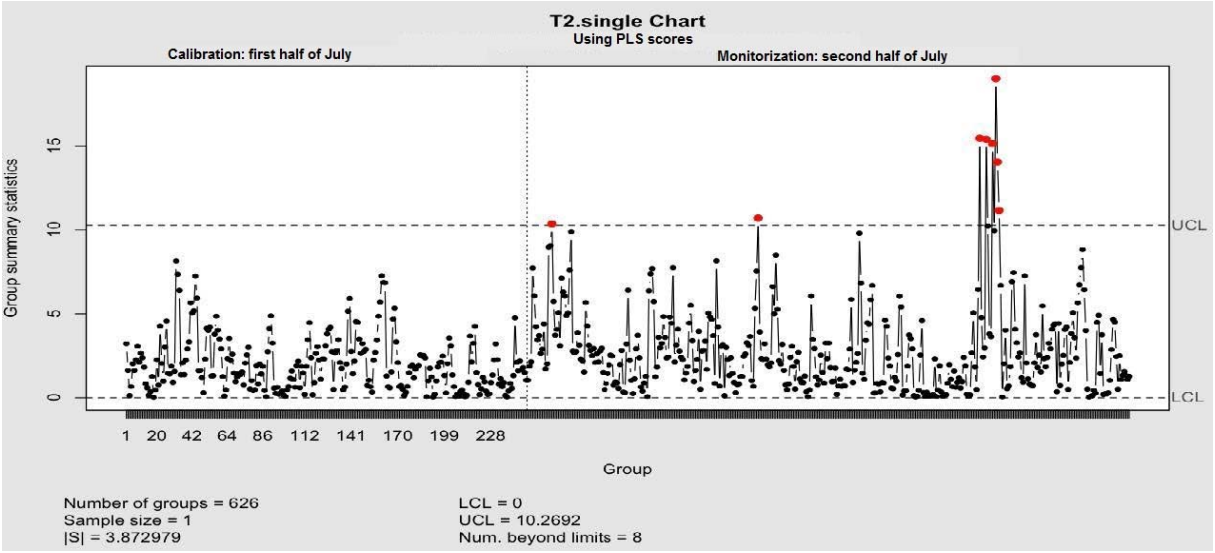


Figura 2: Gráfico de control T2 de Hotelling usando 2 las dos primeras componentes PLS de las 10 variables críticas de calidad HVAC.

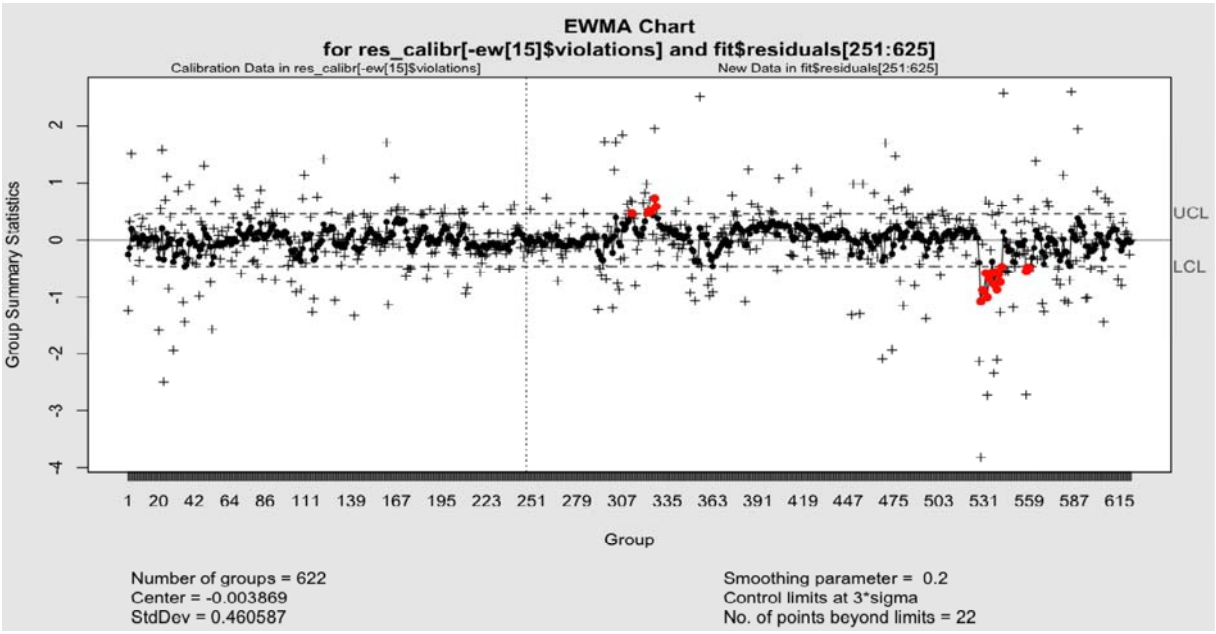


Figura 3: Gráfico EWMA correspondiente a los residuos del modelo ARIMA aplica a la temperatura en el medio de la oficina.

La propuesta de estos métodos clásicos de tipo multivariantes, como las propuestas de las figuras 2 y 3, pueden ser una solución pero presentan ciertos problemas en su aplicación y entendimiento por parte de los responsables de calidad.

Una alternativa para controlar las variables de eficiencia energética es la aplicación de técnicas de FDA. En este contexto cada dato será una curva diaria compuesta por el consumo en una hora (o tomadas cada 5 minutos) mediciones. La Figura 4 muestra las curvas de temperatura diarias suavizadas trazados con la mediana funcional y la mediana recortada. Podemos observar que la temperatura en el medio de la oficina es más constante durante el día que las otras temperaturas medidas en la parte posterior y la entrada.

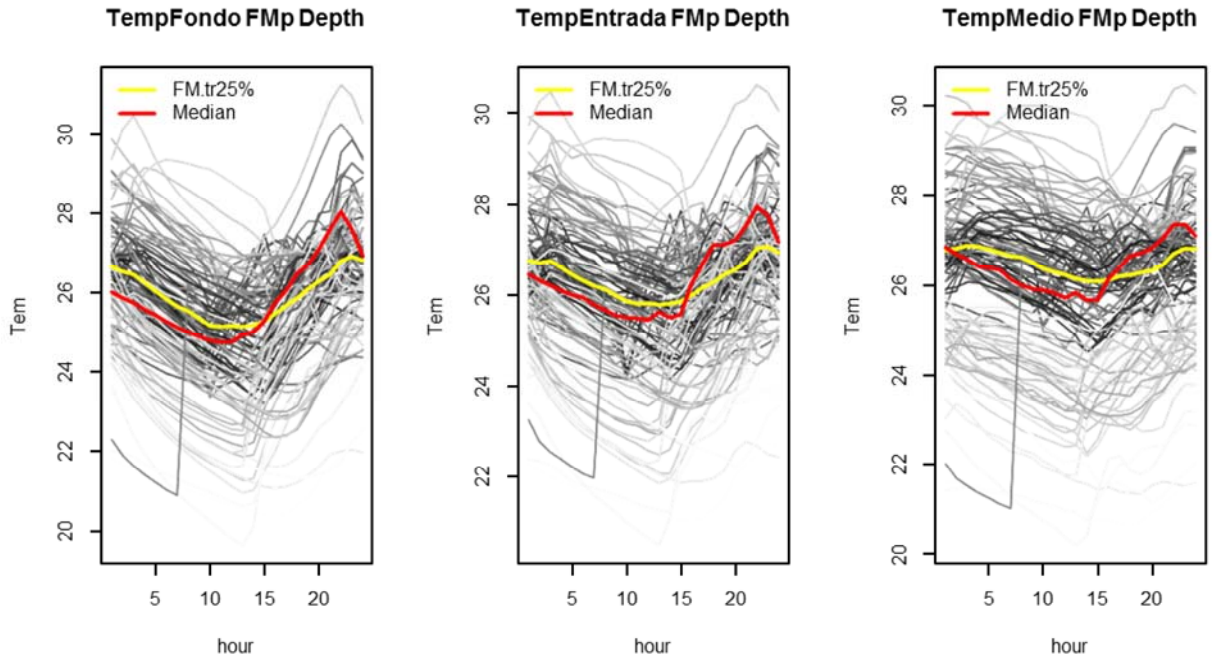


Figura 4: Gráfico izquierda (TempFondo), centro (TempEntrada) y derecha (TempMedio) de la oficina de las temperaturas diarias con sus medianas funcionales.

Además se pueden aplicar técnicas de clasificación FDA: como las DD-plot funcional, como las propuestas por Li et al. (2012) y Cuesta-Albertos et al. (2015). Los DD-plot se aplicaron usando la librería fda.usc (Febrero-Bande y Oviedo de la Fuente, 2012). El objetivo es estimar si el consumo diario de energía agregada (variable escalar) es alta o baja dependiendo de las 3 curvas de temperaturas diarias o dependiendo de las 3 curvas de HR diarias o las 3 curvas de contenido diario CO2.

Para ello se definen dos grupos: un grupo correspondiente a los días en que el consumo de energía fue mayor que la mediana y otro compuesto por los días en que el consumo de energía fue menor que la mediana del consumo. La profundidad de cada dato funcional (curva) se calcula utilizando la profundidad de modo con respecto a las dos distribuciones subyacentes (alto o bajo consumo). Se usó el clasificador logístico para encontrar la frontera que divide el plano en dos regiones: en rojo la región alto consumo y en azul la región bajo consumo. Los resultados se muestran en la Figura 5. Las curvas de temperatura o de CO2 diarias no diferencian entre el consumo de alta y baja energía. Sin embargo, el HR diaria si permite diferencias perfectamente entre los dos grupos. Cada punto en el plano representa el día. Si el punto se coloca en la zona roja, se puede inferir que el consumo de energía correspondiente a la HR medido en ese día fue mayor que la mediana. Los dos grupos pueden ser recalculados una definición correspondiente a los valores atípicos de consumo y otra correspondiente al consumo normal.

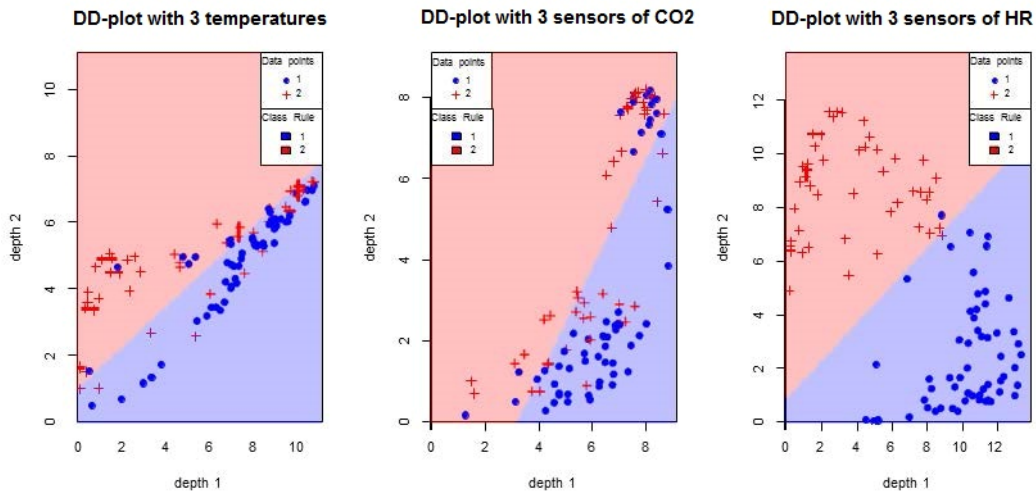


Figura 5: Gráficos DD-plot funcionales correspondiente a las profundidades a las temperaturas (izquierda), CO_2 (Centro) y HR (derecha).

4. CONCLUSIONES

En este trabajo, se propone la aplicación de técnicas de FDA para el control de calidad con grandes volúmenes de datos de eficiencia energética. Como el caso del empleo de los DD-plot. Se muestra como aplicar estas técnicas como una alternativa a la aplicación de gráficos de control estadístico multivariante para resolver los problemas de control y detección de alarmas en la eficiencia energética. Se aplicaron estas técnicas al control de la calidad del aire, el confort térmico y la eficiencia energética de las instalaciones de climatización de un edificio de oficinas.

Se verifica la fuerte relación entre dos variables (el consumo y la HR es funcional). Por otra parte, la relación entre la temperatura y el consumo no es tan fuerte.

Estos estudios abren un nuevo camino en la era de los grandes datos, para la automatización y control en HAVC instalación.

Agradecimientos

Los autores agradecen la ayuda de los Proyecto del ministerio MTM2011-22393 y MTM2013-41383P (fondos FEDER incluídos). También agradecen la ayuda prestada por los responsables de la empresa FRIDAMA que han proporcionado los datos del estudio.

REFERENCIAS

Capizzi G., Masarotto G. (2011). A Least Angle Regression Control Chart for Multidimensional Data. Technometrics, 53 (3):285-296.

Cuesta-Albertos JA, Febrero M, Oviedo de la Fuente M. The DDG-classifier in the functional setting. Submited to Test. 2015. Retrieved in http://arxiv.org/abs/1501.00372

Febrero-Bande M, Oviedo de la Fuente M. Statistical Computing in Functional Data Analysis: The R Package fda.usc. J Stat Softw. 2012; 51, 1–28.

Ferraty F, Romain Y. (2011). The Oxford handbook of functional data analysis. Oxford handbooks. Oxford University Press, Oxford ; New York.

Jones-Farmer LA, Ezell JD, Hazen BT. (2013). Applying control chart methods to enhance data quality. To appear in Technometrics.

Francisco-Fernández M, Tarrío-Saavedra J., Mallik A., Naya S. A comprehensive classification of wood from thermogravimetric curves. Chemom Intell Lab Syst. 2012; 118:159–72.

Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993; 35: 109–135.

Li J, Cuesta-Albertos JA, Liu RY. DD-Classifier: Nonparametric Classification Procedure Based on DD-plot. JASA. 2012; 107: 737-753.

Montgomery DC. (2013). Introduction to statistical quality control. 7th edn. Wiley, Hoboken, NJ.

Ramsay JO, Silverman BW. (2002). Applied functional data analysis: methods and case studies. Springer series in statistics. Springer, New York.

Ramsay JO, Silverman BW. (2005). Functional data analysis. Springer series in statistics, 2nd edn. Springer, New York.

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

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

CLASIFICACIÓN DE FONDOS MARINOS A PARTIR DE DATOS ACÚSTICOS

Javier Tarrío-Saavedra[1], Noela Sánchez-Carnero[2], Andrés Prieto Aneiros[3]

[1]Escola Politécnica Superior. Grupo MODES. Ferrol. Universidade da Coruña
[2]Grupo de Oceanografia Fisica. Universidade de Vigo
[3]Facultade de Informática. Grupo M2NICA. Universidade da Coruña.

RESUMEN

El objetivo de este estudio es la clasificación de fondos marinos en regiones costeras a partir de datos acústicos, en particular aquellos obtenidos mediante técnicas de sónar. Para ello se ha propuesto un enfoque alternativo de clasificación estadística supervisada y no supervisada aplicando técnicas de Análisis de Datos Funcionales (FDA). Se ha aplicado una versión FDA del algoritmo K-medias para identificar los tipos de fondos marinos en la ría de Cedeira. Además, se han aplicado métodos de clasificación supervisada FDA, en particular un modelo funcional lineal generalizado (GLM) y un modelo funcional aditivo generalizado (GSAM), para clasificar tres tipos distintos de suelo a partir de los datos obtenidos por sónar en el área de Cabo de Palos (Murcia). Los resultados indican que la introducción de técnicas FDA podría ser una alternativa al análisis tradicional multivariante para clasificar fondos marinos, sin tener que hacer una extracción selectiva de características relevantes de las curvas, y requiriendo por tanto un menor conocimiento previo de conceptos acústicos.

1. INTRODUCIÓN

Los datos acústicos, obtenidos mediante sónar, aparte de informar acerca de la batimetría y orografía del fondo, aportan información acerca de la textura del suelo (presencia o no de vegetación, tipo de vegetación, tipo de suelo: arenoso, limoso, pedregoso…). Es por esto último que han sido utilizados para identificar tipos de fondos marinos (Rodríguez-Pérez et al., 2014), tarea muy relacionada con el aprovechamiento de recursos (materias primas, flora, fauna) y el conocimiento del medio.

La clasificación de fondos marinos es una tarea compleja que en los últimos años se ha abordado desde el punto de vista de la estadística multivariante. Así, se suelen aplicar métodos de clasificación supervisada y no supervisada a una base de datos compuesta por las características relevantes de los pulsos acústicos obtenidos mediante sónar. La elección de un vector de características representativo de la textura del fondo no es en absoluto trivial. Hoy en día existen varias alternativas que combinan la elección de características con significación física a partir de los pulsos acústicos con métodos de reducción de dimensión como es el Análisis de Componentes Principales, PCA (Legendre et al, 2002; Preston y Kirlin, 2003). Sin embargo, la clasificación automática de fondos marinos es todavía un tema abierto. La aplicación de técnicas de clasificación de Análisis de Datos Funcionales, FDA, (Ramsay and Silverman, 2005) puede ser de gran utilidad en este campo al utilizar como variable regresora las propias curvas de pulsos acústicos, sin tener que partir de un conocimiento previo del problema físico para elegir el vector de variables independientes.

2. OBTENCIÓN DE DATOS

Se han obtenido medidas acústicas a través de 62 transectos realizados en la Ría de Cedeira (Figura 1), utilizando un single-beam echosounder (EA400P Simrad) que trabaja con un transdutor de 38/200

kHz, acoplado a un barco de 6.95 m de eslora. Las frecuencias de 200 kHz y 38 kHz han sido operadas con longitudes de pulso de 256 ms y 1024 ms. La velocidad del barco se ha mantenido entre 4 y 5 nudos (Rodríguez-Pérez et al., 2014). Con respecto a los datos tomados en Murcia, se ha empleado un transductor Simrad 38/200 Combi C, que combina dos transductores y un sensor de temperaturaen un único cuerpo.

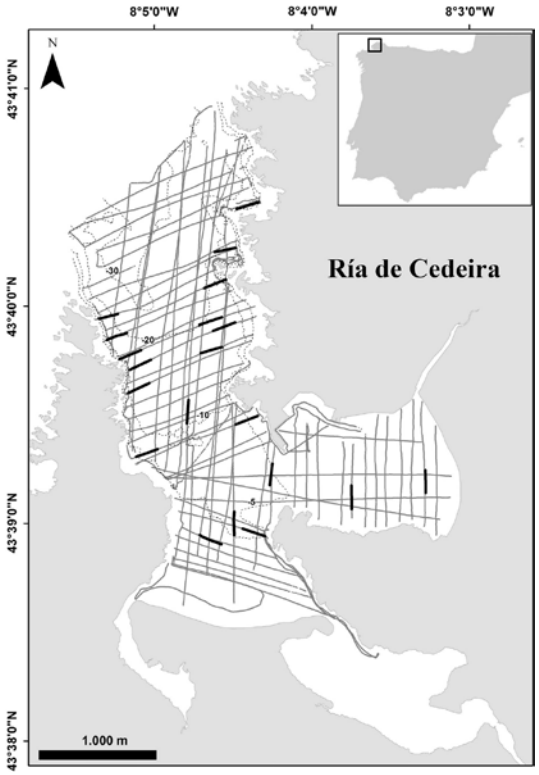


Figura 1. Transectos realizados en la Ría de Cedeira (Rodríguez-Pérez et al., 2014).

Una vez obtenidos los datos, han de hacerse realizarse una seria de transformaciones con el objeto de que el dato transformado aporte sólo información acerca de la textura del fondo. La corrección de profundidad se practica para que dos ecos que han sido obtenidos en un mismo tipo de fondo, con una misma inclinación, aunque a diferentes profundidades, sean prácticamente iguales. Se define una profundidad de referencia y, de acuerdo con el método de Pouliquen (Rodríguez-Pérez et al., 2014), los ecos son también escalados en tiempo y potencia.

3. TÉCNICAS DE CLASIFICACIÓN FDA

El Análisis de Datos Funcionales (FDA) es una nueva rama de la estadística que analiza la información contenida en curvas, super_cies u otra estructura variando sobre un continuo: tiempo, espacio, longitud de onda, etc. El FDA está relacionado con el análisis de datos en forma de funciones (Ferraty y Vieu, 2006). X se dice variable funcional si toma valores en un espacio funcional normado o seminormado, E, y una base de datos funcional $\{X_1,\ldots,X_n\}$ es la observación de n variables funcionales $X_1,\ldots,X_n$ idénticamente distribuidas como X (Tarrío-Saavedra et al., 2011).

Por otro lado, se define clasificación como un conjunto de técnicas de Análisis Multivariante y FDA diseñadas para agrupar o clasificar los elementos de una muestra, definidos por un conjunto de características (variables). Puede ser supervisada y no supervisada.

La clasificación no supervisada es la agrupación de individuos de los que no se conoce su clase (ni el número de grupos diferentes) según sus características. Existen diferentes métodos de clasificación no supervisada como el clúster jerárquico, K-medias, etc. En el caso funcional, existen diversas alternativas entre las que se encuentra una versión K-medias funcional (Febrero-Bande y Oviedo de la Fuente, 2012). Se elige el nº de grupos (K) y se establecen los centros (funciones) al azar. Se asignan los individuos a los

grupos y se vuelven a calcular sus centros (funciones). Se repiten los pasos anteriores hasta que se repite la clasificación, se llega a un máximo número de iteraciones o el desplazamiento de centros es menor que una determinada tolerancia.

La clasificación supervisada tiene por objeto asignar una clase a un individuo usando un modelo de clasificación previamente construido a partir de una muestra de entrenamiento (compuesta por individuos de los que se sabe su clase). Cada individuo está definido por un vector de características o mismo una función. Se define muestra de entrenamiento como un conjunto de individuos, caracterizados por un vector de variables de los que se conoce su clase. El modelo de clasificación o función discriminante se estima a partir de estos datos. Finalmente, la muestra de test es un grupo de individuos de los que no sabemos su clase. El objetivo de la clasificación supervisada es la determinación de las funciones discriminantes a partir de las variables originales, que permiten decidir en qué clase debe estar cada elemento, utilizando como criterio de asignación la proximidad de cada elemento a las distintas clases.

En un problema clásico de clasificación supervisada, se predice la clase *Y* a partir de un vector de características *X* (Wehrens, 2011). En la clasificación FDA, *X* es una función, con lo que el objetivo es predecir *Y* a partir de una varaible funcional *X*. En particular, se estima la probabilidad a posteriori de pertenencia a cada grupo (regla de Bayes):

$$p_g(X) = P(Y = g||\chi = X) = E\left(1_{Y=g}||\chi = X\right) \qquad \hat{Y} = \arg\max \hat{p}_g(X)$$

La clase de cada individuo se puede estimar utilizando clasificadores logísticos o no paramétricos mediante el paquete R fda.usc (Febrero-Bande y Oviedo De la Fuente, 2012): clasificadores k vecinos más próximos, kernel y logística a través de modelos aditivos generalizados y modelos aditivos generalizados.

En este trabajo se aplican los modelos de clasificación FDA lineal generalizados y aditivos generalizados. En el modelo lineal funcional generalizado (GLM) la respuesta escalar y se estima con variables funcionales $\{X_q(t)\}_{q=1}^{Q}$ y también no funcionales $\mathbb{Z} = \{Z_j\}_{j=1}^{J}$,

$$y_i = g^{-1}\left(\alpha + \mathbb{Z}_i\beta + \sum_{q=1}^{Q}\langle X_i^Q(t), \beta_q(t)\rangle\right) + \varepsilon_i$$

donde $g(\,)$ es el link inverso y ε_i los errores de media cero y varianza σ^2.

En estos modelos se requiere representar $X(t)$ y $\beta(t)$ en una base adecuada. Ramsay and Silverman (2005) propusieron su representación en función de una base B-spline, Fourier o Wavelets. Por otro lado, Cardot et al. (1999) introdujeron la regresión con Componentes Principales funcionales (FPC), mientras que Preda et al. (2007) propusieron la regresión con Mínimos Cuadrados Parciales funcionales (FPLS).

El modelo lineal aditivo funcional espectral generalizado, GSAM (Febrero-Bande y González-Manteiga , 2013), tiene la forma

$$\boldsymbol{y_i} = \boldsymbol{g^{-1}}\left(\alpha + \sum_{j=1}^{J} f_j\left(Z_j^J\right) + \sum_{q=1}^{Q} s_q\left(X_i^Q(t)\right)\right) + \boldsymbol{\varepsilon_i}$$

donde $f(\,)$ y $s(\,)$ son funciones suaves.

Como no se puede probar un modelo con la misma muestra que se usa para entrenarlo, para asegurar que un método de clasificación es válido para la población, no sólo para la muestra, es necesario aplicar modelos de validación como el método de validación cruzada. En el caso del presente trabajo, se ha empleado el procedimiento 10-fold cross-validation (la muestra es grande) para obtener una estimación de la proporción de clasificación incorrecta así como su variabilidad. La muestra se divide en 10 bloques de forma aleatoria, se entrena el modelo de clasificación con nueve de ellos y se deja fuera el restante. Una

vez estimado el modelo, se predicen las clases de los individuos del bloque que quedó fuera y se obtiene una proporción de clasificación incorrecta. Esto se repite 10 veces, hasta haber sacado fuera una vez cada uno de los bloques que componen la muestra.

4. RESULTADOS

En primer lugar se ha aplicado un método de clasificación no supervisada K-medias FDA para evaluar la posibilidad de identificar los distintos fondos marinos de la Ría de Cedeira a partir de curvas acústicas. Presenta la ventaja de ser un procedimiento más automático que los actualmente utilizados, no precisando un conocimiento profundo de la naturaleza física de los datos. En la Figura 2 se muestran cada una de las curvas acústicas obtenidas para cada muestra de terreno. Cada curva está compuesta de dos ecos relacionados con la incidencia de la onda sonora en el fondo marino y su posterior rebote, respectivamente. La Figura 2 muestra el grupo asignado a cada curva, de cuatro posibles (que debería relacionarse con dos tipos de roca, arena gruesa y arena fina). Se han calculado los centros representativos, en este caso también curvas acústicas, las medias funcionales.

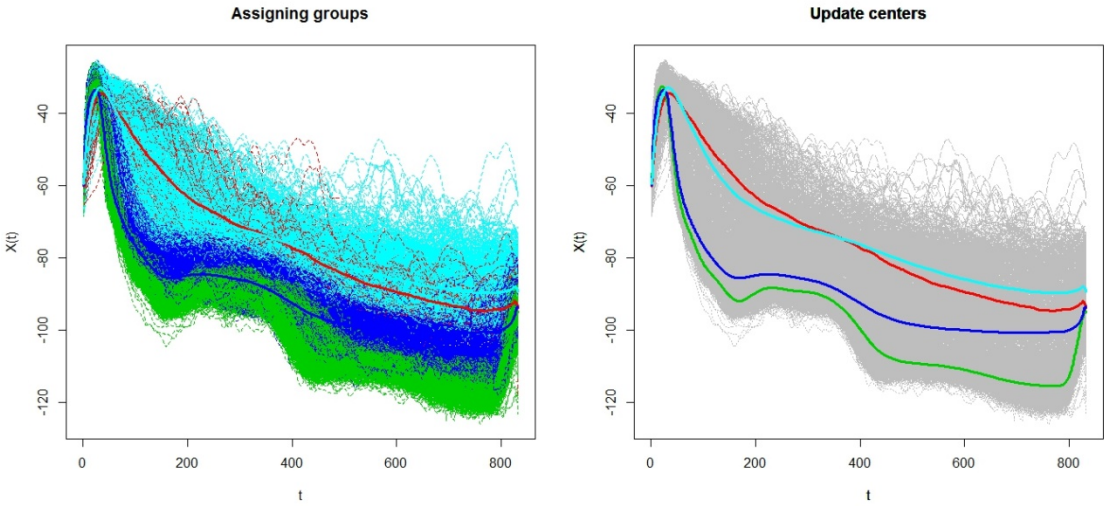


Figura 2. Panel izquierdo: ecos de cada muestra y grupo asignado (cuatro en total). Panel derecho: medias funcionales de cada grupo (centros).

En la Figura 3 se observa el mapa de la Ría de Cedeira con los transectos realizados por el barco. Se ha marcado el grupo al que pertenece cada muestra a partir de su curva de ecos utilizando el método K-medias funcional y la distancia L_1. Si se definen tres clases, se distinguen correctamente dos tipos de roca (en rojo y negro) y un tipo de arena si se comparan los resultados con otros estudios previos (Rodríguez-Pérez et al., 2014). Si se clasifican las curvas acústicas en cuatro clases, se distinguen tres tipos de roca y uno de arena. Esto es debido a que existen más diferencias desde un punto de vista acústico entre los grupos de rocas que entre los tipos de arena.

En el caso de Cabo de Palos (Murcia) se han evaluado métodos de clasificación supervisada FDA para identificar el tipo de fondos marino partiendo de un número definido de clases posibles: arena (CP01), arena con vegetación (CP02), rocas (CP03). De la primera, segunda y tercera clase se han obtenido 9295 curvas, 11235 y 4967, respectivamente, en total 25497 curvas evaluadas en 260 puntos/tiempos. Primeramente se comprobó que los datos acústicos son una característica discriminante del tipo de fondo. Para ello se realizó un contraste ANOVA funcional por el método de proyecciones aleatorias (Cuesta-Albertos y Febrero-Bande, 2010 ; Tarrío-Saavedra et al., 2011), resultando que la media de cada grupo son significativamente diferentes (ver también Figura 4).

El siguiente paso es la aplicación de los modelos de clasificación supervisada FDA mencionados. La Tabla 1 muestra las proporciones de clasificación incorrecta. *Y* es la respuesta cualitativa, clase. *X*, los ecos, variable funcional. Se evalúan los siguientes modelos de clasificación funcional GLM y GSAM mediante un procedimiento 10 fold cross-validation:

a) GLM1, donde X y β se representan suavizados con dos bases B-spline, de 15 y 7 elementos, respectivamente.
b) GLM2: donde X y β se representan suavizados con dos bases B-spline de 7 elementos.
c) GLM3: X se representa mediante una base de 10 FPC.
d) GSAM1: X se representa por una base de 10 FPC.
e) GSAM2: X se representa por una base de 10 FPC.

La Tabla 2 muestra que, utilizando el modelo GSAM2, se clasifican correctamente el 98.7% de los ecos, prácticamente la totalidad de los ecos. Por tanto, a la vista de las altas proporciones de identificación correcta, se demuestra que la aplicación de técnicas FDA de clasificación supervisada podría ser de gran utilidad para la clasificación de fondos marinos.

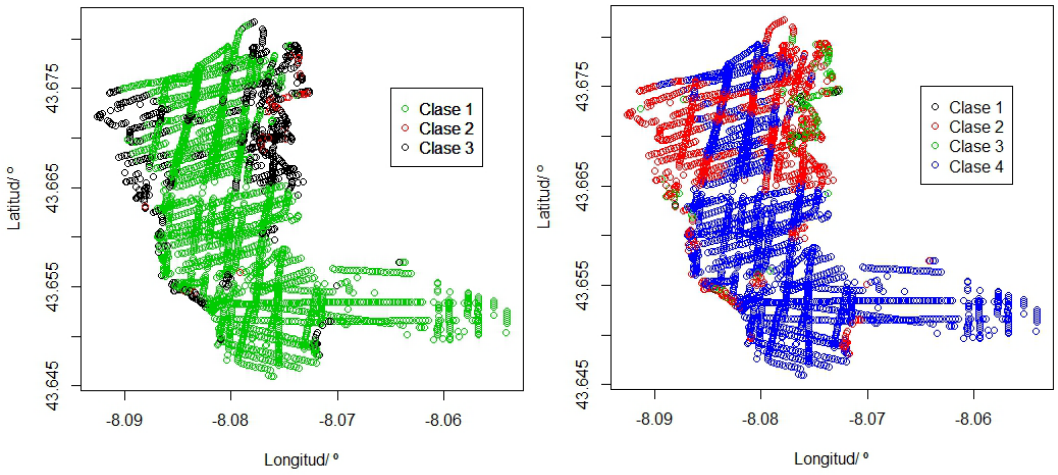


Figura 3. Mapa de la Ría de Cedeira con los transectos realizados por el barco. Se ha marcado el grupo al que pertenece cada muestra a partir de su curva de ecos utilizando el método K-medias funcional. Panel izquierdo: fijando 3 clases o grupos. Panel derecho: fijando 4 grupos diferentes.

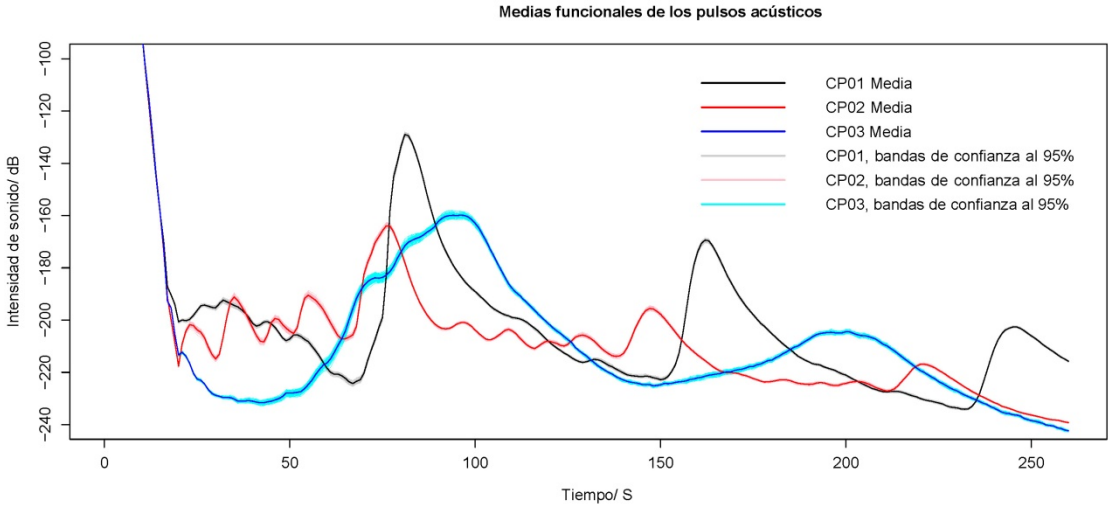


Figura 4. Medias funcionales y bandas de confianza al 95% para las curvas acústicas correspondientes a las tres clases de suelo de Cabo de Cabo de Palos (Murcia).

Tabla 1. Proporciones de clasificación incorrecta (media y desviación típica) con 10-fold cross-validation.

	Proporción de clasificación incorrecta	
Métodos	Media	Desviación estándar
GSAM1	0.050	0.003
GLM1	0.110	0.006
GLM2	0.151	0.007
GSAM2	0.013	0.002
GLM3	0.082	0.007

5. CONCLUSIONES

En este trabajo se han aplicado técnicas FDA de clasificación supervisada y no supervisada para la clasificación de tipos de fondo marino a partir de los datos obtenidos mediante una sonda de sónar. Se ha conseguido clasificar correctamente (entre fondo rocoso, arenoso y arenoso con vegetación) más del 98% de las muestras, utilizando para ello un modelo de clasificación supervisada funcional GSAM. Por tanto, el uso de técnicas FDA podría ser de gran utilidad en este campo.

Agradecimientos

Los autores agradecen la ayuda de los proyectos Simulación numérica de problemas hidroacústicos de alta frecuencia en medios mariños – SIMNUMAR (EM2013/052) y MTM2013-41383P (fondos FEDER incluidos).

REFERENCIAS

Cardot H., Ferraty F., Sarda P. (1999). Functional Linear Model. Statistics and Probability Letters, 45(1), 11-22.

Cuesta-Albertos J.A., Febrero-Bande M. (2010). A simple multiway anova for functional data. Test. 19, 537-57.

Febrero-Bande M., Oviedo de la Fuente M. (2012). Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51(4):128.

Febrero-Bande M., González-Manteiga W. (2013). Generalized additive models for functional data. TEST, 22(2):278292.

Ferraty F., Vieu P. (2006). Nonparametric functional data analysis. Springer Series in Statistics, New York.

Francisco M., Tarrío J., Mallik A., Naya S.(2012). A comprehensive classification of wood from thermogravimetric curves. Chemometrics and Intelligent Laboratory Systems. 118, 159-172.

Legendre P., Ellingsen K. E., Bjørnbom E., Casgrain P. (2002). Acoustic seabed classification: improved statistical method. Canadian Journal of Fisheries and Aquatic Sciences, 59, 1085-1089.

Preda C., Saporta G., Lévéder C. (2007). PLS classification of functional data. Computational Statistics, 22, 223-235.

Preston J. M., Kirlin T. L. (2003). Comment on “Acoustic seabed classification: improved statistical method”, Canadian Journal of Fisheries and Aquatic Sciences, 60, 1299-1300

Ramsay J., Silverman B.W. (2005). Functional Data Analysis. Springer.

Rodríguez-Pérez, D., Sánchez-Carnero, N., Freire, J. (2014). A pulse-length correction to improve energy-based seabed classification in coastal areas. Continental Shelf Research, 77, 113.

Tarrío, J., Naya, S., López, J., Artiaga, R. (2011). Application of functional ANOVA to the study of thermal stability of micronano silica epoxy composites. Chemometrics and Intelligent Laboratory Systems. 105, 114-124.

Wehrens, R. (2011). Chemometrics with R. Multivariate Data Analysis in the Natural Sciences and Life Sciences. Berlin Heidelberg: Springer.

Comunicacións orais

Sesión 4 de Estatística

UN POTENTE AROMA DE ESTADÍSTICA

Mónica Bécue-Bertaut[1] y Josep-Anton Sanchez Espigares[2]

[1] Fundació Miquel Augustí, Departament Estadística i Investigació Operativa
Universitat Politècnica de Catalunya
[2] Grup d'Investigació ABDB, Departament Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

RESUMEN

La evaluación sensorial de los alimentos es una actividad muy antigua. En el campo de los vinos, la descripción sensorial es objeto de una atención especial incluso elevada a la categoría de actividad artística. Las opiniones de los expertos son escuchadas, valoradas e incluso leídas, si se considera la abundante cantidad de guías de vinos publicadas. En ocasiones, estas descripciones son muy ricas y destacan las características esenciales de los vinos tal y como son percibidas por el catador. No obstante, son opiniones personales y su reproducibilidad, e incluso su repetibilidad son muy bajas. No se puede calificar estas descripciones como "medidas" convencionales.

Con el objetivo de poder disponer de medidas sensoriales, el mundo del vino ha adoptado las metodologías desarrolladas por la industria agroalimentaria, muy especialmente las basadas en el concepto de perfil descriptivo cuantitativo. Primero, se establece una lista de descriptores sensoriales que conforma la ficha de la degustación. A continuación un conjunto de catadores entrenados da a cada descriptor una valoración numérica (nota). El conjunto de notas da lugar a diferentes tratamientos estadísticos.

Esta metodología, muy satisfactoria, es la base de descripciones sensoriales que pretenden alcanzar el estatus de "medida". En general, el comportamiento del jurado considerado en conjunto suele ser excelente o suficientemente bueno. A pesar de ello, esta metodología no da información de la importancia relativa de los diferentes descriptores en la percepción del producto, como por ejemplo, que un catador valore como muy importante el amargor y otro, la acidez.

Por ello, se ha propuesto evaluar globalmente las diferencias entre productos. El método del "napping®" (Pagès, 2005), del que se presenta a continuación una aplicación, pertenece a este último punto de vista y su objetivo es recoger directamente una medida global de similaridades/disimilaridades sensoriales entre productos. Para interpretar las distancias sensoriales, son necesarios términos que describen los vinos. Resulta sencillo recoger, en el mismo momento del "napping®", las palabras que los catadores asocian libremente a los vinos (*Flash Profiling*).

Esta comunicación describe el protocolo seguido en la recogida de datos y la metodología estadística utilizada para tratar dichos datos, basada en el *análisis factorial múltiple (AFM;* Escofier & Pagès, 1992) y el *análisis factorial múltiple para tablas de contingencia AFMTC (*Bécue & Pagès, 2008*).* Se ilustra con los resultados de la cata realizada en el marco del XII Congreso Galego de Estatística e Investigación de Operacións (SGAPEIO2015).

Palabras y frases clave: Sensometría; Napping; Flash Profiling; Análisis factorial múltiple;

REFERENCIAS

Bécue-Bertaut M., Pagès, J. (2008). Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data *Computational Statistics and Data Analysis*

Escofier, B., Pagès, J. (1992) Anàlisis factoriales simples y múltiples. Universidad del País Vasco, Servicio Editorial / Euskal Herriko Unibersitatea, Argitaparen Zerbitzua. Bilbo.

Pagès, J. (2005) Collection and analysis of perceived product interdistances using multiple factor analysis: application to the study of 10 white wines from the Loire Valley. Food Quality and Preferences. 16, 64 –649.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

VARIACIÓN DE LA DENSIDAD DE LA MADERA A LO LARGO DEL FUSTE EN EUCALYPTUS GLOBULUS. INFLUENCIA DE LOS INCENDIOS FORESTALES

Guillermo Riesco Muñoz

Departamento de Enxeñería Agroforestal, Universidade de Santiago de Compostela. Escuela Politécnica Superior. Campus Universitario s/n. 27002 Lugo (España). guillermo.riesco@usc.es

RESUMEN

La magnitud del problema de los incendios forestales en Galicia hace que cada año se vean afectadas por el fuego numerosas plantaciones de eucaliptos. Los ejemplares que son afectados pero no destruidos, siguen teniendo interés comercial y son objeto de aprovechamiento maderero aunque la madera procedente de estos pies sufra una depreciación importante. En este trabajo se ha analizado el peso específico anhidro (densidad anhidra) de la madera y su variación a lo largo del fuste, comparando un grupo de 15 árboles sanos con otro grupo de 10 árboles afectados por el fuego, a fin de averiguar si la depreciación de la madera está justificada por una significativa pérdida de propiedades del material. Se encontró que la representación gráfica mediante nubes de puntos, donde estos no representan pares de valores de mediciones individuales sino promedios de mediciones, permitía explorar cuál es la altura del fuste a partir de la que empieza a haber diferencias apreciables entre la densidad anhidra de pies verdes y la de pies afectados por el fuego. Dicha exploración gráfica preliminar se contrastó con regresión percentílica (percentiles del 95 y 5 %), encontrándose para la muestra analizada que hasta los 11 m de altura en el fuste, correspondiente a un diámetro con corteza de unos 9 cm, no hay diferencias significativas de densidad (al 10 % de nivel de significación) entre la madera procedente de árboles sanos y la procedente de árboles afectados por el fuego.

Palabras y frases clave: propiedad física, análisis exploratorio, regresión percentílica.

1. INTRODUCCIÓN

Las masas forestales de la especie *Eucalyptus globulus* Labill. ocupan un total de 433.916 ha en Galicia, ya sea en masas puras o en mezcla con otras especies, considerando tanto las plantaciones como las masas procedentes de rebrote. Dicha superficie representa el 31 % de toda la superficie forestal arbolada de la comunidad, según las estadísticas oficiales más recientes, que son las procedentes del Cuarto Inventario Forestal Nacional (Dirección General de Medio Natural y Política Forestal, 2011). Por otra parte, la incidencia de los incendios forestales es tan elevada en Galicia que con frecuencia las masas arboladas sufren los daños del fuego, incluyendo a las formaciones de eucaliptos. De hecho, *Eucalyptus globulus* es una de las especies más afectadas por los incendios forestales, tanto a nivel nacional como en Galicia (Ministerio de Medio Ambiente, 2006).

Después de un incendio forestal se impone la rápida extracción de los productos maderables, tanto para evitar el peligro de plagas para masas forestales vecinas como para evitar el progresivo deterioro de la madera afectada y la consiguiente depreciación de la misma. Asimismo, la retirada de los troncos quemados facilita las labores posteriores de gestión forestal y obedece también, aunque en menor medida, a razones estéticas. Además, determinados árboles en zonas quemadas pueden parecer sanos pero su posible debilidad fisiológica les hace ser los primeros candidatos a ser atacados por plagas defoliadoras e insectos xilófagos, sobre todo si se dan condiciones climáticas desfavorables, por lo que la corta de árboles vivos puede estar justificada si se sospecha un daño por fuego en el interior del tronco. Por otro

lado, la permanencia de pies quemados supone una peligrosa carga de combustible en la zona, que incrementa la probabilidad y la severidad potencial de un incendio futuro (Bautista *et al*., 2004).

Ante el paso de un incendio el árbol puede quedar intacto o bien afectado por el fuego, o incluso quemado, influyendo en ello tanto las características del árbol, por ejemplo el espesor de la corteza, como el "comportamiento del incendio". Dicho comportamiento (calor desprendido por unidad de superficie, intensidad, longitud de llama, velocidad de propagación) se puede estimar mediante modelos implantados en sistemas informáticos. Entre estos sistemas destaca BehavePlus (Carlton *et al*., 2001), basado en la ecuación de propagación de Rothermel (Rothermel, 1972; Andrews, Chase, 1989), que es el modelo de comportamiento del fuego de superficie de mayor prestigio y reconocimiento internacional.

Los eucaliptos sanos suelen ser transformados por la industria de la celulosa, para la fabricación de pasta al sulfato, mientras que los pies afectados por el fuego tienen el mismo destino industrial aunque sufren una depreciación. Cuando la calidad y dimensiones de las trozas de eucalipto lo permiten, la madera es transformada por industrias de aserrado y chapa, pero si la madera está afectada por el fuego es dirigida a la industria de desintegración, en las que la materia prima está menos cotizada. En uno u otro caso esta depreciación se atribuye en parte a una supuesta reducción en la calidad de la madera procedente de árboles dañados por el fuego, reducción que debería ser apreciable en aquellas propiedades físicas y mecánicas que comúnmente se utilizan como indicadores de calidad de esta materia prima.

En el presente trabajo se ha analizado el peso específico anhidro de la madera (densidad anhidra) y su variación a lo largo del fuste, comparando madera de árboles sanos y de árboles afectados por el fuego, a fin de averiguar si la depreciación de la madera está justificada por una significativa pérdida de propiedades del material.

2. MATERIAL Y MÉTODOS

Para la toma de muestras en las que basar la realización de este trabajo se apearon 15 ejemplares verdes y aparentemente sanos de eucalipto (*Eucalyptus globulus* Labill.) así como 10 ejemplares de la misma especie afectados por el fuego y aparentemente secos. Con ello se superaba el tamaño de muestra mínimo de 5 árboles que establece la norma UNE-EN 384:2010 (AENOR, 2010). Los árboles de la muestra, cuyas principales características se indican en la tabla 1, se encontraban próximos a la edad de corta final que se suele aplicar a esta especie en Galicia (González-Río *et al*., 1997) y procedían de seis masas forestales, localizadas en los siguientes municipios: Lorenzana y Ribeira de Piquín en Lugo, La Coruña y Mesía en la provincia de La Coruña, Catoira y Pontevedra en la provincia de Pontevedra. El diámetro normal de los árboles de la muestra se midió con forcípula de ± 0,5 cm de precisión y la altura total con hipsómetro Blume-Leiss con precisión de ± 0,5 m. La edad se determinó a partir de la fecha de plantación.

		mínimo	media	máximo
Pies verdes (n = 15)	Diámetro normal (cm)	11,3	14,6	17,3
	Altura total (m)	9,4	15,1	20,0
	Edad (años)	9	10,1	12
Pies dañados (n = 10)	Diámetro normal (cm)	10,5	13,0	16,3
	Altura total (m)	14,8	16,2	18,6
	Edad (años)	10	12,1	13

Tabla 1: Características dendrométricas de los ejemplares de *Eucalyptus globulus* de la muestra. n = número de pies de la muestra.

Para analizar las propiedades de la madera, de cada fuste apeado se extrajeron rodajas de la sección transversal completa a diferentes alturas en el tronco: en la base, a 1 m de altura, a 2 m de altura, etc. hasta obtener en algunos ejemplares rodajas a 11 m de altura, lo que representaba el 67 % de la altura total del fuste. Se pretendía obtener rodajas tan arriba como fuera posible en el fuste, para hacer así más apreciable la previsible tendencia de la variable densidad a lo largo del tronco. No obstante, no fue posible obtener rodajas a alturas superiores a 11 m ya que su pequeño diámetro imposibilitaba la extracción posterior de las probetas de ensayo normalizadas que se requieren para la medida de la densidad.

Las rodajas se cortaron con un espesor aproximado de 10 cm y seguidamente se labraron en carpintería para obtener de ellas probetas prismáticas de pequeñas dimensiones (dimensiones nominales: 2 cm de dimensión radial, 2 cm de dimensión tangencial y 4 cm de dimensión longitudinal) y libres de defectos, que fueron desecadas hasta estado anhidro en una estufa a 103 ºC durante una semana. Tras el secado, y enfriado posterior de las probetas en un ambiente desecante, se determinó para cada pieza su peso con precisión de centésima de gramo y sus tres dimensiones principales (dimensión radial, tangencial y longitudinal), con precisión de centésima de milímetro. Se consideró que cada probeta era un prisma perfecto por lo que el volumen de cada pieza se obtuvo como producto de sus tres dimensiones principales. El peso específico anhidro se obtuvo por cociente entre el peso medido en estado anhidro y el volumen calculado para la probeta, también en estado anhidro, expresado en kilogramos por metro cúbico. En promedio se obtuvieron unas 11 probetas válidas por rodaja y en total fueron válidamente ensayadas 2.145 probetas, magnitud de muestra que superan ampliamente los mínimos establecidos en la norma UNE 56528:1978 (AENOR, 1978).

El análisis sobre probetas de pequeñas dimensiones sin defectos permite caracterizar la madera de una especie y comparar madera de diversas procedencias ya que la presencia de defectos o singularidades en la madera es una fuente de variación que impediría detectar la influencia de otros factores en las propiedades del material.

Para la representación gráfica de pares de valores se empleó la hoja de cálculo Microsoft Excel 2010. Para la obtención de histogramas y estadísticos descriptivos de las variables se empleó el programa estadístico IBM SPSS Statistics versión 20. Para la obtención de la regresión percentílica se empleó el programa estadístico R (Koenker, 2015; R Core Team, 2014).

3. RESULTADOS Y DISCUSIÓN

La distribución de los valores y los principales estadísticos descriptivos de la muestra de peso específico anhidro se pueden encontrar en la Figura 1 y en la Tabla 2, que se obtuvieron tras validar y depurar los resultados de los ensayos. La dispersión de resultados es similar a la que indica la norma UNE 56528:1978 (AENOR, 1978) para las propiedades gravimétricas obtenidas sobre probetas de pequeñas dimensiones sin defectos. Tanto la Figura 1 como la Tabla 2 no muestran una distinción clara entre los grupos madera verde/madera dañada para la variable peso específico anhidro.

variable	*n*	mínimo	media	máximo	*CV* (%)
ρ_0 (verde)	1.454	334	624	1.042	14
ρ_0 (afectado)	691	489	616	939	10
ρ_0 (total)	2.145	334	621	1.042	13

Tabla 2: Características de la muestra de datos de peso específico anhidro. *n* = tamaño de muestra; *CV* (%) = coeficiente de variación (relación entre la media y la desviación típica en porcentaje); ρ_0 = peso específico anhidro (kg/m^3).

Por otra parte, en la representación de la nube de puntos correspondiente a los pares de valores peso específico anhidro de la probeta frente a altura de la probeta en el fuste (Figura 2) no son claramente distinguibles tendencias diferentes en los dos grupos de probetas (madera verde/madera dañada). Debido a ello se combinaron ambos grupos de probetas en uno solo, para el que se obtuvo el coeficiente de correlación lineal de Pearson entre peso específico anhidro y altura en el fuste, que resultó ser $R = 0{,}13$, valor muy pobre aunque altamente significativo ($p < 0{,}01$) debido probablemente al elevado número de pares de valores en el que se basa el cálculo. Debido al amplio rango de variación de alturas totales en la muestra (Tabla 1) se probó a utilizar la altura relativa en el fuste (altura en el fuste dividida por la altura del árbol muestreado, expresada en porcentaje) en lugar de la altura absoluta en metros. Sin embargo, en este caso el valor del coeficiente de correlación era aún más bajo ($R = 0{,}11$; $p < 0{,}01$). La correlación lineal se muestra por tanto insuficiente para la nube de puntos de la Figura 2 ya que se trata de una colección de puntos con una pendiente inapreciable, salvo en el primer metro de altura de fuste, en el que se da una gran reducción de densidad para después estabilizarse. Además, la distribución no sigue un patrón lineal y existe una elevada dispersión de valores de peso específico para cada altura de fuste

(coeficiente de variación promedio del 8 %). Debido a ello se renunció a modelizar la nube de puntos de la Figura 2, ni para la muestra en su conjunto ni para cada grupo de probetas por separado.

Para encontrar una representación gráfica en la que fueran más claramente observables las tendencias entre pares de variables se procedió a representar el valor medio del peso específico anhidro para cada altura de fuste en cada árbol (valor medio del peso específico por rodaja), resultando la representación de la Figura 3. Debido a que esta nueva representación tampoco revelaba tendencias divergentes entre grupos de probetas se representó en la figura 4 la media del peso específico anhidro para todas las probetas de todos los árboles para cada altura. En este caso de máxima simplificación (un valor para cada altura de fuste en cada grupo de datos) sí se aprecia una tendencia a una menor densidad en la madera de pies afectados por el fuego a partir de los seis metros de altura en el fuste. Para contrastar la significación estadística de esta diferencia de tendencias entre los dos grupos de probetas se procedió a efectuar una regresión percentílica, que consistió en obtener por regresión ponderada la curva del percentil del 95 % y la del percentil del 5 % para cada nube de puntos (probetas procedentes de árboles sanos/probetas procedentes de árboles afectados). Para ello se ajustó para cada conjunto de puntos un polinomio completo de orden 3 por el método de ajuste de mínimos cuadrados ordinarios, siendo las curvas percentílicas ecuaciones de la misma naturaleza ($y = a + bx + cx^2 + dx^3$). Las curvas obtenidas aparecen de forma analítica y de forma gráfica en la Tabla 3 y en la Figura 5 respectivamente. Se observa (Tabla 3) que en la curva del percentil 5 % para madera de pies afectados por el fuego los parámetros *b*, *c* y *d* no son significativamente distintos de cero por lo que debería recalcularse la ecuación, si bien se desiste de esta posibilidad dado que dicha curva no afecta a la distinción entre poblaciones de madera sana/madera afectada, que es el objetivo buscado, y que depende de la curva del percentil superior.

		parámetro	valor	error típico	*t*	Sig.
Madera de pies verdes	Curva del percentil 95 %	*d*	-1,61405	0,16636	-9,70199	0,00000
		c	29,17649	2,80919	10,38609	0,00000
		b	-140,20758	13,53972	-10,35528	0,00000
		a	911,99805	20,15138	45,25734	0,00000
	Curva del percentil 5 %	*d*	-0,24838	0,13170	-1,88598	0,05950
		c	4,25780	1,93130	2,20463	0,02764
		b	-4,53782	7,89531	-0,57475	0,56555
		a	465,14699	10,43775	44,56392	0,00000
Madera de pies afectados por el fuego	Curva del percentil 95 %	*d*	-0,65047	0,21918	-2,96770	0,00310
		c	13,74701	4,04367	3,39963	0,00071
		b	-90,87358	23,09181	-3,93532	0,00009
		a	874,21282	42,59010	20,52620	0,00000
	Curva del percentil 5 %	*d*	-0,22333	0,20431	-1,09308	0,27474
		c	2,87355	3,37990	0,85019	0,39552
		b	-5,39073	15,76452	-0,34195	0,73249
		a	517,46147	19,94507	25,94433	0,00000

Tabla 3: Características de la curvas percentílicas. *a*, *b*, *c* y *d* = parámetros del modelo; *t* = t de Student; Sig. = nivel de significación.

En la figura 5 se observa que, en el rango de alturas de fuste analizado, la curva ajustada para el percentil 5 % de madera procedente de árboles sanos nunca toma valores superiores a la curva del percentil 95 % de madera procedente de pies dañados. Por tanto, el resultado que apuntaba la figura 4 hacia una posible diferenciación de tendencias en relación al peso específico anhidro frente a la altura en el fuste no es estadísticamente aceptable al 10 % de nivel de significación para ningún intervalo de alturas del fuste en el rango de alturas analizado. De este modo, hasta los 11 m de altura en el tronco (equivalente en la muestra a un diámetro de fuste de unos 8,7 cm) la madera procedente de eucaliptos sanos y la madera procedente de eucaliptos afectados por el fuego no presentan diferencias significativas de densidad, por lo que la depreciación que sufre el eucalipto a causa de verse afectado por el fuego no es justificable atendiendo a cambios de densidad de la materia prima.

Extrapolando las curvas obtenidas en la figura 5 para alturas de fuste superiores a los 11 m podría predecirse una diferenciación significativa entre densidades por la acción del fuego (pérdida de densidad

de la madera dañada), si bien el volumen de madera que se vería afectado sería irrelevante dado el escaso diámetro de los fustes a alturas próximas al ápice.

AGRADECIMIENTOS

Al profesor don Juan Gabriel Álvarez González, por su colaboración en el análisis estadístico de los resultados experimentales.

REFERENCIAS

AENOR (1978) UNE 56528:1978. Características físico-mecánicas de la madera. Preparación de probetas para ensayos. AENOR.

AENOR (2010) UNE-EN 384:2010. Madera estructural. Determinación de los valores característicos de las propiedades mecánicas y la densidad. AENOR.

Andrews, P. L., Chase, C. H. (1989) BEHAVE: Fire Behavior Prediction and Fuel Modeling System-BURN Subsystem, Part 2. USDA Forest Service. National Wildfire Coordinating Group. Gen. Tech. Rep. INT-260.

Bautista, S., Gimeno, T., Mayor, A. G., Gallego, D. (2004) El tratamiento de la madera quemada tras los incendios forestales. En: Avances en el estudio de la gestión del monte mediterráneo. (Vallejo, R. y Alloza, J. A., eds.). Fundación Centro de Estudios Ambientales del Mediterráneo.

Carlton, D., Andrews, P. L., Bevins, C. (2001) Users´ Guide BehavePlus Fire Modeling System Version 1.00.00. USDA Forest Service. Rocky Mountain Research Station. Systems for Enviromental Management.

Rothermel, R. C. (1972) A mathematical model for predicting fire spread in wildland fuels. USDA Forest Service. Intermountain Forest and Range Experiment Station. Res. Pap. INT-115.

Dirección General de Medio Natural y Política Forestal (2011) Cuarto Inventario Forestal Nacional. Galicia.

González-Río, F., Castellanos, A., Fernández, O., Astorga, R. y Gómez, C. (1997) Manual de selvicultura del eucalipto. Escola Politécnica Superior de Lugo, Celulosas de Asturias S.A.

Koenker, R. (2015) quantreg: Quantile Regression. R package version 5.11. http://CRAN.R-project.org/package=quantreg

Ministerio de Medio Ambiente (2006) http://www.marm.es/

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

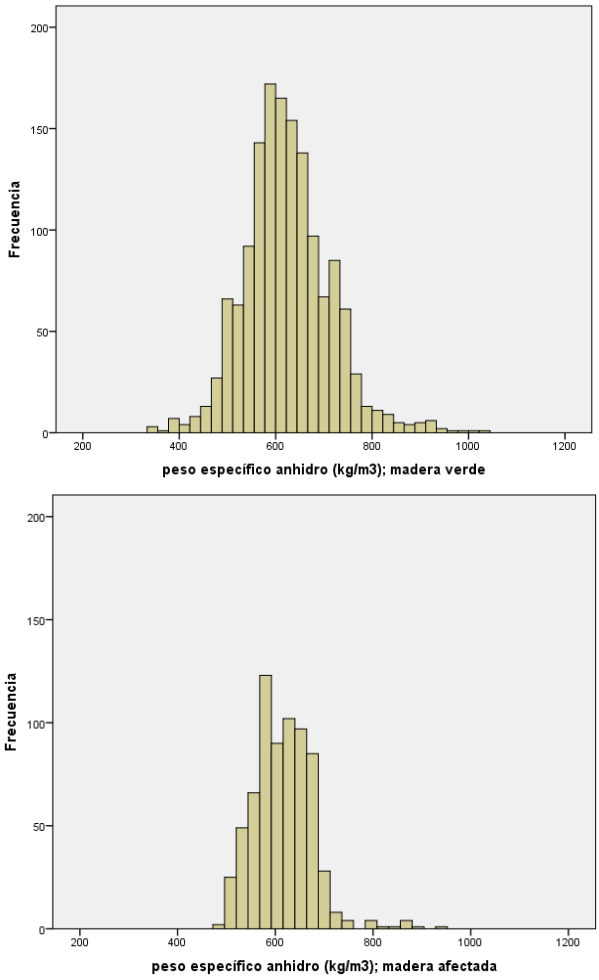


Figura 1: Distribución de la variable peso específico anhidro.

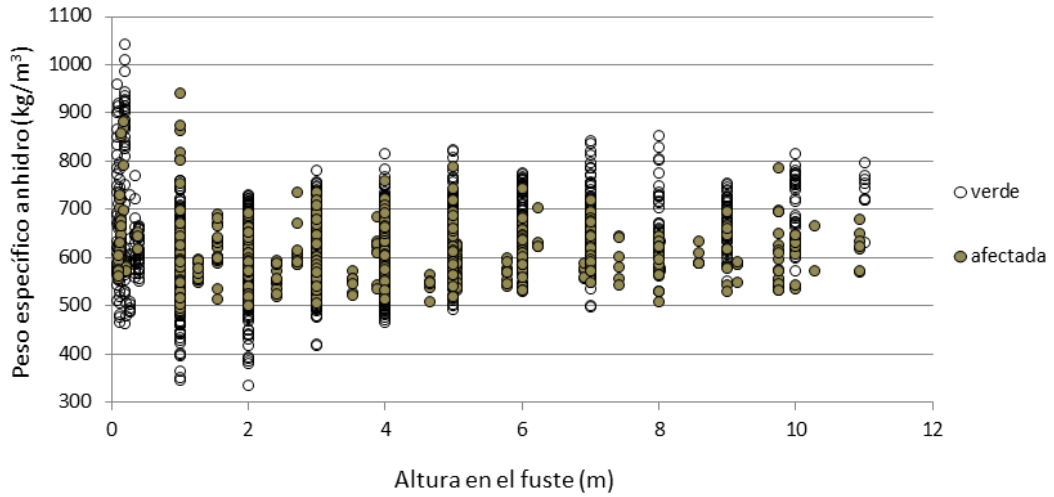


Figura 2: Variación del peso específico anhidro a lo largo del fuste. Cada punto representa una probeta ensayada.

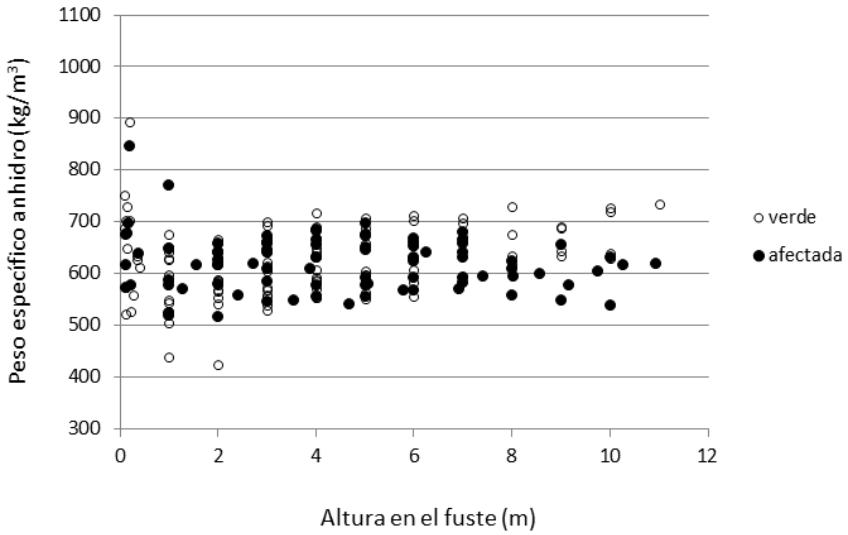


Figura 3: Variación del peso específico anhidro a lo largo del fuste. Cada punto representa la media del peso específico anhidro de todas las probetas ensayadas en una rodaja (probetas procedentes de cierta altura en un árbol determinado).

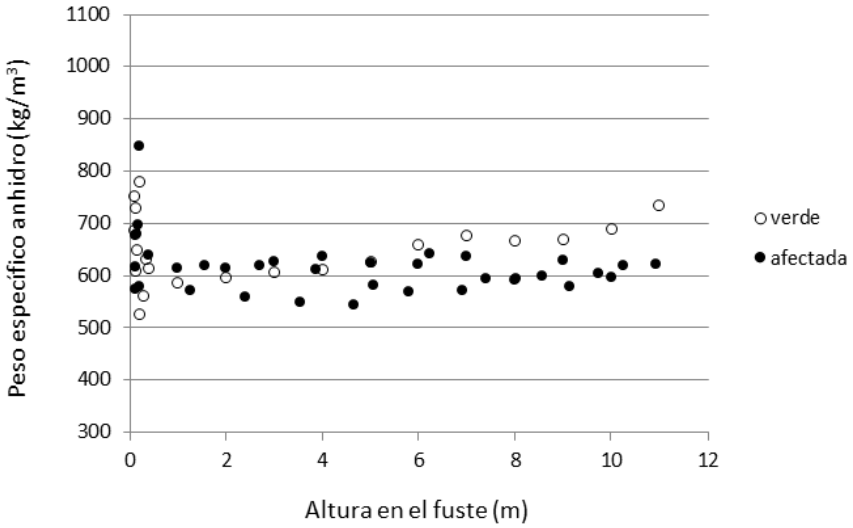


Figura 4: Variación del peso específico anhidro a lo largo del fuste. Cada punto representa la media del peso específico anhidro de todas las probetas de todos los árboles para cierta altura en el fuste.

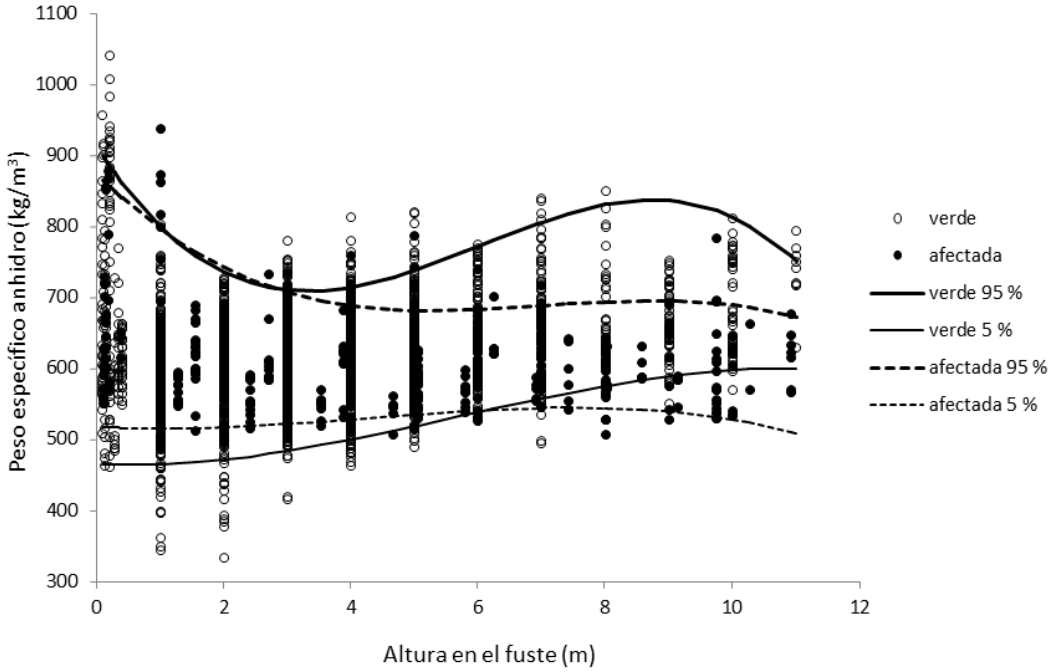


Figura 5: Variación del peso específico anhidro a lo largo del fuste. Cada punto representa una probeta ensayada. Las líneas de trazo continuo representan las curvas de ajuste del percentil 95 % y 5 % para las probetas procedentes de árboles sanos. Las líneas de trazo discontinuo representan las curvas de ajuste del percentil 95 % y 5 % para las probetas procedentes de árboles afectados por el fuego.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Avaliação do Desempenho Estatístico de um Novo Método Combinado para obter os Instantes de Amostragem em Controlo de Qualidade

Manuel do Carmo[1,2], Paulo Infante[2,3] e Jorge M. Mendes[4]

[1] Universidade Europeia|Laureate International Universities, Lisboa, Portugal
[2] CIMA da Universidade de Évora, Évora, Portugal
[3] DMAT, ECT da Universidade de Évora, Évora, Portugal
[4] MagIC e NOVA IMS da Universidade Nova de Lisboa, Lisboa, Portugal

RESUMO

Neste trabalho, analisamos a eficiência estatística de um novo método de amostragem, em controlo de qualidade, que combina dois outros métodos: um em que os instantes de amostragem são definidos em função da estatística amostral (adaptativo), e outro cujos parâmetros são definidos no início do processo a ser monitorizado, mas que permanecem constantes ao longo do mesmo (predefinido). Desta forma, os instantes de amostragem, inicialmente calendarizados de acordo com as expectativas de ocorrência de uma alteração, tomando como base a distribuição do tempo de vida do sistema, são adaptados em função do valor da estatística amostral calculada no instante anterior.

Utilizando cartas de controlo para a média do tipo *Shewhart*, realizamos uma análise do desempenho estatístico do método apresentado, considerando diferentes alterações da média, diferentes valores para a dimensão amostral e diferentes pesos para os instantes de amostragem definidos pelo esquema de intervalos adaptativo e pelo esquema de intervalos predefinidos.

Os resultados, obtidos nesta fase por simulação, permitem concluir que o método apresenta um bom desempenho estatístico para diferentes alterações da média (inclusive grandes alterações) quando comparado com outros esquemas amostrais adaptativos, em termos do tempo médio de mau funcionamento do sistema e do número médio de itens inspeccionados.

Palavras e frases chave: Intervalos adaptativos; Intervalos Predefinidos; ,Simulação; Controlo de Qualidade.

1. INTRODUÇÃO

Actualmente, qualquer organização, de produção ou serviços, faz avaliações rigorosas da qualidade dos bens que disponibiliza ao consumidor, recorrendo a ferramentas estatísticas adequadas. Para monitorizar uma, ou mais, das suas características, a carta de controlo tem sido uma das ferramentas mais utilizadas, pois permite a distinção entre a variação intrínseca do processo e a variação com origem em causas externas.

O desempenho das cartas de controlo está directamente relacionada com a dimensão amostral e com o modo como são definidos os instantes de recolha das amostras. Para melhorar o desempenho do esquema clássico (amostras de tamanho fixo recolhidas em intervalos igualmente espaçados no tempo) surgiram métodos que definem o tamanho das amostras e/ou os instantes de amostragem em função do valor da estatística amostral (designados adaptativos) e outros em que estes parâmetros são determinados no início do controlo do processo em função da distribuição do tempo de vida (designados predefinidos). Se por um lado é desejável analisar o menor número de amostras (de modo a reduzir custos de amostragem e associados a falsos alarmes), por outro lado é fundamental que a carta seja eficaz em detectar as alterações na qualidade (de modo a que sejam reduzidos os custos de mau funcionamento).

Em Carmo *et al.* (2013) e em Carmo (2014) é apresentado um novo método adaptativo de amostragem no qual os intervalos entre amostras são obtidos com base na função densidade da distribuição de *Laplace* reduzida. Este método, denominado LSI (*Laplace Sampling Intervals*), revelou-se

particularmente eficiente na detecção de alterações da média moderadas e grandes, pouco sensível à limitação do menor intervalo de amostragem e robusto para a não normalidade da característica da qualidade. Em determinadas situações, este método é sempre mais eficiente do que o método de amostragem VSI (*Variable Sampling Intervals*).

Tendo como base o método de amostragem LSI e um método no qual os intervalos de amostragem são definidos antes do início do controlo do processo com base na taxa cumulativa de risco do sistema (Rodrigues Dias & Infante (2008)), em Carmo *et al.* (2014) e em Carmo (2014) é apresentado um método, denominado CAPSI (*Combined Adaptive and Predetermined Sampling Intervals*), que combina o método de intervalos predefinidos com o método de intervalos adaptativos. Neste método, os instantes de amostragem são definidos pela média ponderada dos instantes dos dois métodos, atribuindo-se maior peso ao método adaptativo para alterações moderadas (onde o método predefinido é menos eficaz) e maior peso ao método predefinido nos restantes casos (onde o método adaptativo é menos eficaz). Desta forma, os instantes de amostragem, inicialmente calendarizados de acordo com as expectativas de ocorrência de uma alteração, tomando como base a distribuição do tempo de vida do sistema, são adaptados em função do valor da estatística amostral calculada no instante anterior. Utilizando cartas de controlo para a média, do tipo *Shewhart*, e considerando que a característica da qualidade a ser monitorizada tem distribuição normal, os autores mostram que este método é sempre mais eficiente do que o método periódico clássico, o que não acontece com nenhum outro esquema adaptativo, e do que o método de amostragem VSI (Reynolds *et al.* (1988)) para alguns pares de amostragem usualmente considerados.

Neste trabalho, recorrendo à simulação, apresentamos um estudo de comparação do desempenho estatístico, em termos de AATS (*Adjusted Average Time to Signal*) e de ANOS (*Average Number of Observations to Signal*), do método de amostragem CAPSI com o dos métodos adaptativos VSS (Prabhu *et al.* (1993) e Costa (1994)), que considera dimensões amostrais adaptativas, VSSI (Prabhu *et al.* (1994)), que considera instantes de amostragem e dimensões amostrais adaptativos e VP (Costa (1999)), que considera, também, os limites de controlo adaptativos. Os resultados obtidos revelam, em geral, um muito bom desempenho do método CAPSI, quando se usam cartas de controlo para a média, posicionando-o como uma forte alternativa aos procedimentos de amostragem em comparação.

Por fim devemos realçar que o método CAPSI faz variar um só parâmetro, o que é uma vantagem em aplicações práticas, enquanto que os restantes métodos fazem variar mais do que um parâmetro.

REFERÊNCIAS

Carmo, M. (2014). *Políticas de Amostragem em Controlo Estatístico da Qualidade.* PhD Thesis, NOVA IMS, Universidade Nova de Lisboa, Lisboa.

Carmo, M.; Infante, P.; Mendes, J. M. (2013). Alguns resultados da robustez de um método de amostragem adaptativo em controlo de qualidade In Maia, M., Campos, P. & Silva, P. D. (Eds.), *Estatística: Novos Desenvolvimentos e Inspirações, SPE* (pp. 95-108). Porto.

Carmo, M.; Infante, P.; Mendes, J. M. (2014). Cartas de controlo com intervalos de amostragem dinâmicos inicialmente predefinidos para um risco cumulativo constante. In Pereira, I., Freitas, A., Scotto, M., Silva, M. E. & Paulino, C. D. (Eds.), *XXI Congresso Anual da Sociedade Portuguesa de Estatística, SPE* (pp. 183-195). Aveiro.

Costa, A. F. B. (1994). X-bar charts with variable sample size. *Journal of Quality Technology, 26*(3), 155-163.

Costa, A. F. B. (1999). X-bar charts with variable parameters. *Journal of Quality Technology, 31*(4), 408-416.

Prabhu, S. S.; Montgomery, D. C.; Runger, G. C. (1994). A combined adaptive sample size and sampling interval X-bar control scheme. *Journal of Quality Technology, 26*(3), pp.164-176.

Prabhu, S. S.; Runger, G. C.; Keats, J. B. (1993). An adaptive sample size X-bar chart. *International Journal of Production Research, 31(2)*, 2895-2909.

Reynolds, M. R.; Amin, R. W.; Arnold, J. C.; Nachlas, J. A. (1988). X-bar charts with variables sampling intervals. *Technometrics, 30*(2), 181-192.

Rodrigues Dias, J.; Infante, P. (2008). Control charts with predetermined sampling intervals. *International Journal of Quality and Reliability Management, 25*(4), 423-435.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

MODELOS MIXTOS DE POISSON: APLICACIÓN A DATOS DE INCENDIOS FORESTALES EN GALICIA

Miguel Boubeta[1], María José Lombardía[1], Manuel Marey-Pérez[2] y Domingo Morales[3]

[1] Departamento de Matemáticas, Universidade da Coruña.
[2] Departamento de Enxeñería Agroforestal, Universidade de Santiago de Compostela.
[3] Centro de Investigación Operativa, Universidad Miguel Hernández de Elche.

RESUMEN

Los incendios forestales están considerados en la actualidad como una de las principales causas de pérdida de masa forestal. Esta problemática tiene especial incidencia en las zonas Mediterráneas, donde los incendios aumentan de forma considerable en los periodos de primavera y verano. Los modelos mixtos de Poisson suponen una alternativa atractiva para modelizar variables respuesta de naturaleza discreta tal como el número de incendios forestales bajo una unidad territorial. Las razones de su éxito se deben a su buen comportamiento a altos niveles de desagregación, donde las modelizaciones más convencionales poseen un elevado error. En este trabajo se aborda la modelización del número de incendios en Galicia (2006-2008) por áreas forestales a través de diferentes alternativas: el modelo de Poisson mixto con efecto de área (Boubeta et al., 2015), el modelo de Poisson mixto con efectos de área y tiempo (independiente y AR(1)-correlado) y el modelo de Poisson mixto con correlación espacial. Se ha realizado un análisis comparativo de diferentes métodos de estimación de los parámetros tales como el método de los momentos (Jiang, 1998) u otros métodos basados en la aproximación de la verosimilitud tales como PQL o el método de Laplace (Saei and Chambers, 2003; Demidenko, 2004). Bajo este contexto, hemos desarrollado los correspondientes Empirical Best Predictors (EBP) y se han comparado los estimadores propuestos con otros estimadores existentes en la literatura tales como el estimador directo o el plug-in. Se realizarán varios estudios de simulación para estudiar el comportamiento de los diferentes estimadores utilizados. Finalmente, como medida de precisión de los EBP's proponemos un MSE bootstrap dado que su comportamiento es similar a la versión analítica y menos costoso computacionalmente.

Palabras y frases clave: Incendios forestales, modelos mixtos de Poisson, empirical best predictor, error cuadrático medio, bootstrap.

REFERENCIAS

Boubeta, M., Lombardía, M. J., Marey-Pérez, M. and Morales, D. (2015). Prediction of forest fires occurrences with area-level Poisson mixed models. Journal of Environmental Management, 154, 151-158.
Demidenko, E. (2004) Mixed models: theory and applications. Wiley.

Jiang, J. (1998) Consistent estimators in generalized linear models. Journal of the American Statistical Association, 93, 720-729.

Saei, A. and Chambers, R. (2003) Small area estimation under linear and generalized linear mixed models with time and area effects. Methodology working paper No. M03/15. Southampton Statistical Sciences Research Institute, University of Southampton, UK.

Comunicacións orais

Sesión 5 de Estatística

XI Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

CONSTRUCCIÓN DE UN INDICADOR SINTÉTICO DE COYUNTURA ECONÓMICA EN GALICIA

Esther López Vizcaíno[1], Patricio Sánchez Fernández[2] y Carlos Iglesias Patiño[1]

[1] Instituto Galego de Estatística
[2] Universidade de Vigo

RESUMEN

El presente trabajo construye un indicador sintético que pretende proporcionar una herramienta para el seguimiento y valoración de la evolución de la coyuntura económica de Galicia. Siguiendo la literatura existente se trata de un indicador sintético construido mediante la aplicación a un conjunto de series económicas observadas el modelo factorial dinámico, que permite reducir la dimensión de las series iniciales. De este modo se obtiene una interpretación más simple y compacta a través un conjunto reducido de factores comunes. Los resultados obtenidos permiten seguir la evolución conyuntural de la economía gallega de manera consistente con la evolución del PIB.

Palabras y frases clave: Incluya aquí las palabras y frases clave. Máximo de seis.

1. INTRODUCCIÓN

Una de las consecuencias de la reciente crisis económica viene dada por la necesidad de realizar análisis de los perfiles económicos que permitan superar los inevitables retrasos y pérdidas de información que se producen con el uso de medidas estándar de la actividad económica agregada basada en los macroagregados económicos de la Contabilidad Nacional. Como señalan Cuevas y Quilis (2012; 312) "se requiere información oportuna y precisa sobre las condiciones macroeconómicas generales" que evite el obligado retraso de las medidas macroeconómicas más estándares. Por ejemplo en el caso de Galicia, los datos de la Contabilidad son trimestrales y se publican con dos meses de decalaje. Un mes antes de que los datos se publiquen, ya estarían disponibles muchos de los indicadores mensuales que se utilizan para estimar el dato proporcionado por la Contabilidad.

Los indicadores a corto plazo para el seguimiento de la economía tienen su precedente a mediados del siglo XX con los indicadores cíclicos de Burns y Mitchell (1946) que fueron continuados más recientemente por Stock y Watson (1991). Desde entonces, son frecuentes los trabajos de este tipo realizados por instituciones tanto oficiales como académicas. Para el caso español destacan las propuestas de Camacho y Pérez-Quirós (2009), que es ampliada para el contexto europeo, y las de Cuevas y Quillis (2010) y Camacho y Doménech (2011).

En la línea de estos indicadores se enmarca el presente trabajo de diseño y construcción de un indicador sintético a escala regional. Su pretensión final reside en proporcionar una herramienta para el seguimiento en el corto plazo de la evolución de la coyuntura económica de una región mediante la

extracción de las señales comunes de un conjunto de variables de la economía (Cendejas et al. 2014). En concreto, para este trabajo se toma como referencia territorial la región de Galicia donde se consigue representar las series iniciales mediante un conjunto reducido de factores comunes, lo que permite ganar interpretabilidad en las series económicas.

Con el objetivo de obtener un indicador alternativo y a la vez complementario a la Contabilidad Regional, las series que se emplearán en este trabajo son series mensuales, por tanto el indicador coyuntural también será mensual.

El problema de trabajar con muchas series de tiempo es que los datos están correlacionados dentro de la misma serie y entre las series. Si se quieren modelizar estas series, lo habitual es que se necesite un número de parámetros grande, debido a la presencia de la correlación. Por tanto, si se logra reducir el número de series, se puede simplificar considerablemente la estructura del modelo, además descubrir los factores latentes en series económicas es muy importante para el seguimiento de la coyuntura económica.

2. DESCRIPCIÓN DE LA INFORMACIÓN Y DESARROLLO METODOLÓGICO

Las series iniciales de las que se parte se encuentran recogidas en la Tabla 1 y su representación gráfica en el Gráfico 1. Todas ellas cumplen el requisito de estar disponibles puntualmente y proporcionar una medida sintética de algún aspecto de la evolución de la coyuntura económica en línea con lo señalado por Cuevas y Quillis (2010). En general elegimos como periodo de inicio en la mayor parte de las series el mes de enero del año 1995, por ser el período a partir del cual está disponible el Índice de Produción Industrial en Galicia, indicador que se considera muy relacionado con la coyuntura económica de una región. De todos modos, en algunas de las series el período de inicio disponible en las fuentes de información es posterior. Por tanto, en este trabajo tenemos un problema de paneles de datos no equilibrados" (Cuevas y Quilis, 2010; 44).

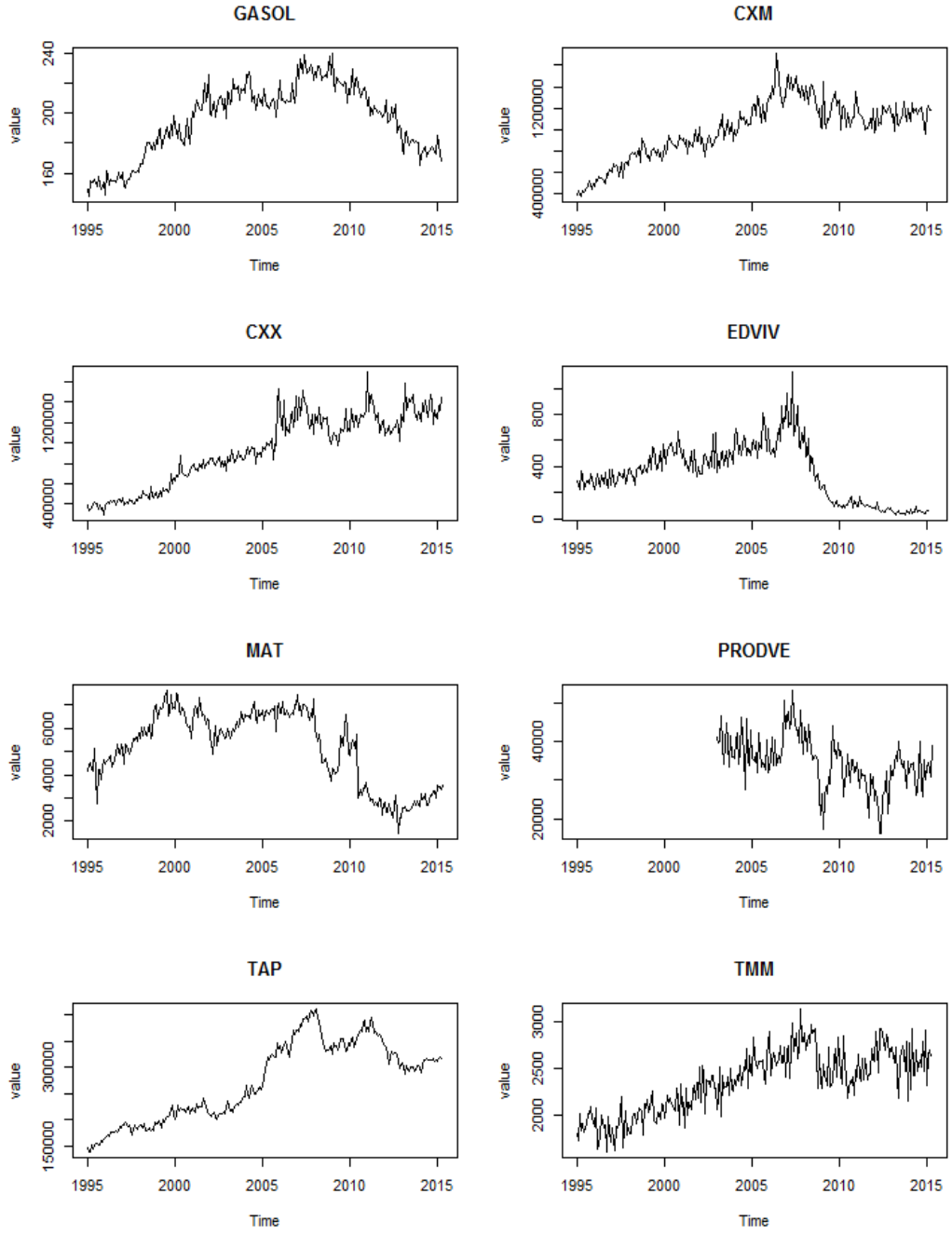
Además, todas las series están deflactadas y corregidas de estacionalidad y calendario. Con el objetivo de emplear en el análisis series estacionarias se tomaron las primeras diferencias de la serie en logaritmos. Estas diferencias se aproximan a la taxa de variación mensual.

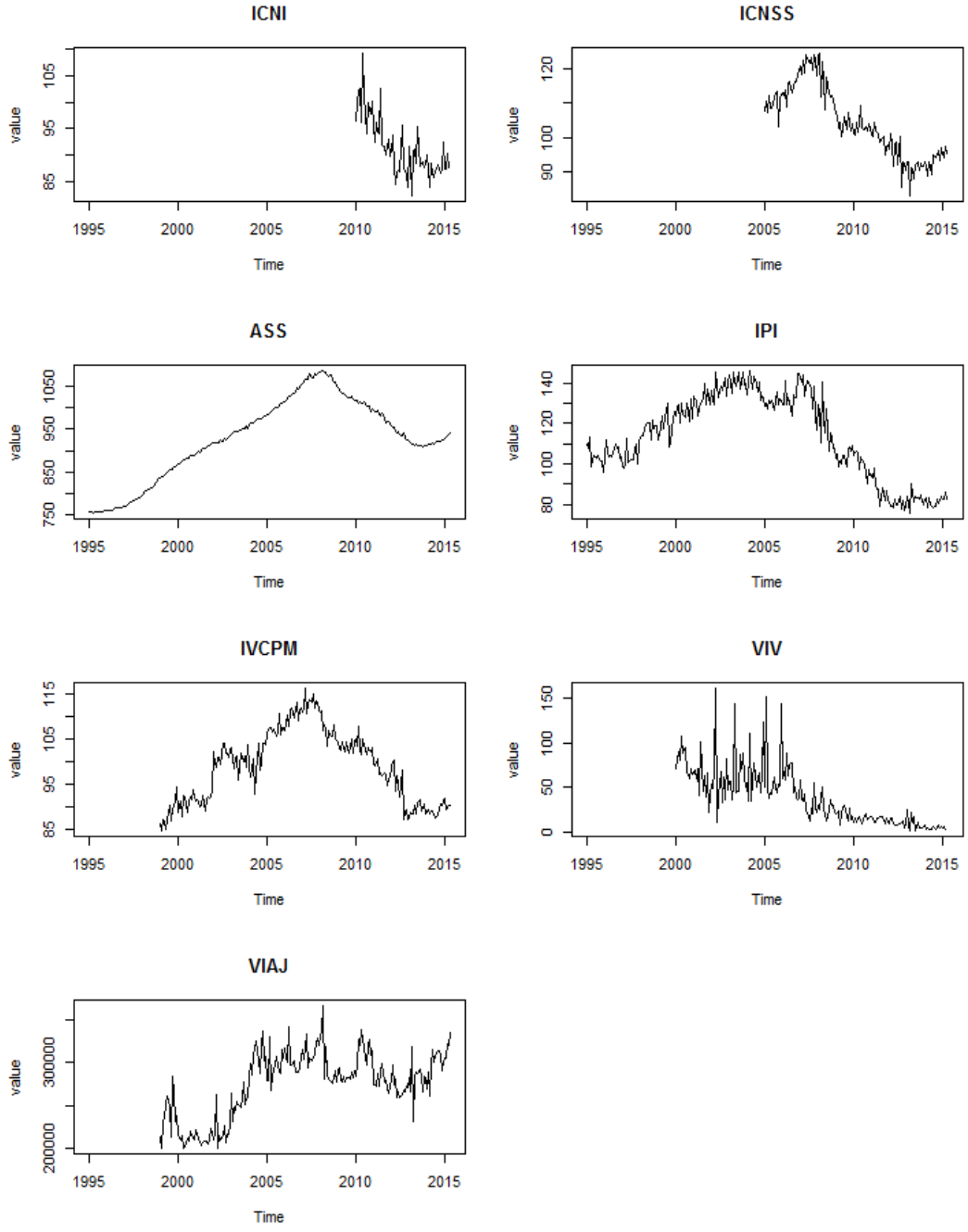
Tabla 1 : Series utilizadas en el análisis

Abrev.	**Variable**	**Fonte**	**Ano de inicio**	**Unidade**
GASOL	Consumo de gasolina y gasóleo	Ministerio de Industria	1995	Miles de toneladas
CXX	Exportacións de bienes deflactado por el Índice do valor unitario de exportación	AEAT/IGE	1995	Índice de valor
CXM	Importaciones de bienes deflactado por el Índice de valor unitario de exportación	AEAT/IGE	1995	Índice de valor
EDVIV	Superficie de las viviendas a crear de nueva planta	IGE	1995	Metros cuadrados
MAT	Matriculación de vehículos turismo	DGT	1995	Número

PRODVE	Producción de vehículos turismo	CITROEN	2003	Número
TAP	Transporte aéreo de pasajeros	MF	1995	Número
TMM	Transporte marítimo de mercancías (cargadas + descargadas)	Puertos del Estado	1995	Miles toneladas
VIV	Viviendas visadas	MF	2000	Número
ICNI	Índice de cifras de negocios en la industria deflactado por el Índice de Precios Industriales (IPRI)	INE	2002	Índice de valor deflactado
ASS	Afiliaciones a la Seguridad Social	MTAS	1995	Número
IPI	Índice de Produción Industrial deflactado por el IPRI (serie enlazada)	INE/IGE	1995	Índice de valor deflactado
IVCPM	Índice de ventas del comercio al por menor deflactado por el IPC (serie enlazada)	INE/IGE	1995	Índice de valor deflactado
VIAJ	Viajeros entrados en los hoteles	INE/IGE	1995	Número
ICNSS	Índice da cifra de negocios en el sector servicios deflactado por el IPC de los servicios	INE	2005	Índice de valor deflactado

Gráfico1. Representación de las series empleadas en el análisis


GASOL
value
160
200
240
1995
2000
2005
2010
2015
Time
CXM
value
400000
1200000
1995
2000
2005
2010
2015
Time
CXX
value
400000
1200000
1995
2000
2005
2010
2015
Time
EDVIV
value
0
400
800
1995
2000
2005
2010
2015
Time
MAT
value
2000
4000
6000
1995
2000
2005
2010
2015
Time
PRODVE
value
20000
40000
1995
2000
2005
2010
2015
Time
TAP
value
150000
300000
1995
2000
2005
2010
2015
Time
TMM
value
2000
2500
3000
1995
2000
2005
2010
2015
Time



En este trabajo se va a intentar explicar la variación temporal de un conjunto de *n* series observadas usando una combinación lineal de un conjunto de *m* paseos aleatorios desconocidos, donde *m*<<*n*. Por tanto, el modelo de análisis factorial dinámico (DFA) empleado tiene la siguiente estructura:

$$y_t = Zx_t + v_t \quad donde \quad v_t \sim MVN(0, R)$$

$$x_t = x_{t-1} + w_t \quad donde \quad w_t \sim MVN(0, Q)$$

La idea general es que las observaciones (***y***) se modelen como una combinación lineal de tendencias desconocidas (***x***) de dimensión *m* y de cargas factoriales (***Z***) de dimensión *n* x *m*. Por tanto, la estructura dinámica viene reflejada a través de los factores comunes ***x***. $\boldsymbol{w_t}$ hace referencia a la dinámica particular de cada serie. El elemento Z_{ij} representa el peso del *j*-ésimo factor sobre la *i*-ésima serie observada. La matriz ***Q*** se toma diagonal, con la diagonal igual a 1 y la matriz ***R*** se asume con valores desconocidos.

El problema que tienen las series que estamos empleando es que no tienen el mismo periodo temporal, es decir, no tienen el mismo momento de inicio y final. Así, por ejemplo, mientras el índice de producción industrial está disponible desde enero de 1995, el índice de cifra de negocios del sector servicios sólo está disponible desde enero del año 2005. Por tanto, para la estimación del modelo factorial dinámico se emplea un procedimiento iterativo que permite ir rellenando los huecos de las series.

Para la estimación del modelo se utiliza el paquete MARSS de R (Holmes et al, 2014). Las funciones del paquete estiman los parámetros del modelo (***Z***,***Q*** y ***R***) utilizando el algoritmo de ExpectationMaximization (EM) . Para la estimación de los factores se emplea el filtro de Kalman.

3. RESULTADOS

Una vez aplicado el modelo factorial dinámico se reducen todas las series iniciales a un factor y se obtienen las siguientes cargas factoriales en las series empleadas:

Tabla 2 : Cargas factoriales

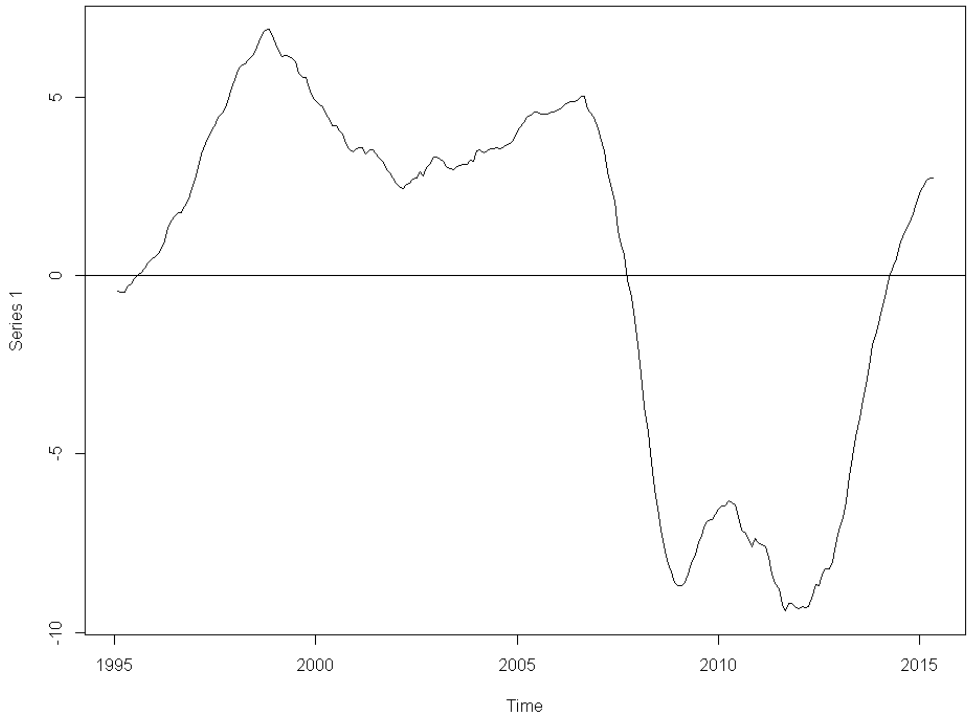
variables	cargas factoriais
GASOL	0,015
CXX	0,007
CXM	0,007
EDVIV	0,014
MAT	0,011
TAP	0,024
TMM	0,003
ASS	0,107
IPI	0,013
PRODVE	0,003
IVCPM	0,008
ICNSS	0,017
VIV	0,002
VIAJ	0,006

Las cargas factoriales estimadas para cada uno de los indicadores miden el grado de correlación dinámica entre el factor común y cada uno de los indicadores económicos. Como se puede observar en la Tabla 2, todos ellos son positivos . En este sentido se puede considerar a todos los indicadores utilizados en el modelo como procíclicos. Aunque todas las series contienen información sobre la evolución de la

coyuntura económica de Galicia, hay algunas diferencias en valor absoluto en las cargas factoriales. Así, estas estimaciones muestran que las afiliaciones a la Seguridad Social muestran el mayor valor de las cargas factoriales, seguido del Transporte aéreo de pasajeros, la cifra de negocios en el sector servicios, la edificación de viviendas y el índice de producción industrial.

Con estas cargas factoriales se obtiene el factor común que se presentan en el siguiente gráfico.

Gráfico 2: Indicador de coyuntura



Una vez realizada la estimación del factor común es importante hacer un análisis de la capacidad explicativa del factor sobre el crecimiento del PIB trimestral como indicador coyuntural de referencia comúnmente aceptado. La existencia de coherencia cíclica y la estabilidad de la relación permitirán aceptar el factor común como índice sintético válido.

REFERENCIAS

Burns, A. e Mitchell, W. (1946). Measuring business cycles. National Bureau of Economic Research.

Camacho, M. e Doménech, R. (2011), "MICA-BBVA: A Factor Model of Economic and Financial Indicators for Short-term GDP Forecasting", Journal of the Spanish Economic Association.

Camacho, M. e Perez-Quiros, G. (2009), Ñ-Sting: España Short Term Indicator of Growth, Banco de España, Working Paper Nº. 0912.

Cendejas, J.L., S. de Lucas, M.J. Delgado y I. Álvarez (2011), "Testing for structural breaks in factor loadings: An application to international business cycles", *Economic Modelling*, v. 28, No. 1-2, pp. 259-263.

Cendejas Bueno, J. L., de Lucas Santos, S., & Delgado Rodríguez, M. J. (2014). Evaluación del cambio estructural en los indicadores cíclicos. *Estadística española*, 56(183), 39-60.

Cuevas, Á., e Quilis, E. M. (2012). "A factor analysis for the Spanish economy". SERIEs, 3(3), 311-338.

Holmes, E.E., Ward, E. J. e M. D. Scheuerell (2024). Analysis of multivariate time series using the MARSS package. http://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf

Stock, J.H. y M.W. Watson (1991), "A probability model of the coincident economic indicators", en K. Lahiri y G.H. Moore (eds.), Leading economic indicators. New approaches and forecasting records, Cambridge University Press, Cambridge.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics, 20(2), 147-162.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Indicador Mensal de Actividade Económica da Região Autónoma dos Açores (IAE - Açores)

Autor 1[1]

[1] Isabel Cristina Monjardino - Serviço Regional de Estatística dos Açores

RESUMO

O SREA iniciou, em 2015, a divulgação do Indicador de Actividade Económica (IAE) regional, resultante do trabalho desenvolvido nos últimos anos, com o apoio técnico do INE de Portugal, no sentido de, à semelhança do que é feito pelo INE ao nível do território nacional, o SREA construir um IAE para a Região Autónoma dos Açores (RAA).

O objectivo deste projecto é colmatar uma lacuna existente nas estatísticas regionais, dotando a Região e os decisores da política económica regional de uma ferramenta essencial à análise económica conjuntural. Esta análise é dificultada quando o decisor da política económica se vê confrontado com dados que fornecem informações contraditórias acerca do estado actual da economia e o PIB, a medida mais representativa da actividade económica como um todo, se apresenta apenas com uma periodicidade anual e é calculado com bastante atraso em relação ao período a que respeita.

A necessidade de sintetizar o conjunto de informação económica de periodicidade infra-anual, conduz à construção de um indicador compósito que reflicta a evolução da economia no curto-prazo.

O IAE - Açores é um indicador compósito coincidente, construído para acompanhar a evolução, no curto prazo, da economia regional, a partir de séries de referência escolhidas como *proxy* da actividade económica regional.

Na sua construção foi seguida a chamada metodologia tradicional, ou seja, uma abordagem na óptica do VAB, que assenta na selecção das séries mensais/trimestrais mais representativas das actividades que compõem o VAB e na sua agregação, ponderada pelo peso do VAB de cada uma dessas actividades no VAB total da Região.

Palavras e frases chave: Indicadores compósitos coincidentes, Conjuntura económica, Ciclos económicos

1. INTRODUÇÃO

Questão de base: acompanhamento da evolução da economia no curto prazo e identificação dos ciclos económicos e dos pontos de viragem da economia.

Conceito de ciclo económico – flutuações recorrentes da actividade económica agregada, que se traduzem em movimentos generalizados de: recessão, contracção, recuperação, expansão e, novamente recessão. A sua duração varia entre 1 a 12 anos.

Métodos de medição e acompanhamento dos ciclos:

- de tipo estatístico (metodologias de decomposição ciclo-tendência, modelização ex: ARIMA, etc.)
- de tipo estatístico-económico (nos quais se inclui a abordagem dos indicadores)

Breve historial da abordagem dos indicadores:

Arthur Burns e Wesley Mitchel (1938) – primeiro estudo sobre a utilização de indicadores no seguimento da conjuntura económica

Geoffrey Moore (1950) e Julius Shiskin (1961)– desenvolvimento dos primeiros indicadores compósitos

1968 – os EUA publicam, pela 1ª vez, índices compostos de indicadores cíclicos

1987 – OCDE - *Composite Leading Indicator* (CLI)

1997 - Comissão Europeia - *Economic Sentiment Indicator* (ESIN)

2000- Comissão Europeia – *Business Climate Indicator* (BCI)

Em Portugal:

Banco de Portugal (BdP):

Indicador Coincidente de Actividade Económica (Dias,1993)

Novo Indicador Coincidente de Actividade Económica (Azevedo, Koopman e Rua, 2003)

Instituto Nacional de Estatística (INE):

Indicador de Actividade Económica

Indicador de Clima Económico

Os indicadores cíclicos podem ser simples ou compósitos (resultantes da agregação de diversos indicadores simples) e estes últimos podem ser avançados (se antecipam a evolução do ciclo), coincidentes (se acompanham o ciclo) ou atrasados (se a sua evolução é posterior á do ciclo).

Principais objectivos da utilização de indicadores compósitos:

- Como indicadores sintéticos da evolução geral da actividade económica, no curto prazo;
- Como ferramenta para identificar pontos de viragem do ciclo económico;
- Como mecanismo de previsão a curto prazo do crescimento económico.

Metodologias de construção dos indicadores compósitos:

- Stock-Watson (BdP)
- Componentes Principais e Análise Factorial Clássica (INE)
- Metodologia Derivada
- Análise Factorial Dinâmica
- Metodologia tradicional – abordagem pelo VAB (SREA)

O IAE– Açores é um indicador compósito, coincidente, construído a partir da metodologia tradicional, que pretende acompanhar, mensalmente, a evolução do "estado geral" da economia da Região Autónoma dos Açores.

2. APONTAMENTO METODOLÓGICO E PRINCIPAIS RESULTADOS

Na obtenção do IAE- Açores foi seguida a abordagem na óptica do VAB, que assenta na selecção das séries mensais/trimestrais mais representativas das actividades que compõem o VAB e na sua agregação, ponderada pelo peso do VAB de cada uma dessas actividades no VAB total da Região. Esta metodologia é a mais simples e foi considerada a mais indicada para a RAA.

No cálculo do IAE – Açores, de acordo com a metodologia adoptada, as etapas seguidas foram:

i) **Selecção** - constituição das séries de variáveis económicas infra-anuais (mensais ou trimestrais), seleccionadas de acordo com critérios de maior correlação com o VAB da actividade a que respeitam, existência de séries longas, disponibilidade em tempo útil e significado económico. Para o cálculo do IAE – Açores foram utilizadas as seguintes séries: "Leite entregue nas fábricas", "Gado abatido", Pesca descarregada", "Produção de energia", "Produtos lácteos produzidos", "Consumo de Energia na Indústria", "Venda de cimento", "Nº de empregados na Construção Civil", "Passageiros desembarcados via aérea", "Dormidas nos estabelecimentos hoteleiros e TER", "Empréstimos bancários concedidos", "Nº de prédios transaccionados", " Levantamentos Multibanco";

ii) **Filtragem** - eliminação todas as actividades que não apresentam flutuação cíclica (por exemplo, a Administração Pública), periodicidade mensal ou trimestral;

iii) **Tratamento das séries** – Detecção de *outliers*, estimação de *missing values*, ajustamento da sazonalidade. Normalização ou cálculo das taxas de variação homólogas, para tornar comparáveis todas as séries;

iv) **Agregação** – por média ponderada, tomando como ponderador de cada variável o peso da actividade que ela representa, no VAB total do último ano disponível;

v) **Apresentação** - os dados apresentam-se ajustados da sazonalidade, calibrados pela variação do PIB e alisados pelo método de médias móveis de 3 meses.

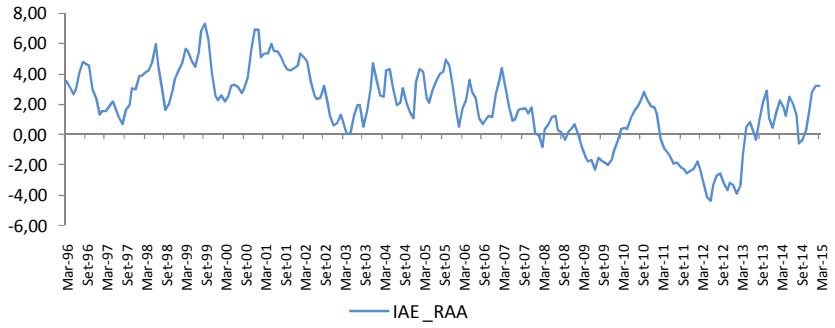
Principais resultados

O IAE é um indicador compósito construído para acompanhar a evolução, no curto prazo, da economia regional, a partir de séries de referência escolhidas como proxy da actividade económica regional.

Assim, a partir de um conjunto de 13 séries infra-anuais (mensais ou trimestrais) seleccionadas com base nos critérios referidos no Apontamento Metodológico, construiu-se uma série longa de IAE, desde 1996 a 2015.

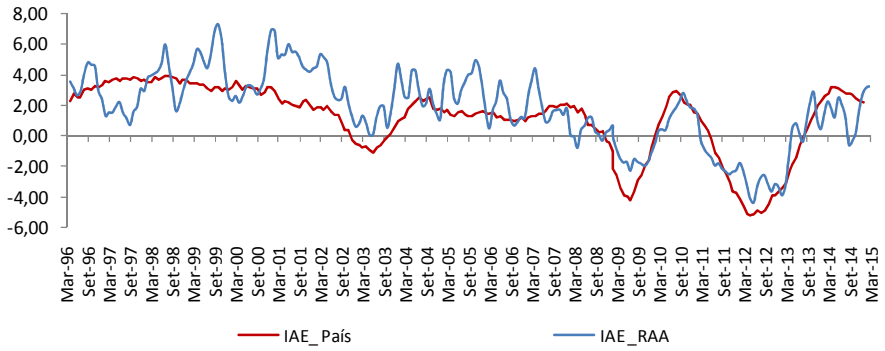
A partir do Gráfico 1 pode-se observar a evolução, ao longo desse período de tempo, do valor mensal do IAE da RAA.

Gráfico 1 - Evolução do IAE - Açores
1996 -2015



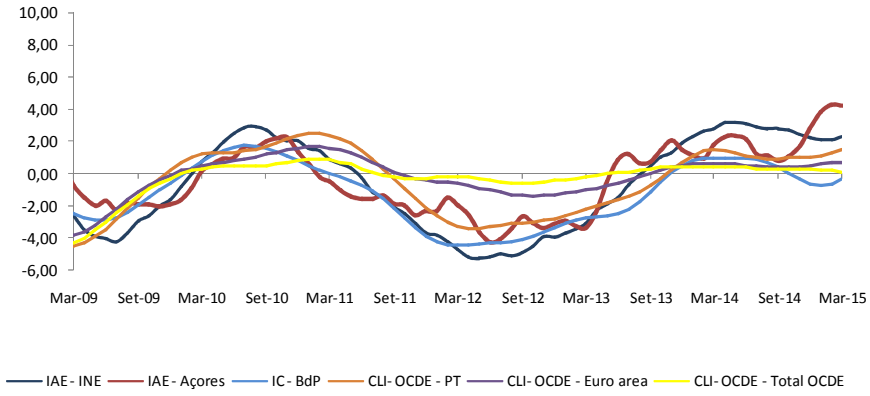
Comparando com a evolução do IAE nacional, verifica-se uma evolução consistente entre os dois indicadores (nacional e regional), o que faz sentido dada a estreita ligação e relação de dependência da economia da RAA face à nacional. A maior volatilidade do IAE Açores tem a ver com as diferentes metodologias utilizadas e ainda com a dimensão mais reduzida da economia regional. (Gráfico 2)

Gráfico 2 - IAE -Açores vs IAE nacional



No Gráfico 3 apresentam-se, para além dos IAE nacional e regional, o Indicador Coincidente do Banco de Portugal, calculado para o País e o CLI (*Composite Leading Indicator*) da OCDE, para Portugal, a zona Euro e o total do espaço da OCDE. É visível a consistência da evolução de todos estes indicadores, nos últimos 6 anos.

Gráfico 3 - IAE -Açores vs IAE nacionais e internacionais
Março 2009 - Março 2015



3. CONCLUSÕES

Aspectos positivos

→ Os indicadores compósitos sintetizam realidades complexas, pelo seu carácter multidimensional, neste caso, o "estado geral da economia";

→ São mais fáceis de interpretar, facilitando a análise dos decisores;

→ Facilitam o *benchmarking* ;

→ São uma forma de obter informação sobre o sentido da evolução cíclica da economia, enquanto o PIB não está disponível;

→ O IAE – Açores, em particular, sendo um projecto pioneiro na RAA, constitui uma mais-valia para a informação estatística económica infra-anual produzida pelo SREA, organizando-a e sintetizando-a num indicador único e para a actividade deste Serviço, enquanto autoridade estatística regional, elevando a produção estatística regional a um novo patamar de complexidade .

Fragilidades e limitações

→ Os indicadores compósitos podem dar lugar a conclusões simplistas;

→ Podem produzir sinais confusos se forem deficientemente construídos ou mal interpretados;

→ A selecção das variáveis e a atribuição dos ponderadores podem ser alvo de questionamento;

→ Um IAE isolado não é suficiente para avaliar devidamente a evolução da conjuntura económica; é necessário complementá-lo com outros indicadores de conjuntura.

REFERÊNCIAS

BCE (2001), " O conteúdo informativo dos indicadores compósitos do ciclo económico da área do euro", BCE, Boletim Mensal, Novembro de 2001.

RUA, António (2002), "Indicadores compósitos para a actividade económica na área do euro", Banco de Portugal, Boletim económico, Setembro de 2002

OLIVEIRA, Pedro Miguel (2002), "Indicadores de actividade e ciclos económicos: teoria e evidência para Portugal", tese de Mestrado em Economia, UTL - ISEG

DIAS, Francisco C. (2003), " O indicador coincidente para a economia portuguesa: uma avaliação histórica dos seus 10 anos de existência", Banco de Portugal, Boletim Económico, Setembro de 2003.

RUA, António (2004), "Um novo indicador coincidente para a economia portuguesa", Banco de Portugal, Boletim Económico, Junho de 2004.

MELRO, Francisco (2005), "Um indicador avançado trimestral para o indicador coincidente do PIB", INE, Gabinete de Estudos

OCDE (2008), "OECD system of composite leading indicators".

De la TIGERA, Victor Alfredo Bustos (2009), "Indicadores Sintéticos para seguir la evolucion en el tiempo de fenómenos multidimensionales: una proposta metodológica", INEGI, Boletin de los Sistemas Nacionales Estadistico y de Informacion Geográfica, Vol.2, Núm3, Septiembre-Deciembre 2009

CARDOSO, Fátima e RUA, António (2011), "As contas nacionais trimestrais em tempo real: uma análise das revisões da última década", Banco de Portugal, Boletim económico, Outono de 2011

ESTEVES, P.S. e RUA, António (2012), "Previsões económicas de curto prazo para Portugal: uma síntese metodológica", Banco de Portugal, Boletim económico, Outono de 2012

HEATH, Jonathan (2012), "Lo que indican los indicadores: como utilizar la informacion estadística para entender la realidad económica de México", INEGI

RUA, António (2015), "Indicadores coincidentes mensais do Banco de Portugal revisitados", Banco de Portugal, Revista de Estudos Económicos, vol I, nº1

Estatística, Turismo e sustentabilidade – um estudo de caso aplicado a uma região portuguesa

Ermelinda Oliveira (i), José R. Pires Manso (ii)

(i) Instituto Politécnico da Guarda / ESTG, Departamento de Gestão, UDI, Guarda, e*rmelindaol@ipg.pt*,
(ii) Universidade da Beira Interior, Departamento de Gestão e Economia, NECE, Covilhã, *pmanso@ubi.pt*

RESUMO

O objetivo desta comunicação/paper é apresentar um trabalho científico que envolve a realização, colheita e tratamento dos dados de um inquérito com recurso a métodos multivariados de análise aplicado ao sector turístico em que os objetivos são apreciar a perceção dos habitantes da região sobre os diversos impactos do turismo, questionar a existência de fatores de desenvolvimento do turismo no território de estudo e apreciar as atitudes dos residentes face ao turismo como motor de desenvolvimento sustentável. O turismo é um sector de atividade económica cada vez mais importante para qualquer país e em particular para Portugal e as suas regiões e de entre elas a Serra da Estrela e municípios envolventes de alguma forma coincidente com a antiga região de turismo da Serra da Estrela. A sustentabilidade dos países e das regiões, em particular dos que sofrem o efeito da indústria turística, assume também cada vez mais importância pois é indispensável a preservação de recursos e de um ambiente saudável para os vindouros, tornado evidente depois da publicação do relatório Burdtland (1987) – "Our Common Future" cujos "3 principais pilares do desenvolvimento sustentável são o desenvolvimento económico, proteção ambiental e igualdade social", das Conferências do Rio (1992, 2012) com metas materializados na Agenda 21 ou agenda para o século XXI e reafirmados regularmente nos Objetivos do Desenvolvimento do Milénio (UNDP, 2015), nomeadamente "assegurar a sustentabilidade ambiental" e "desenvolver uma parceria global para o desenvolvimento" e também do Protocolo de Kyoto (1997) (um acordo internacional ligado ao UNFCCC-Convenção das Nações Unidas para as alterações climáticas, que compromete as partes através da fixação de objetivos internacionalmente aceites em termos de limites de emissões). Em termos metodológicos o artigo começa por propor um inquérito dirigido aos residentes na região das Beiras e Serra da Estrela sobre os impactos do turismo e sua importância para o desenvolvimento sustentável, seguido da recolha e tratamento dos resultados desse inquérito recorrendo a técnicas que vão das mais simples – contagens e frequências relativas – a outras mais exigentes como as técnicas multivariadas de análise, análise fatorial (FA) e estimação de um Modelo de Equações Estruturais (SEM), elementos importantes para a identificação de um modelo conceptual relacionado com a perceção dos impactos do turismo sobre a sustentabilidade da região. Os resultados mostram que a metodologia estatístico-econométrica seguida é correta para estudar este fenómeno e que o turismo é uma alavanca para o desenvolvimento sustentável das regiões, nomeadamente das regiões do interior de Portugal continental; que relativamente à região das Beiras e Serra da Estrela (BSE), a atividade turística ainda apresenta fragilidades que exigem de todos os intervenientes ações urgentes e eficazes; que os residentes reconhecem que as BSE apresentam realidades e dinâmicas distintas; que atendendo a essa realidade, é necessário unir esforços em prol da formulação de uma estratégia integrada para o seu desenvolvimento; que os desafios de desenvolvimento que a Região das BSE enfrentam exigem um espaço próprio de entendimento daquilo que são os seus recursos humanos, os seus valores culturais, os seus valores ambientais, os seus recursos endógenos, as suas relações exógenas sem pôr em risco a sua própria identidade; que face às conclusões apresentadas nesta investigação é um imperativo, garantir o desenvolvimento do turismo da Região enquadrado numa estratégia de desenvolvimento, como forma responsável de assegurar a sua sustentabilidade económica, sociocultural e ambiental.

Palavras-chave: Desenvolvimento, Sustentabilidade, Território, Turismo, Recursos, Residentes, Perceções, Impactos, Atitudes

1. Introdução

Turismo e desenvolvimento sustentável são temas de investigação cada vez mais importantes. Os trabalhos seminais sobre turismo e desenvolvimento económico são o de Lanza e Pigliaru (2000) e o de Balaguer e Cantavella-Jordà (2002) publicados no início do século. Desde então a literatura empírica acerca desta relação espalhou-se rapidamente. No geral são investigações que usam dados seccionais/"cross-country" ou cronológicas/ "time series" que mostram o efeito positivo do turismo sobre o desenvolvimento económico (Lanza et al., 2003; Eugenio-Martin et al., 2004; Dritsakis, 2004; Lee e Chang, 2008; Holzner, 2011, Massida e Mantana, 2013). Um cada vez maior número de economistas e outros investigadores sociais consideram o turismo como um dos principais 'drivers' do desenvolvimento. As receitas crescentes do turismo são fontes importantes de rendimento familiar, de colheita de impostos e de divisas externas (Sinclair, 1998; Lee e Chang, 2008). O turismo para o desenvolvimento sustentável tem muitas implicações importantes, sendo também considerado, desde há muito tempo, como um meio eficaz de redução dos níveis de pobreza (de Kadt, 1979), de geração de emprego (Lim, 1997) e de empoderamento (empowerment) das mulheres (Tucker e Boonabaana, 2011). O estudo de Castro-Nuño et al. (2013) juntou evidência que o turismo contribui para o crescimento económico e conclui que investigações futuras centradas em grupos específicos de países irão provavelmente ampliar as suas descobertas. Em concordância com este último ponto de vista Chou (2013) argumenta que como o crescimento das despesas em turismo pode ser 'country-specific' é necessário ter em atenção a natureza heterogénea dos países em investigação. A quota do turismo no PIB dos países pós-comunistas tem sido pequena comparada com à média geral, contudo, estes países têm vindo a observar um grande crescimento nas receitas do turismo ao longo dos últimos anos. Entre estes países ainda nao há evidência estável acerca do link entre turismo e desenvolvimento económico. Autores como Hall (1991; 1995), Richards (1996) e Ibrahimova (2014) e investigaram o processo que evoluiu do turismo social para o turismo orientado pelo mercado nas ex economias de direção central concentrando as suas análises sobre a natureza da descrição mas sem fazerem uma avaliação crítica do comportamento do crescimento do turismo nesses países. Nas suas análises empíricas investigaram a relação de longo prazo entre turismo e desenvolvimento económico usando uma abordagem multivariada com as receitas reias per capita e as taxas de crescimento do PIB real com base num painel de 23 países ao longo do período 1991-2012. Turismo é uma grande oportunidade de negócio também para as montanhas e áreas rurais, regiões muito vulneráveis às alterações e à degradação causada pelo turismo (ICIMOD, 2010). Especificidades próprias destas regiões como más acessibilidades, fragilidade, vulnerabilidade e diversidade são barreiras ao seu desenvolvimento (Jodhaet al., 1992; Berkes e Gardner, 1997); o turismo pode ajudar a converter essas especificidades em oportunidades económicas e activos (Nepal e Chipeniuk, 2005). O turismo, como outras atividades de desenvolvimento traz alterações ambientais positivas e negativas (Buckley, 2009). O turismo contribui positivamente para o desenvolvimento local. No entanto, Syamlal (2008) acrescenta que o crescimento rápido causa frequentemente degradação ambiental e perda da identidade cultural tradicional. As áreas turísticas sofrem alterações socio-culturais logo que as culturas locais e dos visitantes se encontram frente uma à outra (Mathieson e Wall, 1982; Sharpley, 1994). O desenvolvimento turístico influencia negativamente o ambiente dos destinos (Gossling, 2002; Ramdas e Badaruddin, 2014). Assim, o crescimento e as atividades económicas dependem da qualidade e características do ambiente tendendo a afetar o ambiente, a capacidade de absorção de grande quantidade de gente e da intensidade de utilização do local (Smith, 1989; WTO, 1990; Gartner, 1996, respectivamente).

O resto do paper está estruturado da seguinte forma. A próxima secção apresenta uma rápida revisão da literatura acerca da relação turismo e crescimento sustentável, a secção 3 apresenta a metodologia e os dados da amostragem, a secção 4 apresenta os resultados empíricos e a sua interpretação e a secção 5 conclui o artigo.

2. Revisão da literatura

Shan e Wilson (2001) formulou uma hipótese conhecida como "*tourism led growth* (TLG)" que conseguiu bastante atenção depois de 2001. Foram feitos estudos "cross-country" por Pablo-Romero e Molina (2013) de forma a eliminar os efeitos dos ciclos económicos e as possíveis alterações estruturais na relação entre a especialização turística e crescimento económico mas os problemas de heterogeneidade e de endogeneidade misturados com possíveis correlações contemporâneas espúrias entre os "time-averaged data" tornaram-se barreiras difíceis de ultrapassar. Brau et al. (2007), usando uma amostra seccional de dados, testaram esta hipótese usando uma amostra de 143 países divididos em grupos: pequenos, membros da OCDE, produtores de petróleo e especializados ou não em turismo. O período temporal considerado foi 1980-2003. A correlação entre a média das receitas do turismo e a taxa média de crescimento do PIB real nos pequenos países é positiva e mais elevada apenas quando eles são especializados em turismo; a correlação também positiva entre o turismo especializado e crescimento

económico mantem-se mesmo quando se usa como variáveis de controlo o rendimento per capita inicial e o grau de abertura comercial do país. Singh (2008) testou a existência de ligação entre turismo e desenvolvimento económico e a contribuição da indústria entre 37 pequenas ilhas em desenvolvimento. O autor usou um modelo de regressão cross-secional para estimar a relação entre estas variáveis, inicialmente, e entre turismo e multiplicadores do rendimento para as economias insulares, mais tarde. Po e Huang (2008) testaram a existência de uma relação linear entre turismo e crescimento económico usando uma amostra de 88 países divididos em 3 grupos em função do peso ou cota das receitas turísticas no PIB. Os resultados juntaram evidência empírica a favor da existência de uma relação significativa e positiva. Lanza et al. (2003) estudaram a ligação entre turismo e crescimento económico usando um painel de dados relativo a 13 nações da OCDE ao longo do período 1977–1992 e concluíram que o turismo tem um efeito mais forte nos países menos industrializados do que nos países industrializados onde o turismo internacional é um luxo. Eugenio-Martín et al. (2004), usaram uma amostra referente a países da América Latina para investigar a relação entre turismo e crescimento económico para o período 1985-1998. Para isso usaram o método Islam (1995) e controlaram a endogeneidade das variáveis explicativas. Os resultados mostraram que a chegada de turistas tem impacto positivo sobre o crescimento do PIB real per capita apenas nos países com rendimentos baixos e médios. Holzner (2011) e outros juntam mais evidência acerca do link entre turismo e crescimento económico num painel de países. Fayissa et al. (2008) juntam evidência de impacto positivo sobre o crescimento económico em 42 países sub-Saharianos usando o estimador de Arellano e Bond (1991) com autocorrelação. Este é um dos papers que foca o problema da heterogeneidade num painel de países. Por sua vez Pratt (2015) explora o nexo turismo-crescimento em pequenas ilhas em desenvolvimento, recorrendo a um modelo de Equilíbrio Geral para estimar os impactos macroeconómicos e os impactos económicos do turismo nestas nações. Estudos semelhantes de Lee e Chang (2008), Proenca e Soukiazis (2008), Narayan et al. (2010), Dritsakis (2012) e Apergis e Payne (2012) reportam evidência de efeito positivo do turismo no crescimento economico nos países não-OCDE, do sul da Europa, ilhas do Pacifico, países Mediterrânicos e das Caraíbas, respectivamente. Payne e Mervar (2010) para a Croácia, Surugiu e Surugiu (2013) para a Roménia e Chou (2013) para 10 países da europa central e do leste examinaram a relação entre turismo e crescimento económico, mas geralmente com pequenas amostras de dados, omitindo os membros da ex-URSS, factos que conduziram a algum viés regional entre os resultados empíricos e a robustez dos resultados estimados (Salahodjaev e Safarova, 2015).

Sustentabilidade, turismo e desenvolvimento: A sustentabilidade é um conceito conhecido dos profissionais do turismo e dos investigadores, profissionais estes que forneceram um conjunto de propostas sobre os desafios lançados ao desenvolvimento sustentável e abriram um caminho de sucesso para a sustentabilidade destas áreas de negócio. A sustentabilidade tornou-se um conceito prático considerado benéfico para o progresso do turismo e o desenvolvimento. Diversos autores afirmam que levar a cabo ações ambientalmente amigas do ambiente ajuda as entidades ligadas ao turismo a criar as suas próprias imagens como operações que têm em conta a vida ambiental dos seres humanos (Dolnicar e Leisch, 2007, Dolnicar et al., 2010; Mihalicˇ, 2013, Weaver, 2013, Chiu et al., 2014). O ambiente natural tem sido usado como um recurso para o desenvolvimento do turismo (Maher, 2012). Os turistas são atraídos por "*breathtaking landscapes*" nas áreas protegidas como as dos Parques Naturais, com os seus glaciares e rios. Mas estes turistas algumas vezes não têm consciência dos estragos que podem criar no ambiente que visitam. Miller et al. (2009) reportaram que os turistas não alteram o seu comportamento face ao ambiente, incluindo nas atividades como passeios a pé, de automóvel e camping em ambientes sensíveis, forçando a natureza a levar muito tempo a recuperar desses danos. Torna-se necessário desenvolver práticas apropriadas que estejam de acordo com os conceitos de sustentabilidade antes que a chegada maciça de turistas às áreas protegidas possa causar mais danos irreparáveis à natureza. Estruturas pragmáticas para o turismo sustentável são fundamentais, mas a inconsistência na definição do conceito de sustentabilidade é também apontada como uma limitação ao seu desenvolvimento (Johnston e Tyrrell, 2005, e Ko, 2005). Font (2001) e Font e Harris (2004) notam que vários programas de certificação da indústria foram desenvolvidos mas que deixaram os consumidores muito confusos. À falta de normas e critérios de sustentabilidade universal aceitáveis por todas as partes interessadas na questão do turismo de sustentabilidade, foi desenvolvida uma prática verde (*'green-wash')*, uma preocupação estratégia de marketing que fornece sugestões sustentáveis em termos ambientais como forma de aumentar as vendas turísticas sem implementar medidas de sustentabilidade (Font, 2001, e Sloan et al., 2012). Como resultado, os pesquisadores convencionais têm vindo a ser pressionados para insistirem no significado da Ética (Holden, 2009) e do Código de Conduta (Twynam e Johnston, 2002) na implementação de políticas e ações sustentáveis. É urgente construir um modelo sustentável operacional amplamente aceite pelos intervenientes do turismo, e adotar pontos de vista coletivos sobre a sustentabilidade das partes interessadas parece ser agora um tema candente na investigação em turismo (Chen, Joseph S., 2015). Enquanto as investigações iniciais sobre o turismo sustentável se facm nas atitudes dos residentes locais

(Burns e Sancho, 2003), os mais recentes (Imran et al., 2014 e Waligo et al., 2013) colocam-no em planos integrados com uma rede e uma parceria entre todos os intervenientes no processo de desenvolvimento de políticas de turismo sustentável. Enquanto Ballantyne et al. (2009) defendem que a mobilização de turistas é vital para o progresso da gestão sustentável, Vernon et al. (2005) defendem uma abordagem conjunta para desenvolver políticas sustentáveis envolvendo residentes, governos e empresas de turismo; por sua vez Buckely (2012) afirma que a regulação é o principal motor para implementar a sustentabilidade, a que Yasarata et al. (2010) juntam o planeamento político como a chave para promover a sustentabilidade do turismo. Como afirma J. S. Chen, (2015), as partes interessadas - turistas, moradores locais, autoridades públicas e empresários – são uma rede importante e integrada de tomada de decisões para o desenvolvimento sustentável. Impactos negativos do turismo em áreas protegidas e noutras zonas rurais têm sido referidas por diversos autores como Cohen (1978); Stephenson (1993); Croall (1995); Atik et al. (2009); Farooquee et al. (2008), Barros et al. (2013), Malik et al. (2011), Zabortseva e Yevstropieva (2009) e Rashid e Romshoo (2012). A erosão, sujidade, geração de resíduos e contaminação de água estão entre esses impactos negativos (Kuniyal et al., 1998; Pickering e Buckley, 2003; Stephenson, 1993); a recreação ao ar livre tem também impacto negativo sobre o solo, a flora, a fauna e qualidade da água (Monz et al., 2010). A cultura ambiental pode ser danificada; as relações socioculturais e alterações do comportamento também podem modificar o padrão de utilização dos recursos levando à dilapidação e destruição de recursos e do ambiente, outros impactos negativos a partir dos quais é geralmente difícil recuperar (O'Riordan e Turner, 1983; Wenjun, 2004). É necessário regular e preservar o turismo e o ambiente para as gerações vindouras (WCED, 1987; Bramwell e Lane, 1993), encontrar um equilíbrio que preserve também a defesa dos negócios turísticos e a defesa da qualidade de vida dos residentes locais. A sustentabilidade, as atividades turísticas e o ambiente desenvolveram-se muito ao longo dos últimos anos (Clarke, 1997; Swarbrooke, 1999; Saarinen, 2006; Grabara e Bajdora, 2013; Huiqin e Linchun, 2011; Hunter, 1997; Simpson, 2001). O desenvolvimento sustentável do turismo promove simultaneamente o bem-estar das comunidades locais e as necessidades do turismo (Hunter, 1997; Ko, 2001). TCC é um conceito relacionado com o turismo sustentável (Tribe et al., 2000) que mede o número máximo de pessoas que podem visitar um destino turístico num dado momento, sem prejuízos físicos, económicos e socioculturais e a satisfação do interesse dos visitantes (WTO, 1981). Este conceito garante que os dois podem co-existir e que podem ser uma estrutura útil para a análise dos impactos e dos limites ao desenvolvimento (Butler, 1996). TCC. O principal objetivo do turismo sustentável é estabelecer o TCC de um destino (Howie, 2003). Foi Wagar (1964) que desenvolveu formalmente o conceito de TCC. Contudo, o conceito foi evoluindo para destinos turísticos individuais em todo o mundo (Manning et al., 2002; McCool e Lime, 2001; Nghi et al., 2007; Barancok e Barancokova, 2008; Sayan e Atik, 2011; Lone, Lone, e Malik, 2013). TCC é considerado um instrumento apropriado para a gestão, uma vez que ajuda a preservar os recursos (Queiroz et al, 2014), sendo a avaliação do TCC uma das melhores técnicas para o planeamento e gestão do turismo e recreação combinado com outras ferramentas (Zacarias et al., 2011; Queiroz et al., 2014). TP é outro conceito interessante nesta temática. O conjunto das possibilidades que os ambientes naturais e sociais colocam à disposição das atividades turísticas constitui o conceito TP (Glavan, 2006), a condição base para o desenvolvimento (Hall e Page, 2004). Há também quem o defina como a soma de recursos naturais e humanos (Ielenicz e Comanescu, 2006). Diversas abordagens metodológicas foram adotadas para separar diferentes regiões TP (Al Mamun e Mitra, 2012; Iatu e Bulai, 2011; Malik e Bhat, 2013; Nestoroska, 2012). O setor do turismo tem tido um crescimento quase descontrolado orientado pela procura (East et al., 1998). O turismo tem influência no desenvolvimento de uma região nos seus aspetos económico, social, cultural, e ambiental (Arrow et al., 1993). Contudo, estes impactos são difíceis de destrinçar e de analisar (Briassoulis, 2002). Entre as causas da degradação ambiental temos o crescimento populacional, a urbanização, o turismo, por vezes a exploração comercial das florestas, e o desenvolvimento económico (Rawat e Sharma, 1997). As fragilidades e vulnerabilidades dos ambientes das montanhas e de outras zonas rurais tornam-nas particularmente sensíveis a estas perturbações (Berkes e Gardner, 1997). O turismo de montanha e rural pode criar oportunidades alternativas de negócio, diversificação das economias, favorecer o crescimento populacional, e ajudar a minimizar os problemas de pobreza da região ou destino (Singh, 1989; Sinclair e Ham, 2000). O impacto do turismo ambiental de montanha tem sido referido em diversos artigos (Godde et al., 2000; Kreutzmann et al., 2009; Kruk e Banskota, 2007; Nyaupane e Chhetri, 2009). O turismo não regulado favorece alterações ad hoc ao uso da terra (Bhat et al., 2007; Malik et al., 2013), a geração de resíduos sólidos, deflorestação (Malik et al., 2011), delapidação da qualidade da água (Rashid e Romshoo, 2012) e problemas ambientais. A crescente atividade turística e o crescimento da população resultaram em mais resíduos sólidos gerados em várias partes do mundo (Khan, 2014). A elevada concentração de poluentes do ar nos meses de verão está relacionada com os picos da atividade turística (Jehangir et al., 2011). A degradação da água é também atribuída a turismo não planeado em termos de assoreamento, sedimentação e poluição dos lagos (Fazal e Amin, 2012 Mushtaq et al., 2013). A

quantidade crescente de lixos, poluição e problemas ambientais deixada nos itinerários de *trekking* e nos *campings* têm impacto nas pessoas e na vida selvagem simultaneamente (Geneletti e Dawa, 2009). Mesmo a subida da temperatura nalguns locais turísticos é considerada uma seria ameaça à sustentabilidade do turismo de inverno (Dar et al., 2013).

3. Metodologia e dados

Nesta investigação vamos usar a metodologia dos sistemas de equações estruturais (SEM), uma poderosa e geral técnica multivariada de análise que inclui como casos especiais vários outros métodos de análise. Entre as aplicações da modelação de equações estruturais podemos incluir: (i) a modelação causal, ou *path analysis*, que testa a existência de relações causais entre as variáveis e a linearidade dos modelos causais (os modelos causais podem envolver quer variáveis *manifest*, quer variáveis latentes quer ambas); (ii) análise fatorial confirmatória, uma extensão da análise factorial na qual são testadas hipóteses específicas acerca da estrutura das cargas e inter-relações fatoriais; (iii) análise fatorial de segunda ordem, uma variante da análise fatorial em que a matriz das correlações dos fatores comuns é ela própria um fator analisado de forma a fornecer fatores de segunda ordem; (iv) modelos de regressão, uma extensão da análise regressão linear em que os pesos da regressão podem ser restringidos de forma a serem iguais uns aos outros ou iguais a valores numéricos específicos; (v) modelos de estrutura covariante que testa a hipótese de a matriz da covariância ter uma forma particular; (vi) modelos com estrutura correlacionada que testa a hipótese de a matriz das correlações ter uma forma particular. Muitas formas diferentes de modelos caiem nestas categorias referidas, donde ser difícil de caracterizar a modelação estrutural. A maioria dos modelos de equações estruturais pode ser expressa através de *path diagrams*. A modelação estrutural permite análises muito complicadas. Esta ideia pode generalizar-se, em vários aspetos, a várias variáveis interrelacionadas por um grupo de equações lineares. As regras tornam-se mais complexas, os cálculos mais difíceis, mas a mensagem básica permanece a mesma: testar se as variáveis estão interrelacionadas através de um conjunto de relações lineares através do exame das variâncias e covariâncias das variáveis.

O método testa se um conjunto de variâncias e covariâncias se ajusta a uma estrutura específica. A modelação estrutural funciona da seguinte forma: (i) torna-se necessário indicar a forma como se acredita que as variáveis estão inter-relacionados, muitas vezes com o uso de um *path diagram;* (ii) depois aprecia-se, através de algumas regras internas complexas, quais são as implicações disso para as variâncias e covariâncias das variáveis; (iii) em seguida testa-se se as variâncias e covariâncias se ajustam ao modelo; (iv) depois apresentam-se os resultados dos testes estatísticos, e também as estimativas dos parâmetros e dos erros padrões para os coeficientes numéricos nas equações lineares, e (vi) na base desta informação, decide-se se o modelo se ajusta bem aos seus dados.

Os dados tratados foram colhidos através do recurso a um inquérito aos habitantes da região das Beiras e Serra da Estrela (BSE) com 385 elementos repartidos pelas três sub-regiões – Beira Interior Norte (170), Cova da Beira (143) e Serra da Estrela (71) – antigas nuts III, que compõem aquela em função do seu peso relativo na população. Aos dados assim recolhidos e depois de introduzidos no AMOS/SPSS aplicou-se a modelação das equações estruturais cujos resultados são os que se deixam na próxima secção. A escolha do universo não incidiu na população total residente mas apenas na população residente com idade igual ou superior a 15 anos de idade. Esta metodologia, também utilizada noutros estudos sobre esta temática, (Barros, 2011; Guerreiro et al, 2008), justifica-se pelo pressuposto de serem estes residentes que mais conhecimentos têm da região das Beiras e Serra da Estrela, do desenvolvimento da atividade turística na região e, por último, por serem considerados mais capazes de percecionarem os impactos (positivos e negativos) do desenvolvimento do turismo. Neste estudo, a determinação da amostra de residentes a inquirir seguiu o descrito pelo método de amostragem por quotas. Na aplicação da técnica de amostragem por quotas teve-se em conta três critérios: o concelho de residência, género e o grupo etário dos residentes. O primeiro critério aplicado foi o do concelho de residência. A mesma metodologia incidiu sobre todos os concelhos da região BSE, tendo resultado uma amostra de 384 inquéritos. A Tabela 1 apresenta o número de inquéritos aplicados em cada um dos sub-destinos turísticos do território desta região.

Sub-destinos turísticos	Inquéritos aplicados	
	N	%
Beira Interior Norte	170	44,27
Cova da Beira	143	37,24
Serra da Estrela	71	18,49

Tabela 1. Número de inquéritos aplicados por sub-destino turístico.

Nesta investigação elegeu-se o inquérito por questionário como instrumento de recolha dos dados. A administração do questionário final foi efetuada durante os meses de abril a agosto de 2014 e aplicados nas sedes de municípios de cada concelho. Os dados recolhidos através da implementação de um inquérito por questionário, junto da amostra final definida, foram sujeitos a um tratamento estatístico.

4. Resultados

Após a recolha dos dados, procedeu-se a uma Análise Fatorial Exploratória (AFE) dos valores tomados pelos indicadores em estudo. A análise dos índices de consistência interna obtidos conduziu à exclusão de alguns itens o que melhorou substancialmente a fiabilidade da escala. Aos 97 indicadores resultantes voltaram-se a aplicar as técnicas de Análise Fatorial. O valor de 0,912 para a medida KMO (Keiser-Meyer-Ohlin) e um valor de $p<0,001$ associado ao Teste de Bartlett, indicam que o modelo de análise aplicado se adequa à amostra em estudo (Reis, 2008). Para a extração dos eixos recorreu-se à Análise em Componentes Principais (ACP) e para a determinação do número de eixos a reter, utilizou-se o critério do *scree plot* de Cattell e a regra do valor próprio superior a 1. A aplicação dos testes levou à seleção de onze componentes que conjuntamente explicam 62,5% da variância total. Após rotação Varimax obteve-se a distribuição dos itens pelos fatores. De modo a facilitar a interpretação removeram-se as cargas fatoriais inferiores a 0,3 e os itens foram numerados (ordem numérica e alfabética) pelas onze componentes retidas.

A estrutura fatorial obtida evidencia que o Fator 1 reúne os itens relativos aos impactos positivos do turismo, este de âmbito económico, ambiental e sociocultural. Este fator agrupa 24 itens, dos quais 11 estão relacionados com os impactos económicos, 10 com os impactos socioculturais e 3 com os impactos ambientais. Numa análise por grupos, os itens relativos aos impactos económicos positivos apresentam maiores níveis de saturação, sendo o item *contribuído para gerar localmente mais receita fiscal* que apresenta o valor mais elevado (0,850). O Fator 2 concentra informação relativa à gestão do território, enquanto destino turístico, ao nível da questão ambiental, da preocupação da envolvência de todos os atores locais nas questões relativas ao desenvolvimento da atividade turística, da criação e divulgação de novos produtos, bem como, da preservação e divulgação do património natural, histórico e etnográfico. Os 15 itens agrupados neste fator apresentam níveis de saturação elevados, concretamente, entre 0,684 a 0,779. No Fator 3 encontram-se os itens relacionados com os impactos económicos, ambientais e socioculturais negativos do turismo. Nesta dimensão todos os onze indicadores apresentam saturações bastante elevadas, sendo que o item relativo à qualidade dos serviços prestados apresenta o maior nível de saturação (0,785) e o indicador relacionado com a insegurança e a criminalidade o menor nível de saturação (0,626). O Fator 4, que agrupa 6 itens, está relacionado com os recursos naturais, nomeadamente os recursos paisagísticos, termais, faunísticos, fluviais, parques protegidos e, por último, a qualidade do solo, ar e água. Neste fator o item recursos faunísticos é o que possui carga mais elevada (0,842). O Fator 5 concentra informação relativa às infraestruturas e aos serviços básicos de apoio ao turismo (acessibilidades para e dentro da região, serviços de limpeza e segurança, entre outras). Este fator agrupa 9 itens com níveis de saturação superiores a 0,5, sendo o item drenagem e tratamento de resíduos sólidos o que ostenta a saturação mais elevada. O Fator 6 concentra 5 itens relacionados com o património histórico-monumental, nomeadamente, as aldeias históricas, os museus, monumentos civis e religiosos. Neste fator o item mais relevante é referente às aldeias históricas. O Fator 7 concentra informação relativa aos atributos dos recursos humanos, nomeadamente à diversidade e qualificação da oferta de mão-de-obra e à oferta formativa local na área do turismo. Este fator agrupa 5 itens todos com níveis de saturação bastante elevados, sendo o item parcerias entre escolas e empresas do setor o que regista maior saturação. O Fator 8 engloba itens relacionados com os serviços turísticos, especificamente, oferta de alojamento, oferta termal, postos de informação turística, oferta de desportos de inverno e de postos de informação turística, agrupando 7 itens numerados com níveis de saturação entre 0,411 a 0,730. O Fator 9 concentra as questões relacionadas com o património etnográfico, especificamente com a gastronomia, festividades e eventos, tradições locais e a prática da atividade agrícola tradicional. O item com maior nível de carga é o referente às atividades agrícolas tradicionais com o valor 0,766. O Fator 10 reúne os itens relativos à interação-visitante e agrupa os primeiros quatro itens (A1 a A4), todos com saturações superiores a 0,5. O Fator 11 concentra informação relativa aos atributos do turismo como motor de desenvolvimento sustentável e reúne 5 indicadores. Da análise fatorial exploratória resultaram os 11 fatores, atrás referenciados, apresentando-se na Tabela 2 os respetivos itens agrupados por dimensões.

Dimensões	**Sigla**	**N.º indicadores**	**Alpha de Cronbach**
Escala completa		97	0,956
Interação residente-visitante	IRV	4	0,82

Impactos positivos do turismo	IPT	24	0,955
Impactos negativos do turismo	INT	11	0,91
Recursos humanos	RH	5	0,908
Serviços turísticos	ST	7	0,866
Infraestruturas e serviços de apoio ao turismo	IESAT	9	0,893
Recursos naturais	RN	6	0,923
Património etnográfico	PE	6	0,869
Património histórico-monumental	PHM	5	0,818
Gestão do destino	GD	15	0,951
Turismo como motor de desenvolvimento regional sustentável	TMDS	5	0,766

Tabela 2: Alpha de Cronbach para a totalidade dos indicadores em análise e respetivas subdimensões

Como medida de consistência interna, seguindo Greene (2012), utilizou-se o coeficiente Alpha de Cronbach, coeficiente que estima a confiabilidade de um questionário através da análise das respostas dadas, calculada através da correlação média das perguntas, esta calculada a partir da "variância dos itens individuais e da variância da soma dos itens de cada avaliador de todos os itens de um questionário que utilizem a mesma escala de medição. Neste caso, essa medida propiciou correlações satisfatórias que oscilam entre 0,766 (TMDS) e 0,956 (escala completa) e que se podem classificar entre boas/Good e Excelentes/Excelent (George e Mallery, 2003; Kline, 2000). A Tabela 2 descreve os valores desagregados do coeficiente.

Após a validação do modelo através da Análise Fatorial Exploratória, os dados obtidos foram submetidos a um modelo de equações estruturais para confirmar (ou não) as hipóteses de investigação previamente colocadas. Aos dados obtidos ajustou-se o modelo apresentado na Figura 3 recorrendo-se à análise de equações estruturais, mediante a utilização do *software* AMOS 5.0.

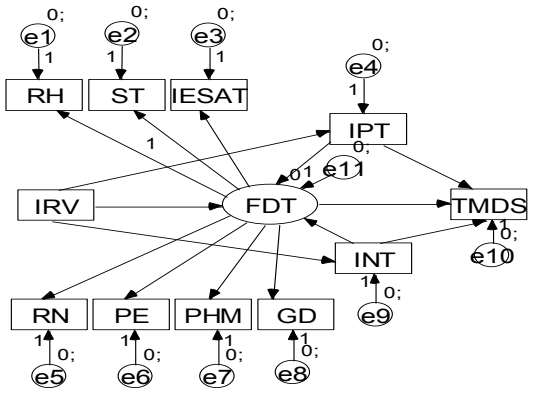


Figura. 3: Representação gráfica do Modelo de medida

A Tabela 3 apresenta a informação relativa às cargas fatoriais não estandardizadas e estandardizadas, o erro padrão associado a cada coeficiente e a razão crítica (razão entre a carga fatorial e o seu erro padrão, ou t observado); o valor da estimativa 1 é considerado significativo quando esta razão é maior que o módulo de ±1,96.

Hipótese			Estimativa1	S.E.	C.R.	p-value	Estimativa2
IPT	<---	IRV	,214	,034	6,242	***	,303
INT	<---	IRV	,138	,034	4,097	***	,205
FDT	<---	IRV	-,053	,028	-1,904	,057	-,110
FDT	<---	IPT	,323	,047	6,915	***	,469
FDT	<---	INT	-,033	,040	-,823	,410	-,046
RH	<---	FDT	1,000				,677
ST	<---	FDT	1,173	,100	11,655	***	,847
IESAT	<---	FDT	,734	,077	9,552	***	,569
RN	<---	FDT	,973	,106	9,223	***	,523
PE	<---	FDT	,767	,083	9,256	***	,559
PHM	<---	FDT	,738	,089	8,264	***	,496
GD	<---	FDT	,483	,076	6,340	***	,385

Hipótese			Estimativa1	S.E.	C.R.	p-value	Estimativa2
TMDS	<---	FDT	,261	,079	3,280	,001	,198
TMDS	<---	IPT	,021	,157	,132	,895	,023
TMDS	<---	INT	,011	,065	,163	,870	,011

Tabela 3: Estimativas de Máxima Verosimilhança relativas aos Coeficientes Estruturais para o modelo ajustado. Nota: Standardized Regression Weights: (Group number 1 - Default model) ***- Estimativa 1--- e estimativa 2---.

A Tabela 3 mostra que há uma relação estatisticamente significativa, uma vez que o p--value ou probabilidade de rejeitar a hipótese nula é inferior ao n.s. de 5% previamente fixado (Hair et al., 2010), entre:

- A interação residente-visitante e a perceção dos impactos positivos do turismo (IRV→IPT);
- A interação residente-visitante e a perceção dos impactos negativos do turismo (IRV→INT);
- A perceção dos impactos positivos do turismo e a satisfação com os fatores de desenvolvimento do turismo (IPT→FDT);
- A satisfação com os fatores de desenvolvimento do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável (FDT→TMDS).

Todavia, a Tabela 3 também evidencia a existência de relações estatisticamente não significativas (p-values > 0,05) entre:

- A interação residente-visitante e a satisfação com os fatores de desenvolvimento do turismo (IRV→FDT);
- A perceção dos impactos negativos do turismo e a satisfação com os fatores de desenvolvimento do turismo (INT→FDT);
- A perceção dos impactos positivos do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável (IPT→TMDS);
- A perceção dos impactos negativos do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável (INT→TMDS).

A tabela evidencia que a relação entre as variáveis «Fatores de Desenvolvimento do Turismo» e «Serviços Turísticos» (FDT→ST), com $\beta = 0{,}847$ e, ainda, a relação entre as variáveis «Fatores de Desenvolvimento do Turismo» e «Recursos Humanos» (FDT→RH) com $\beta = 0{,}677$, são as relações que apresentam estimativas estandardizadas mais elevadas, concluindo-se que existe uma grande associação entre esses pares de variáveis (Hair et al., 2010). A Tabela 4 contém as medidas da qualidade do ajustamento para o modelo obtido, indicando uma qualidade do ajustamento bastante satisfatória.

N	384
Qui-quad (χ^2)	30,730; gl=21
CFI	0,993
RMSEA	0,035
TLI	0,981
NFI	0,978

Tabela 4: Índices de Qualidade do Ajustamento; Notas: CFI - Comparative Fit Index; RMSEA - Root Mean Square Error of Approximation; TLI - Tucker- Lewis Index; NFI –Normalized Fit Index.

O valor de medida RMSEA – iniciais de Root Mean Square Error of Approximation - de 0.035, segundo Browne e Cudeck (1993), por ser menor que 0,05 indica um ajustamento muito bom (valores de RMSEA de 0,08 ou inferiores indicam bons ajustamentos do modelo e superiores a 0,10 indicam mau ajustamento). Os intervalos de variação dos valores de NFI - Normalized Fit Index e CFI - Comparative Fit Index são de zero a um, sendo que valores acima de 0,95 indicam ajustamentos bons. Neste caso, o Índice de Ajustamento Comparativo (CFI) de valor 0,993 e o Índice de Ajuste Normalizado (NFI) de valor 0,978 indicam ambos bons ajustamentos do modelo, como se verifica na tabela 4. O valor da estatística do Qui-quadrado (x2), que mede também a bondade de ajustamento global do modelo, obteve o valor de 30,730 para 21 graus de liberdade, que permite igualmente concluir que o ajuste é adequado. A Figura 4 ilustra o modelo estrutural com as hipóteses de investigação.

Após a apresentação dos resultados, segue-se a validação das hipóteses de investigação. A hipótese H1a indica que "Existe uma relação direta e positiva entre a interação residente-visitante e a perceção dos impactos positivos do turismo". O teste desta hipótese (H1a) destina-se a averiguar se há uma relação direta e positiva entre a interação residente-visitante e a perceção dos impactos positivos do turismo. Como $\beta = 0{,}303$ e $p < 0{,}001$, esta hipótese é suportada pela investigação (Greene, 2012 e Hair et al., 2010).

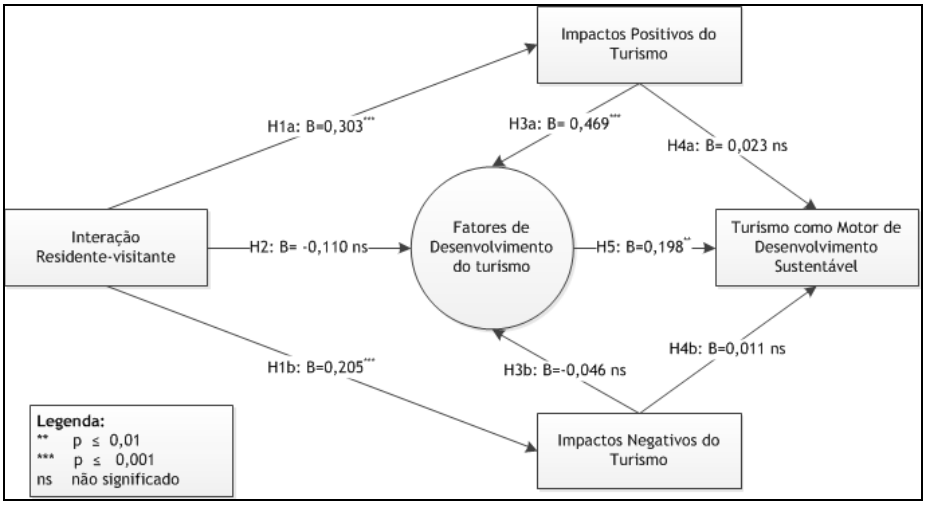


Figura 4: Modelo estrutural com as hipóteses

Por sua vez a hipótese H1b refere que "Existe uma relação direta e positiva entre a interação residente-visitante e a perceção dos impactos negativos do turismo". A hipótese H1b que prevê uma relação direta e positiva entre a interação residente-visitante e a perceção dos impactos negativos do turismo é suportada pela investigação, pois β = 0,205, (valor positivo) e p< 0,001(valor significativo). A hipótese H2 refere que "Existe uma relação direta e positiva entre a interação residente-visitante e a satisfação global com os fatores de desenvolvimento do turismo". A hipótese H2 que prevê uma relação direta e positiva entre a interação residente-visitante e a satisfação global com os fatores de desenvolvimento do turismo não é suportada pela investigação, pois β = - 0,110 (valor negativo) e p = 0,057 (valor não significativo). A hipótese H3a estipula que "Existe uma relação direta e positiva entre a perceção dos impactos positivos do turismo e a satisfação global dos fatores para o desenvolvimento do turismo". A hipótese H3a que prevê uma relação direta e positiva entre a perceção dos impactos positivos do turismo e a satisfação global com os fatores de desenvolvimento do turismo é confirmada pela investigação (β = 0,469 e p <0,001). A hipótese H3b estabelece que "Existe uma relação direta e negativa entre a perceção dos impactos negativos do turismo e a satisfação global dos fatores para o desenvolvimento do turismo". A hipótese H3b que prevê uma relação direta e negativa entre a perceção dos impactos negativos do turismo e a satisfação global com os fatores de desenvolvimento do turismo não é suportada nesta investigação (β = - 0,046; p = 0,410), embora a relação entre as variáveis seja negativa. A hipótese H4a estabelece que "Existe uma relação direta e positiva entre a perceção dos impactos positivos do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável". A hipótese H4a que prevê uma relação direta e positiva entre a perceção dos impactos positivos do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável não é suportada nesta investigação (β = 0,023 e p = 0,895). A hipótese H4b admite que "Existe uma relação direta e negativa entre a perceção dos impactos negativos do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável". A hipótese H4b que prevê uma relação direta e negativa entre a perceção dos impactos negativos do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável não é suportada nesta investigação (β = 0,023 e p = 0,870). A hipótese H5 estabelece que "Existe uma relação direta e positiva entre a satisfação global com os fatores de desenvolvimento do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável". A hipótese H5 que prevê uma relação direta e positiva entre a satisfação global com os fatores de desenvolvimento do turismo e a atitude face ao turismo como motor de desenvolvimento sustentável é suportada nesta investigação (β = 0,198 e p = 0,001).

5. Conclusões

O artigo aqui apresentado pretende estudar a questão do turismo como motor do desenvolvimento sustentável aplicado a uma região do interior de Portugal (BSE). Em termos metodológicos o artigo começa por propor um inquérito dirigido aos residentes na região das Beiras e Serra da Estrela sobre os impactos do turismo e sua importância para o desenvolvimento sustentável, seguido da recolha e tratamento dos resultados desse inquérito recorrendo a técnicas que vão das mais simples – contagens e frequências relativas – a outras mais exigentes como as técnicas multivariadas de análise, análise fatorial (FA) e estimação de um Modelo de Equações Estruturais (SEM), elementos importantes para a identificação de um modelo conceptual relacionado com a perceção dos impactos do turismo sobre a

sustentabilidade da região. Os resultados mostram que a metodologia estatístico-económétrica usada para estudar este fenómeno é correta e que o turismo é uma alavanca para o desenvolvimento sustentável das regiões, nomeadamente das regiões do interior de Portugal continental; que relativamente à região das Beiras e Serra da Estrela (BSE), a atividade turística ainda apresenta fragilidades que exigem de todos os intervenientes ações urgentes e eficazes; que os residentes reconhecem que as BSE apresentam realidades e dinâmicas distintas; que atendendo a essa realidade, é necessário unir esforços em prol da formulação de uma estratégia integrada para o seu desenvolvimento; que os desafios de desenvolvimento que a Região das BSE enfrentam exigem um espaço próprio de entendimento daquilo que são os seus recursos humanos, os seus valores culturais, os seus valores ambientais, os seus recursos endógenos, as suas relações exógenas sem pôr em risco a sua própria identidade; que face às conclusões apresentadas nesta investigação é um imperativo, garantir o desenvolvimento do turismo da Região enquadrado numa estratégia de desenvolvimento, como forma responsável de assegurar a sua sustentabilidade económica, sociocultural e ambiental.

Referências

Al Mamun, A., and Mitra, S. (2012). A methodology for assessing tourism potential: Case Study Murshidabad District, West Bengal, India. International Journal of Scientific and Research Publications, 2(9), 1–8.

Apergis, N., e Payne, J. E. (2012). Tourism and growth in the Caribbean—evidence from a panel error correction model. *Tourism Economics*, *18*(4), 449–456.

Arellano, M., e Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The review of economic studies*, *58*(2), 277-297.

Arrow et al., (1993) - Arrow, K., Solow, R., Portney, P., Learner, E., Radner, R., and Schuman, H. (1993). Report of the NOAA panel on contingent valuation. Federal Register, 58, 4602–4614.

Atik et al. (2009) - Atik, M., Sayan, S., and Karagüzel, O. (2009). Impact of recreational trampling on the Natural Vegetation in Termessos National Park, Antalya-Turkey. Journal of Agricultural Science, 15(3), 249–258.

Balaguer, J., & Cantavella-Jordà, M. (2002). Tourism as a long-run economic growth factor: The Spanish case. *Applied Economics*, *34*, 877–884.

Ballantyne et al., (2009), R. Ballantyne, J. Packer, K. Hughes. Tourists' support for conservation messages and sustainable management practices in wildlife tourism experiences, Tourism Manage., 30 (5) (2009), pp. 658–664

Barancok, P., e Barancokova, M. (2008). Evaluation of the tourist path carrying capacity in the Belianske Tatry Mts. Ekologia, 27(4), 401–420.

Barros et al. (2013) - Barros, A., Gonnet, J., and Pickering, C. (2013). Impacts of informal trails on vegetation and soils in the highest protected area in the Southern Hemisphere. Journal of Environmental Management, 127, 50–60.

Beck, N., e Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American political science review*, *89*(03), 634-647.

Berkes, F., e Gardner, J. (1997). Mountain environments and sustainability. In F. Berkes, and J. Gardner (Eds.), Sustainability of mountain environments in India and Canada (pp. 1–18). Winnipeg: Natural Resources Institute, University of Manitoba.

Bhat et al., (2007) - Bhat, M.S.,Malik, M.I., and Kuchay, N.A. (2007). Impact of tourism on land use/land cover — A case study of Lidder watershed of Kashmir Valley. In M.S.S. Rawat (Ed.), Management strategies for the Indian Himalaya: Development and conservation (pp. 61–71). Srinagar, Garhwal: Media House.

Bramwell, B., e Lane, B. (1993). Sustainable tourism: An evolving global approach. Journal of Sustainable Tourism, 1(1), 1–5.

Brau et al. (2007) - Brau, R., Lanza, A., & Pigliaru, F. (2007). How fast are small tourism countries growing? Evidence from the data for 1980–2003. *Tourism Economics*, 13(4), 603–613.

Briassoulis, H. (2002). Sustainable tourism and the question of the commons. Annals of Tourism Research, 29(4), 1065–1085.

Brida et al. (2014) - Brida J.G., Cortes-Jimenez I. e Pulina M.) (2014).Has the tourism-led growth hypothesis been validated? A literature review. *Current Issues in Tourism*, 1.

Buckely, R. (2012). Sustainable tourism: research and reality, Ann. Tourism Res., 39 (2) (2012), pp. 528–546

Buckley, R. (2009). Ecotourism. Principles and practices (CABI publishing, Wallingford).

Burns, P. M. M.M. Sancho (2003). Local perceptions of tourism planning: the case of Cuellar, Spain, Tourism Manage., 24 (3) (2003), pp. 331–339

Butler, R.W. (1996). The concept of carrying capacity for tourist destinations: Dead or merely buried? Progress in Tourism and Hospitality Research, 2(3), 283–292.

Castro-Nuño et al. (2013) - Castro-Nuño, M., Molina-Toucedo J.A. & Pablo-Romero M.P.) (2013). Tourism and GDP: a meta-analysis of panel data studies. *Journal of Travel research*, *52*, 745-759.

Chen, Joseph S., 2015

Chiu et al., (2014) - Y.T. H. Chiu, W.I. Lee, T.H. Chen (2014), Environmentally responsible behavior in ecotourism: antecedents and implications, Tourism Manage., 40 (2014), pp. 321–329

Chou M. C. (2013). Does tourism development promote economic growth in transition countries? A panel data analysis.*Economic Modelling*, *33*, 226-232.

Clarke, J. (1997). A framework of approaches for sustainable tourism. Journal of Sustainable Tourism, 5(3), 224–233.

Cohen, E. (1978). Impact of tourism on the physical environment. Annals of Tourism Research, 5(2), 215–237.

Croall, J. (1995). Preserve or destroy: Tourism and the environment. London: Calouste Gulbenkian Foundation.

Dar et al. (2013) - Dar, R.A., Irfan, R.I., Romshoo, S.A., and Marazi, A. (2013). Sustainability of winter tourism in a changing climate over Kashmir Himalaya. Environmental Monitoring and Assessment, 185(12), 2549–2562.

deKadt, E. (ed.). (1979). *Tourism: Passport to Development*? New York: Oxford University Press.

Dolnicar e Leisch (2007), S. Dolnicar, F. Leisch, Selective marketing for environmentally sustainable tourism, Tourism Manage., 29 (4) (2007), pp. 672–680

Dolnicar et al. (2010) - S. Dolnicar, C. Laesser, K. Matus (2010), Short-haul city travel is truly environmentally sustainable, Tourism Manage., 31 (4) (2010), pp. 505–512

Dritsakis, N. (2004). Tourism as a long-run economic growth factor: An empirical investigation for Greece using causality analysis. *Tourism Economics, 10(3),*305–316.

Dritsakis, N. (2012). Tourism development and economic growth in seven Mediterranean countries: A panel data approach. *Tourism Economics*, *18*(4), 801–816.

East et al. (1998) - East, P., Luger, K., and Inmann, K. (1998). Himalayan tourism on the sustainable trail. In P. East, K. Luger, and K. Inmann (Eds.), Sustainability in mountain tourism: Perspective for the Himalayan countries (pp. 1–14). Delhi: Book Faith India.

Eugenio-Martín et al. (2004) - Eugenio-Martín, J. L., Morales, N. M., & Scarpa, R. (2004). Tourism and economic growth in Latin American countries: A panel data approach. FondazioneEni Enrico Mattei Working Paper Series, Nota di Lavoro, 26.

Farooquee et al. (2008) - Farooquee, N.A., Budal, T.K., and Maikhuri, R.K. (2008). Environmental and socio-cultural impacts of river rafting and camping on Ganga in Uttarakhand Himalaya. Current Science, 94, 587–594.

Fayissa et al. (2008) - Fayissa, B., Nsiah, C., & Tadasse, B. (2008). Impact of tourism on economic growth and development in Africa. *Tourism Economics*, *14*(4), 807–818.

Fazal, S., e Amin, A. (2012). Hanjis activities and its impact on Dal Lake and its environs — A case study of Srinagar city, India. Research Journal of Environmental and Earth Sciences, 4(5), 511–524.

Font and Harris, (2004), X. Font, C. Harris, Rethinking standards from green to sustainable, Ann. Tourism Res., 31 (4) (2004), pp. 986–1007

Font, X (2001). Environmental certification in tourism and hospitality: progress, process and prospects, Tourism Manage., 23 (3) (2001), pp. 197–205

Gartner, W.C. (1996). Tourism development: Principles, processes, and policies. New York: Van Nostrand Reinhold.

Geneletti, D., e Dawa, D. (2009). Environmental impact assessment of mountain tourism in developing regions: A study in Ladakh, Indian Himalaya. Journal of Environmental Impact Assessment Review., 29(4), 229–242.

Glavan, V. (2006). Tourism potential and its development. Bucharest: Ed. Fundației România de Maine.

Godde et al., (2000) - Godde, P.M., Price,M.F., and Zimmerman, F.M. (2000). Tourism and development in mountain regions. New York: CABI Publishing.

Gossling, S. (2002). Global environmental consequences of tourism. Global Environmental Change, 12, 283–302.

Grabara, K.J., e Bajdora, P. (2013). Towards to sustainable tourism framework, activities and dimensions. Procedia Economics and Finance, 6, 523–529.

Hall (1991; 1995),

Hall, C.M., e Page, S.J. (2004). The geography of tourism and recreation. New York: Routledge.

Holden, A (2009), The environment-tourism nexus influence of market ethics, Ann. Tourism Res., 36 (3) (2009), pp. 373–389

Holzner, M. (2011). Tourism and economic development: The beach disease? *Tourism Management, 32*, 922–933.

Howie, F. (2003). Managing the tourist destinations. London: Continuum.

Huiqin, L. e Linchun, H. (2011). Evaluation on sustainable development of scenic zone based on tourism ecological footprint: Case study of yellow crane tower in Hubei Province, China. Energy Procedia, 5, 145–151.

Hunter, C. (1997). Sustainable tourism as an adaptive paradigm. Annals of Tourism Research, 24, 850–867.

Iatu, C., e Bulai, M. (2011). New approach in evaluating tourism attractiveness in the region of Moldavia (Romania). International Journal of Energy and Environment, 2(5), 165–174.

Ibrahimova G. (2014). Tourism development challenges in Central and Eastern Europe. LAP (Lambert Academic publishing).

ICIMOD (2010). Two decades of mountain tourism in ICIMOD 1989–2010. Kathmandu: ICIMOD.

Imran et al., (2014) - S. Imran, K. Alam, N. Beaumont (2014), Environmental orientations and environmental behaviour: perceptions of protected area tourism stakeholders, Tourism Manage., 40 (2014), pp. 290–299

Islam, N. (1995). Growth empirics: a panel data approach. *The Quarterly Journal of Economics*, 1127-1170.

Jehangir et al., (2011) - Jehangir, A., Dar, N., Yousuf, A.R., and Sofi, A.H. (2011). Air quality at Sonamarg — A tourist hill station in Kashmir Valley. Journal of Experimental Sciences., 2(6), 18–22.

Jodha, N.S. et al. (1992). Mountain perspective and sustainability: A framework for development strategies. In N.S. Jodha, M. Banskota, and T. Partap (Eds.), Sustainable mountain agriculture, perspective and issues, Vol. 1. (pp. 41–82). New Delhi: Oxford and IBH Publishing.

Johnston and Tyrrell, (2005), A dynamic model of sustainable tourism, J. Travel Res., 44 (2) (2005), pp. 124–134

Khan, N.A. (2014). Municipal solid waste management — A case study of Srinagar city. Scholars Journal of Economics, Business and Management, 1(5), 179–185.

Ko, J. (2001). Assessing progress of tourism sustainability. Annals of Tourism Research, 28, 817–820.

Ko, J. (2005), Development of a tourism sustainability assessment procedure: a conceptual approach, Tourism Manage., 26 (3) (2005), pp. 431–445

Kruk, E., and Banskota, K. (2007). Mountain tourism in : From impacts to sustainability. In H. Bisht, S. Govind, and Rajwar (Eds.), Tourism and Himalayan biodiversity (pp. 15–34). Srinagar: Transmedia Publication.

Lanza, A., e Pigliaru, F. (2000). Tourism and economic growth: Does country's size matter? *RivistaInternazionale di ScienzeEconomiche e Commerciali*, *47*, 77–85.

Lanza et al. (2003) - Lanza, A., Temple, P., &Urga, G. (2003). The implications of tourism specialization in the long run: An econometric analysis for 13 OCDE economies. *Tourism Management, 24*, 315–321.

Lee, C.C., e Chang, C.P. (2008). Tourism development and economic growth: A closer look to panels. *Tourism Management, 29*, 80–192.

Lemmetyinen, A.e. Go, F.M. (2009). The key capabilities required for managing tourism business networks. *Tourism Management*, 30, 31-40.

Lim, C. (1997). Review of International Tourism Demand Models. *Annals of Tourism Research 24*, 835-849.

Lone et al. (2013) - Lone, S., Lone, F.A. e Malik, A. (2013). Carrying capacity assessment for the promotion of ecotourism in Bangus Valley: A future tourist destination of JandK — India. International Journal of Scientific Research, 2(3), 187–188.

M.I. Malik e M.S. Bhat (2015). Tourism Management Perspectives 16 (2015) 11–21 19

Malik et al. (2011) - Malik, M.I., Bhat, M.S., and Kuchay, N.A. (2011). Anthropogenic impact on forest cover in the western Himalayas — A case study of Lidder catchment in Kashmir Valley. Transactions, 33(1), 55–65.

Malik et al., 2013 - Malik, M.I., Bhat, M.S., and Kuchay, N.A. (2013). Watershed based land use land cover change detection analysis for sustainable development of Lidder catchment in Kashmir Himalayas. In M.H. Qureshi (Ed.), Jamia geographical studies, Vol. 2. (pp. 295–312). New Delhi: Manak Publications private limited.

Malik, M.I. (2013). Spatio-temporal analysis of urban dynamics in Kashmir Valley (1901–2011) using geographical information system. International Multidisciplinary Research Journal, 2(8), 21–26.

Malik, M.I. e Bhat, M.S. (2013). Identification of Tourist Potential Regions for Balanced Tourism Development in Pahalgam Tourist Destination of Kashmir Valley. 2(4), 189–190.

Malik, M.I., and Bhat, M.S. (2014). Integrated approach for prioritizing watersheds for management: A study of lidder catchment of Kashmir Himalayas. Environmental Management, 54(6), 1267–1287.

Manning et al., 2002 - Manning, R.,Wang, B., Valliere,W., Lawson, S., and Newman, P. (2002). Research to estimate and manage the carrying capacity of a tourist attraction: A study of Alcatraz Island. Journal of Sustainable Tourism, 10(5), 388–404.

Mashayekhan et al. (2014) - Mashayekhan, A., Calichi, M.M., Rassam, G.H., Hoseini, V., Jalilvand, H., and Mohalleh, S.M.M.S. (2014). Recreation carrying capacity estimations to support forest park management (case study: Telar Forest Park, Ghaemshahr, Iran). World Applied Sciences Journal, 29(3), 421–425.

Massidda, C., e Mattana, P. (2013).A SVECM analysis of the relationship between international tourism arrivals, GDP and trade in Italy. *Journal of Travel Research, 52(1)*, 93–105.

Mathieson, A., e Wall, G. (1982). Tourism: Economic, physical and social impacts. Harlow: Longman.

Mihalicˇ, T. (2013)., Performance of environmental resources of a tourist destination: concept and application, J. Travel Res., 52 (5) (2013), pp. 614–630

Miller et al., (2009) - G. Miller, K. Rathouse, C. Scarles, K. Holmes, J. Tribe (2009), Public understanding of sustainable tourism, Ann. Tourism Res., 37 (3) (2009), pp. 627–645

Monz et al. (2010) - Monz, C.A., Marion, J.L., Goonan, K.A., Manning, R.E., Wimpey, J., and Carr, C. (2010). Assessment andmonitoring of recreation impacts and resource conditions on mountain summits: Examples from the Northern Forest, USA. Mountain Research and Development, 30(4), 332–343.

Mushtaq et al. (2013) - Mushtaq, B., Rajni, R., Yaseen, T., wanganeo, A., and Yousuf, A.R. (2013). Variations in the physico-chemical properties of Dal Lake, Srinagar, Kashmir. African Journal of Environmental Science and Technology, 7(7), 624–633.

Narayan et al. (2010) - Narayan, P. K., Narayan, S., Prasad, A., & Prasad, B. C. (2010). Tourism, and economic growth: A panel data analysis for Pacific Island countries. *Tourism Economics, 16*(1), 169–183.

Nepal, S.K., e Chipeniuk, R. (2005). Mountain tourism: Toward a conceptual framework. Tourism Geographies, 7(3), 313–333.

Nestoroska, I. (2012). Identifying tourism potentials in Republic of Macedonia through regional approach. Procedia — Social and Behavioral Sciences, 44, 95–103.

Nghi et al., (2007) - Nghi, T., Lan, N.T., Thai, N.D., Mai, D., and Thanh, D.X. (2007). Tourism carrying capacity assessment for Phong Nha-Ke Bang and Dong Hoi, Quang Binh Province. VNU Journal of Science, Earth Sciences, 23, 80–87.

Nyaupane, G.P., e Chhetri, N. (2009). Vulnerability to climate change of nature based tourism in the Nepalese Himalayas. Tourism Geographies, 11(1), 95–119.

O'Riordan, T., e Turner, R. (1983). An annotated reader in environmental planning and management. Oxford: Pergamon Press.

Pablo-Romero M.P. e Molina J.A. (2013). Tourism and economic growth: A review of empirical literature. *Tourism Management Perspectives, 8*, 28-41.

Payne, J. E., e Mervar, A. (2010). The tourism–growth nexus in Croatia. *Tourism Economics, 16(4)*, 1089–1094.

Pearce, D. (1989). Tourist development (2nd ed.). Essex: Longman.

Pickering, C.M., e Buckley, R. (2003). Swarming to the summit—Managing tourists at Mt. Kosciusko, Australia. Mountain Research and Development, 23, 230–233.

Po, W. C., e Huang, B. N. (2008). Tourism development and economic growth, a nonlinear approach. *Phisica A*, 387, 5535–5542.

Pratt, S. (2015). The economic impact of tourism in SIDS. *Annals of Tourism Research, 52*, 148-160.

Proenca, S., & Soukiazis, E. (2008). Tourism as an economic growth factor: A case study for Southern European countries. *Tourism Economics, 14*(4), 791–806.

Queiroz et al, 2014) - Queiroz, R.E., Ventura, M.A., Guerreiro, J.A., e Cunha, R.T.D. (2014). Carrying capacity of hiking trails in Natura 2000 sites: A case study from North Atlantic Islands (Azores, Portugal). Journal of Integrated Coastal Zone Management, 14(2), 233–242.

Ramdas, M., e Badaruddin, M. (2014). Impacts of tourism on environmental attributes, environmental literacy and willingness to pay: A conceptual and theoretical review. Procedia — Social and Behavioral Sciences, 144, 378–391.

Rashid, I., e Romshoo, S.A. (2012). Impact of anthropogenic activities on water quality of Lidder River in Kashmir Himalayas. Environmental Monitoring and Assessment, 185(6), 4705–4719.

Rawat, D.S., e Sharma, S. (1997). The development of a road network and its impact on the growth of infrastructure: A study of Almora District in the Central Himalaya. Mountain Research and Development, 17, 117–126.

Richards, G. (1996) Tourism in Central and Eastern Europe: Educating for Quality. Tilburg University Press, Tilburg.

Saarinen, J. (2006). Traditions of sustainability in tourism studies. Annals of Tourism Research, 33(4), 1121–1140.

Sayan, M.S., e Atik, M.S. (2011). Recreation carrying capacity estimates for protected areas: A study of Termessos National Park. Ekoloji, 20(78), 66–74 http://dx.doi.org/10.5053/ekoloji.2011.7811.

Shan, J. e Wilson K. (2001). Causality between trade and tourism: empirical evidence from China. *Applied Economics Letters, 8*, pp.279-283.

Sharpley, R. (1994). Tourism, tourists and society. Huntingdon: ELM.

Simpson, K. (2001). Strategic planning and community involvement as contributors to sustainable tourism development. Current Issues in Tourism, 4(1), 3–41.

Sinclair, A.J.,e Ham, L.H. (2000). Household adaptive strategies: Shaping livelihood security in the western Himalaya. Canadian Journal of Development Studies, 21, 89–112.

Sinclair, M. T. (1998). Tourism and economic development: A survey. *The Journal of Development Studies, 34*, 1–51.

Singh, D. R. (2008). Small island developing states (SIDS). Tourism and economic development. *Tourism Analysis*, 13, 629–636.

Singh, T.V. (1989). The Kullu Valley: Impact of tourism development in the mountain areas. New Delhi: Himalayan Books.

Sloan et al., (2012) - P. Sloan, W Legrand, J.S. Chen (2012), Sustainability in the Hospitality Industry: Principle of Sustainable Operations (2nd ed.) Routledge, New York (2012)

Smith, S.L.J. (1989). Tourism analysis: A handbook. Harlow: Longman Scientific.

Stephenson, J.L. (1993). The impacts of tourism on nature reserves in Madagascar: A case study. Environmental Conservation, 20, 262–265.

Surugiu, C., e Surugiu, M. R. (2013). Is the tourism sector supportive of economic growth? Empirical evidence on Romanian tourism. *Tourism Economics, 19(1)*, 115–132.

Syamlal, G.S. (2008). Carrying capacity study of coastal tourism in Kumarakom, Kerala. Jurnal Ekonomi dan Bisnis, 13(1), 1–15.

Tucker, H. e Boonabaana, B. (2011). A critical analysis of tourism, gender and poverty reduction. *Journal of Sustainable Tourism, 20(3),*437-455.

Twynam e Johnston, (2002), G.D. Twynam, M.E. Johnston, The use of sustainable tourism practices, Ann. Tourism Res., 29 (4) (2002), pp. 1165–1168

Vernon et al., (2005) - J. Vernon, S. Essex, D. Pinder, K. Curry, Collaborative policymaking: local sustainable projects, Ann. Tourism Res., 32 (2) (2005), pp. 325–345

Wagar, J.A. (1964). The carrying capacity of wild lands for recreation. Forest science monograph no. 7. Washington, D.C.: Society of American Foresters.

Waligo et al., (2013), V.M. Waligo, J. Clarke, R. Hawkins, Implementing sustainable tourism: a multi-stakeholder involvement management framework, Tourism Manage., 36 (2013), pp. 342–353

WCED (1987). Our common future. Oxford: Oxford University Press. Wenjun, L. (2004). Environmental management indicators for ecotourism in China's nature reserves: A case study in tianmushan nature reserve. Tourism Management, 25, 559–564.

Weaver, D. B. (2013). Protected area visitor willingness to participate in site enhancement activities, J. Travel Res., 52 (3) (2013), pp. 377–391

Weng, C.C. e Wang, K.K. (2004). Scale and scope economies of international tourist hotels in Taiwan. *Tourism Management, 25*, 761-769.

World Bank (2011). World Development Indicators Database. Retrieved , from 〈http: //siteresources.worldbank.org/DATASTATISTICS/Resources/GNIPC.pdf〉.

WTO (1981). Technical handbook on the collection and presentation of domestic and international tourism statistics. Madrid: World Tourism Organization.

Yasarata et al., (2010), M. Yasarata, L. Altinay, P. Burns, F. Okumus, Politics and sustainable tourism development: can they co-exist? Voices from North Cypus, Tourism Manage., 31 (3) (2010), pp. 345–356

Zabortseva, T.I., e Yevstropieva, O.V. (2009). Current socio-ecological problems of recreational development of Pribaikalsky National Park. Geography and Natural Resources, 30, 359–366.

Zacarias et al., 2011 - Zacarias, D.A., Williams, T.A., and Newton, A. (2011). Recreation carrying capacity estimations to support beach management at Praia de Faro, Portugal. Applied Geography, 31, 1075–1081. 20 M.I. Malik, M.S. Bhat / Tourism Management Perspectives 16 (2015) 11–21

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

MULTILOCALIZACIÓN E ESTATISTICAS REXIONAIS

Jorge Santiso Blanco[1]

[1] Instituto Galego de Esatística

RESUMEN

Factores organizativos e funcionais explican a multilocalización da actividade empresarial, un fenómeno expansivo que abarca tanto globalización como rexionalización. Dende unha perspectiva rexional a multilocalización ten un interese analítico en campos tan amplos da Economía como a unidade de mercado, a fiscalidade, o esquema de organización autonómica ou o propio sistema estatístico rexional.

Esta tendencia crecente da multilocalización non ten o reflexo necesario na metodoloxía para o seu estudio sistemático, a pesares de recentes avances na clasificación das actividades das sedes centrais (Eurostat, 2008) ou na delimitación da actividade a nivel de establecemento (Comisión Europea, 2013), de eficacia condicionada polo Plan General de Contabilidade, que non contempla a desagregación das contas de produción e explotación.

Diferentes estudios teñen abordado os efectos da multilocalización nas economías rexionais, aínda que baixo enfoques dominados polo estudio do impacto das sedes centrais (Bescós, 2003; Ayet y Sanz, 2004; Mella, 2010)

Nesta comunicación trátanse algúns problemas específicos da multilocalización no enfoque das estatísticas de empresas, con propostas concretas que permiten unha aproximación razoable á medición da actividade económica rexional..

Palabras y frases clave: Estatísticas de empresas; Multilocalización empresarial; Sedes centrais.

REFERENCIAS

Arias, C. et Alt. (2010) El problema de la regionalización en las encuestas dirigidas a empresas.

Ayet, C. y Sanz, B. (2004) Las sedes centrales en la economía madrileña.

Comisión Europea (2013) Reglamento (UE) nº 549/2013 del Parlamento Europeo y del Consejo, de 21 de mayo de 2013, relativo al Sistema Europeo de Cuentas Nacionales y Regionales de la Unión Europea. Diario Oficial de la Unión Europea L 174/1, 26.6.2013.

Comunidad de Madrid (1995) Aproximación al estudio de las Sedes Centrales en Madrid

Edwards, R. (2007) Los problemas estadísticos derivados de la globalización: medidas adoptadas por el FMI. Institutos de estadísticas de América Latina: los retos que presentan la modernización, la transparencia y la independencia tecnológica. Santiago de Chile—julio de 2007

Eurostat (2008) Statistical classification of economic activites in the European Community

Guimbao Bescós (2003) El marco input-output: cuestiones relevantes en la aplicación del SEC-95, 4-5.

MellaA et Alt. (2010) Estimación del VAB generado por las sedes centrales en Madrid

XI Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

MODELOS TEMPORALES LINEALES MIXTOS: APLICACIÓN A DATOS DE LA ENCUESTA DE CONDICIONES DE VIDA

María Dolores Esteban[1], Domingo Morales[1] y Agustín Pérez[2]

[1] Centro de Investigación Operativa, Universidad Miguel Hernández de Elche..
[2] Departamento de Estudios Económicos y Financieros, Universidad Miguel Hernández de Elche.

RESUMEN

En este trabajo se aborda la adaptación de modelos de Fay-Herriot temporales para la estimación de indicadores de pobreza y condiciones de vida en pequeñas áreas. Se investiga la aplicabilidad a la Encuesta de Condiciones de Vida (ECV) de los estimadores EBLUP basados en los modelos Fay-Herriot temporales. El objetivo final es la obtención de estimadores de algunos indicadores de esta encuesta referidos a provincias y comarcas de la Comunidad valenciana con desagregación por sexo y edad. Así mismo se contempla el posterior tratamiento para que adquieran consistencia con las estimaciones de la ECV para la comunidad autónoma. Concretamente se trata de los siguientes indicadores: 1. Renta anual media (sin alquiler imputado) por unidad de consumo, 2. Tasa de riesgo de pobreza por sexo y edad, 3. Carencia material (personas) por sexo y edad, 4. Carencia material severa (personas) por sexo y edad, 5. Carencia material (hogares). Para la aplicación de modelos temporales se usarán datos de 2012 y 2013, siendo el objeto del trabajo la estimación de los citados indicadores de pobreza referidos a 2013

Palabras y frases clave: Indicadores de pobreza, modelo Fay-Herriot, modelos temporales de área, predictores lineales óptimos, error cuadrático medio, bootstrap.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

ESTIMATION OF EMPLOYED PEOPLE BY ECONOMIC ACTIVITY IN GALICIA

María José Lombardía[1], Esther López Vizcaíno[2] e Cristina Rueda[3],

[1]Universidade da Coruña
[2]Instituto Galego de Estatística
[3]Universidad de Valladolid

RESUMO

The objective of this work is the estimation in all the quarters the employed people by economic activity in Galicia from third quarter of 2009 to the fourth quarter of 2013. Our domains are the economic activities using the National Classification o Economic Activities (CNAE). When it is necessary to make estimates for unplanned domains they have very low sample sizes and therefore very high sampling errors. To obtain reliable estimates in these areas is necessary to study models that use auxiliary information related to the purpose. We are going to evaluate the performance of different estimators: estimators from the Fay Herriot model, semiparametric and nonparametric models.

Palabras e frases chave: small area estimation, selection models, AIC, employed people.

1. INTRODUCTION

Government authorities dealing with the economy of a region need to know the number of employed people in each of the economic activities and their evolution over time. This is necessary to influence economic sectors in decline or to encourage potentially emerging sectors. It is also important to study the productivity of the different activities and thus promote employment policies in those economically more productive sectors. From a statistical point of view on the economy accounts for Galicia (Spain) estimates of employment in each of the activities are done through a study of the different statistical sources, remaining at the end with which it is more consistent in terms, for example, changes. With this work we go a step further trying to find what is the most appropriate model that determines the evolution of the employed people by economic activity, avoiding craft procedures.

The Labour Force Survey (LFS) is the only existing Spanish statistic with suitable periodicity to carry out the study of the labour market situation and evolution. The LFS is a quarterly survey that is designed to study the economic activity of the population. In particular, the LFS evaluates the number of persons aged 16 or more that are employed, unemployed or inactive and then it analyzes the characteristics of these groups. The problem of the LFS is that when you need to make estimates for unplanned domains (in our case economic activities), we find very low sample sizes of domains and therefore very high sampling errors. The small area estimation techniques to deal with this kind of situations. Some descriptions of the Small Area Estimation (SAE) theory can be found in the monograph of (Rao, 2003).

The aim of this work is to estimate in each quarter the employment people by economic activity in Galicia in the period from the third quarter of 2009 to the fourth quarter of 2013 using data from the LFS. Therefore, domains are considered economic activities (agriculture, forestry, food industry, ...). Economic activities considered are bounded by the National Classification of Economic Activities (NACE) two digits. There are a total of 84 domains for each 18 quarters considered.

2. DATA DESCRIPTION

We deal with data from the LFS of Galicia in the period from the third quarter of 2009 to the fourth quarter of 2013. We have D domains, where each of them is an economic activity. P_d is the population in each of the economic activities that consists of those people employed in this activity d and the unemployed that in their last job were employed in that activity. Our goal is the estimate of total employment by economic activity, namely:

$$Y_d = \sum_{j \in P_d} y_j,$$

where $y_j = 1$ if the person j of domain d is employed and $y_j = 0$ in other case.

The LFS does not produce official estimates at the domain level, but the analogous direct estimates of the total Y_d, the mean $\bar{Y}_d = Y_d/N_d$ and the size N_d are

$$\hat{Y}_d^{dir} = \sum_{j \in S_d} w_j \, y_j, \; \hat{\bar{Y}}_d^{dir} = \hat{Y}_d^{dir}/\hat{N}_d^{dir}, \; \hat{N}_d^{dir} = \sum_{j \in S_d} w_j$$

where S_d is the sample domain and w_j's are the official calibrated sampling weights.

LFS is designed to obtain precise estimates at activity sector (NACE, 1 digit). The problem of the LFS is that when the domains are bellow the planned level. For the fourth quarter of 2013 the minimum sample size in the domains is 1, the first quartile is 12 and the median 31, therefore for some domains with the direct estimator cannot get a reliable estimate of our objective. To obtain reliable estimates in these domains is necessary to propose models that use auxiliary information related to the purpose of study.

In this work we use area level models. These models need only information about the domains totals for the auxiliary variables, information that can be usually found in administrative registers. Therefore, as auxiliary variables for the estimation of the target variable we will use the people registered in the social security system (SS) and the registered work contract (CONT) by economic activity. This information is available for each quarter and for all variables of interest since the third quarter of 2009 to the fourth quarter of 2010. We use like the response variable the logarithm of employed people in each economic activity, $\log(Y_d)$, where Y_d is the direct estimator obtained from the LFS.

4. MODEL SELECTION

The question of model selection has received a lot of attention in the literature in the past (starting with the well-known paper by Akaike, 1973), and also in recent years due, among other reasons, to the increasing complexity of modeling approaches. In particular, the question has received considerable attention in the context of linear mixed models, although not so much in small area problems. One of the most popular approaches to model selection is to use the AIC (Akaike Information Criteria).

The value of AIC for a model M is defined as $AIC(M) = 2\ln(l(M)) - 2D$ where $l(M)$ is the model likelihood and D is a penalization factor, originally (Akaike, 1973) equal to p, the number of parameters in the model. The model with the highest value of AIC is then selected. It's also usual to use the DF instead of p, DF (Degrees of Freedom) coincides with p for simple models as the normal linear regression model and with the number of *free parameters* or the parameters of the final model, in other cases. However, this is not always so simple for more complex models.

A lot of work has been done during the last years in deriving measures of the complexity of models in such cases, to be used, particularly, as a penalization factor in the AIC. One of the proposed solutions is to use the Generalized Degrees of Freedom (GDF), originally defined in Ye (1998) for normal models. The most interesting advantages of these GDF are its applicability to

different modeling procedures and its ease of calculation using a data perturbation and bootstrap approach, more details in You et al. (2014).

Our proposal of the $GDF(M)$, is inspired in Ye (1998) and You et al. (2014), and is defined to be used for a range of models going from a simple linear synthetic model (without random effects) to the most complex nonparametric with the random effect model. Indeed, in this work we have information available on several domains/areas and time periods, then the small area estimators can be derived borrowing strength from one or two of these dimensions. It may happen that the trend patter is more abrupt than desired when estimators that borrow strength across areas instead of time are considered. In fact, as we can see in our particular case, very different patterns, some of them quite abrupt, are exhibited depending on the model and the area characteristics, point out to a different choice of model-based estimator for each area. On the other side, the option to consider time series models (models that borrow strength across time) may generate too smoothed patterns and prevent major trend changes being detected.

In this work we propose a very different and simple strategy to borrow strength across time, as well as across domains, that lacks this latter drawback and makes use of the AIC. The idea is to select the model to estimate each small area using the area-specific component of an aggregated AIC measure across time.

REFERENCES

Akaike, H. (1973) Information theory and the maximum likelihood principle. nternational Symposium on Information Theory, 267–281.

Rao, J.N.K. (2003) Small area estimation. Wiley, New York.

Ye, J. (1998) On measuring and correcting the effects of data mining and model selection. J. Amer. Statist. Assoc., 93, 120–131

You, C., Mller, S. and Ormerod, J. T. (2014). On generalized degrees of freedom with application in linear mixed models selection. Sat. Comput., DOI 10.1007/s11222-014-9488-7.

Comunicacións orais

Sesión 6 de Estatística

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

IMPACTO DO PROGRAMA GALEGO DE DETECCIÓN PRECOZ DO CANCRO DE MAMA (PGDPCM) SOBRE A MORTALIDADE POR CANCRO DE MAMA EN GALICIA. PERÍODO 1980-2012

G Naveira Barbeito1[1], M Pollán Santamaría[2], MI Santiago Pérez[1], AB Fernández Llanes[1], R Zubizarreta Aberdi[1] e A Malvar Pintos[1]

[1] Dirección Xeral de Innovación e Xestión da Saúde Pública, Xunta de Galicia
[2] Centro Nacional de Epidemiología, Instituto de Salud Carlos III y CIBERESP

RESUMO

En Galicia, no ano 1992, púxose en marcha o Programa Galego de Detección Precoz do Cancro de Mama (PGDPCM), dirixido inicialmente ás mulleres galegas de 50 a 64 anos, aínda que na actualidade abarca ata 69 anos de idade. O obxectivo deste traballo foi estimar o impacto do programa sobre a mortalidade por cancro de mama en Galicia, a través dunha análise idade-período-cohorte. Para isto axustouse un modelo de regresión binomial negativa no que a variable resposta foi o logaritmo das taxas específicas de mortalidade por cancro de mama e as variables explicativas foron: a idade, o período, a cohorte, o cribado considerando a participación e a incidencia de cancro de mama. Os resultados obtidos amosan unha redución do 29% (IC95%: (25, 33)) no risco relativo de morrer por cancro de mama para as mulleres que pasaron polo programa (naqueles concellos nos que se observou unha boa participación) en comparación co grupo de mulleres non cribadas. Este efecto podería estar asociado a un impacto positivo do PGDPCM aínda que non é posible diferencialo totalmente doutros efectos.

Palabras e frases chave: Cancro de mama, cribado, mortalidade, modelo idade-período-cohorte.

1. INTRODUCIÓN

Nos anos 70 e 80 varios ensaios clínicos amosaban que o cribado mamográfico reducía a mortalidade por cancro de mama [1]. Baseándose nos seus resultados, na década dos 90 puxéronse en marcha programas de cribado poboacional de cancro de mama en Europa e na maioría das comunidades autónomas españolas [2]. En Galicia, a Dirección Xeral de Innovación e Xestión da Saúde Pública (DXIXSP), entón Dirección Xeral de Saúde Pública, puxo en marcha en 1992 o Programa galego de detección precoz do cancro de mama (PGDPCM) que invita ás mulleres do grupo de idade obxectivo do programa e residentes en Galicia a realizar unha mamografía de cribado cada dous anos. O seu obxectivo principal é reducir a mortalidade por cancro de mama a través da súa detección en mulleres asintomáticas, cun estadio ao diagnóstico máis temperán o que mellora a resposta ao tratamento [3].

O programa comeza en 1992 cunha unidade de exploración que atende a un número limitado de concellos e a partir dese momento esténdese progresivamente ata acadar o 100% de cobertura no ano 1998. A proba de cribado, unha mamografía con dúas proxeccións de cada mama, é valorada por dous radiólogos de forma independente prevalecendo o resultado máis desfavorable dos dous. En caso de que algún dos radiólogos detecte anormalidades na mamografía, realízanse probas complementarias coa finalidade de confirmar ou descartar patoloxía maligna. O grupo de idade obxectivo era inicialmente de 50 a 64 anos e a partir do ano 2009 esténdese ata os 69 anos, despois dun proceso de ampliación progresiva iniciado no ano 2005 seguindo as recomendacións da Unión Europea [4].

A participación no programa mantense por riba do 70% desde 1997 e desde 2006 permanece relativamente estable entorno ao 80%, unha boa participación se temos en conta que a Guía europea de garantía de calidade en cribado e diagnóstico de cancro de mama [5] establece como obxectivo desexable superar o 75%.

Actualmente cuestiónase a posible perda de relevancia dos ensaios clínicos nos que se baseou a posta en marcha dos programas poboacionais de cribado posto que pasaron décadas dende a súa

realización, anos nos que se produciron importantes melloras técnicas nos equipos mamográficos e grandes avances no tratamento e manexo do cancro de mama. Non obstante, estudos observacionais máis recentes, en xeral, corroboran unha redución na mortalidade das mulleres de 50 a 69 anos invitadas a participar no cribado mamográfico [6] [7].

O obxectivo deste traballo foi estimar o impacto do PGDPCM na mortalidade por cancro de mama en Galicia, para as mulleres de 30 anos e máis.

2. MATERIAL E MÉTODOS

Levouse a cabo un estudo ecolóxico poboacional a través dunha análise idade-período-cohorte [8-10], na que se axustou o logaritmo das taxas específicas de mortalidade empregando a regresión binomial negativa. O uso dos modelos idade-período-cohorte está amplamente estendido xa que permite describir a tendencia temporal dun evento, como pode ser a mortalidade ou incidencia de certa enfermidade, en función da:

- Idade do individuo ao producirse o evento. O efecto desta variable debe ser tido en conta xa que a probabilidade de ocorrencia dun evento como pode ser o cancro de mama tende a aumentar a medida que se cumpren anos.
- Período de tempo no que se produciu o evento. O efecto desta variable se relaciona con cambios producidos nun determinado momento e que afectan de modo similar a todas as idades, como poden ser os avances médicos.
- Cohorte de nacemento da persoa. O efecto desta variable asóciase con cambios producidos a nivel xeracional, como poden ser cambios no patrón de hábito tabáquico nas mulleres novas en comparación coas mulleres de idades máis avanzadas.

Estas tres variables atópanse relacionadas do seguinte xeito, cohorte=período-idade. Esta dependencia lineal que presentan leva consigo o denominado "problema de identificabilidade" cando se incorporan conxuntamente nun modelo de regresión, de modo que non existe unha única solución para o axuste. Este problema non inflúe cando o interese se centra en predicir as taxas específicas pero si cando se está interesado en interpretar o efecto da idade, do período ou da cohorte. Existen varias propostas para solucionar este problema, neste traballo empregouse a proposta baseada na eliminación da tendencia lineal indistinguible do efecto período e cohorte ("drift").

Os datos organizáronse por concello, idade e período. A idade foi considerada en grupos quinquenais, identificando por separado aqueles grupos de idade comprendidos entre 30 e 54 anos, os comprendidos entre 55 e 84 anos e o grupo de 85 anos e máis, xa que o comportamento da mortalidade por cancro de mama difire substancialmente entre eses tres grupos de idade. O período foi considerado o ano de defunción. E a cohorte obtívose a partir das anteriores (idade e período). As variables período e cohorte foron engadidas no modelo por medio de funcións splines cúbicas naturais, funcións definidas por pedazos mediante polinomios de grao 3, que lle proporcionan ás mesmas maior flexibilidade. O número de nodos interiores empregados foron 5 para o período e 3 para a cohorte (menos nodos que os empregados para o período polo simple feito de que nos valores máis extremos das cohortes existen relativamente menos eventos).

A parte das variables explicativas idade, período e cohorte, consideráronse as variables:

- Cribado. Esta variable identifica ás mulleres cribadas polo PGDPCM ademais de considerar a porcentaxe de participación das mulleres do concello no programa, xunto coa porcentaxe de mulleres cuberta polo programa para cada grupo de idade. De modo que se crearon 3 categorías para esta variable: ausencia de cribado, cribado con participación menor do 50% e cribado con participación igual ou superior ao 50%.
- Incidencia de cancro de mama. Calculáronse, para cada concello e para o período 1996-2011, as razóns de incidencia estandarizadas (RIE) empregando como taxa estándar a da poboación de Galicia durante todo o período (valores superiores [inferiores] a 1 indican maior [menor] risco de padecer un cancro de mama en comparación a Galicia). Estas razóns foron suavizadas por medio dun modelo xerárquico de Bayes [11], proposto por Besag, York e Mollie [12] e baseado en métodos de Monte Carlo vía cadeas de Markov (MCMC, segundo as súas siglas en inglés), que permite considerar información a priori para poder incorporar a variabilidade dos riscos e unha estrutura de correlación espacial entre as áreas veciñas. A partir dos valores suavizados se crearon 5 categorías para esta

variable: valores menores de 0,75; entre 0,75 (incluído) e 0,85; entre 0,85 (incluído) e 0,95; entre 0,95 (incluído) e 1,05 e superiores ou iguais a 1,05.

Para o axuste do modelo idade-período-cohorte levouse a cabo un proceso de modelización cara diante, de xeito que se axustaron modelos de regresión binomial negativa aniñados por adición secuencial das variables explicativas anteriormente descritas. Os modelos axustados foron:

Modelo I: Idade
Modelo I+D: Idade+Drift
Modelo I+P: Idade+Período
Modelo I+C: Idade+Cohorte
Modelo I+P+C: Idade+Período+Cohorte
Modelo I+P+C+Inc: Idade+Período+Cohorte+Incidencia
Modelo I+P+C+Inc+Cri: Idade+Período+Cohorte+Incidencia+Cribado

A selección do modelo adecuado realizouse a través do test de razón de verosimilitudes que compara pares de modelos aniñados (un completo e outro aniñado no primeiro), por medio da diferenza de "deviance" entre modelos. Esta diferenza segue unha distribución chi-cadrado con k graos de liberdade, onde k é a diferenza entre o número de parámetros dos modelos a comparar. Este test permite contrastar a hipótese nula de que os coeficiente asociados ás novas variables engadidas ao modelo completo, en comparación co modelo aniñado, sexan iguais a cero.

A formulación do modelo completo (que incorpora todas as variables explicativas) sería:

$$\log(\lambda_i) = \log\left(\frac{d_i}{n_i}\right) = \log(d_i) - \log(n_i) = \alpha + \beta_{I_{[30-54]}} \cdot I_{[30-54]\,i} + \beta_{I_{[55-84]}} \cdot I_{[55-84]\,i} +$$

$$+\beta_{I_{[85+]}} \cdot I_{[85+]\,i} + f_P(P_i) + f_C(C_i) + \beta_{Inc} \cdot Inc_i + \beta_{Cri} \cdot Cri_i + v_i, \text{ con i=1,..., N}$$

Simplificando a expresión quedaría do seguinte xeito:

$$\log(d_i) = \log(n_i) + \alpha + \beta_{I_{[30-54]}} \cdot I_{[30-54]\,i} + \beta_{I_{[55-84]}} \cdot I_{[55-84]\,i} +$$

$$+\beta_{I_{[85+]}} \cdot I_{[85+]\,i} + f_P(P_i) + f_C(C_i) + \beta_{Inc} \cdot Inc_i + \beta_{Cri} \cdot Cri_i + v_i, \text{ con i=1,..., N}$$

Onde λ_i son as taxas específicas de mortalidade por cancro de mama, d_i son as defuncións por cancro de mama, n_i é o tamaño da poboación, $I_{[30-54]\,i}$ é a idade en grupos quinquenais para as mulleres de 30 a 54 anos, $I_{[55-84]\,i}$ é a idade en grupos quinquenais para as mulleres de 55 a 84 anos, $I_{[85+]\,i}$ é a idade para as mulleres de 85 anos e máis, P_i é o ano de defunción, C_i é a cohorte de nacemento, Inc_i é a variable incidencia, Cri_i é a variable cribado. O termo $\log(n_i)$ é un factor de ponderación, tamén coñecido como "offset". A exponencial de α estima o valor esperado da taxa de mortalidade no grupo de referencia, onde as variables explicativas toman todas o valor cero. A exponencial de cada un dos parámetros β, estima o efecto da variable asociada, é dicir, o risco relativo da variable polo aumento dunha unidade ou cambio de categoría con respecto á de referencia, segundo sexa unha variable cuantitativa ou cualitativa, respectivamente. Por último, $e^{v_i} \sim Gamma(1/\theta, \theta)$ con media 1 e varianza θ, sendo θ o parámetro de sobredispersión (cando θ toma o valor cero estase ante o modelo de regresión de Poisson). A estimación dos parámetros do modelo realizouse polo método de máxima verosimilitude.

A partir dun modelo similar ao que finalmente foi elixido na análise anterior se axustaron as taxas de mortalidade no grupo de mulleres non cribadas. A partir deste axuste predíxose o comportamento da mortalidade no grupo de mulleres cribadas. Deste modo estimáronse o número de mortes e as taxas de mortalidade esperadas baixo o suposto de que non fora introducido o programa de cribado na poboación. Calculáronse tamén as diferenzas absolutas (DA) e relativas (DR) entre o número de defuncións observadas e esperadas.

Os datos de mortalidade por cancro de mama (CIE 9: 174; CIE 10: C50) para o período 1980-2012 obtivéronse do Rexistro Galego de Mortalidade (DXIXSP). A información relativa á poboación de mulleres obtívose, para o período 1980-1997, de estimacións intercensuais calculadas pola DXIXSP a partir dos Censos de 1981 e 1991 e dos Padróns de 1986, 1996 e 1998. E, para o período 1998-2012, do

Padrón municipal de habitantes (Instituto Galego de Estatística-IGE). E a información sobre o número de mulleres ingresadas en Galicia con diagnóstico ao alta de cancro de mama durante o período 1996-2011obtívose do rexistro poboacional do Conxunto Mínimo Básico de Datos de Altas Hospitalarias (CMBD-AH).

3. RESULTADOS

Durante o período de estudo producíronse 12.430 defuncións por cancro de mama en mulleres galegas de 30 anos e máis, destas o 37,6% (4.676) producíronse antes do ano 1994 (primeiro ano considerado cribado neste estudo para os primeiros concellos incorporados ao PGDPCM) e, para as restantes, o 65,5% (5.076) producíronse no grupo considerado non cribado, o 5,6% (436) no grupo cribado con participación menor do 50% e o 28,9% (2.242) no grupo cribado con participación superior ou igual ao 50%. A idade media ao falecemento foi en aumento ao longo do período de estudo, pasando de valores en torno a 62 anos de idade nos primeiros anos do período ata acadar valores en torno aos 71 anos para os últimos anos do estudo.

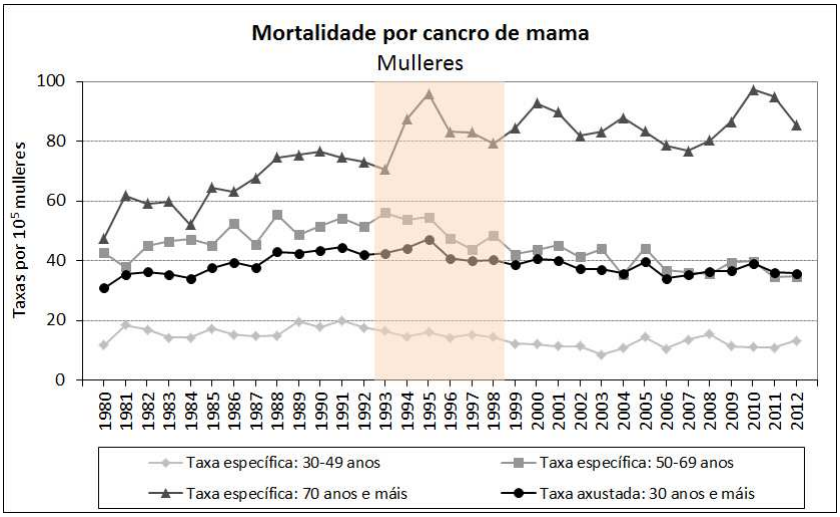


Figura 1. Taxas específicas (de 30 a 49 anos, de 50 a 69 anos e de 70 anos e máis) e axustadas (poboación estándar: Censo de Galicia de 2001) por 100.000 mulleres para a mortalidade feminina por cancro de mama. Sombreado en laranxa amósase o período de progresiva implantación do PGDPCM, en función do ano de finalización da primeira campaña segundo o concello.

As taxas de mortalidade, axustadas por idade, evolucionaron de forma crecente ata mediados dos anos 90 e logo inverteron a súa tendencia. Dun xeito similar comportáronse as taxas específicas das mulleres de 50 a 69 anos, con valores próximos aos das taxas axustadas. As taxas específicas no grupo de idade de 30 a 49 anos foron as máis baixas, situándose entorno aos 15-20 casos por 100.000 mulleres na primeira metade do período e entre 10-15 casos por 100.000 na segunda metade. As taxas específicas no grupo de 70 anos e máis presentaron os valores máis altos en comparación coas taxas específicas dos demais grupos de idade, a tendencia para este grupo foi crecente ata principios dos anos 90, momento no que a serie parece aumentar de nivel e manterse lixeiramente estable, oscilando en torno aos 85 casos por 100.000 mulleres (figura 1).

Tras realizar o axuste dos diversos modelos propostos, levouse a cabo o test de razón de verosimilitudes que permitiu seleccionar, como modelo que mellor axusta os datos, o modelo que inclúe todas as variables explicativas. O resultado do test de sobredispersión leva ao rexeitamento da hipótese nula de datos equidispersos, polo que un modelo de Poisson non sería adecuado para axustar os datos, de modo que finamente se empregou o modelo de regresión binomial negativa. O axuste deste modelo preséntase na táboa 1, sen incorporar os resultados das variables idade, período e cohorte.

O risco de morrer por cancro de mama diminúe un 25% ou máis, de forma estatisticamente significativa, naqueles concellos cunha RIE suavizada inferior a 0,85 en comparación ao risco dos concellos cunha RIE suavizada en torno a 1. Para os concellos onde o valor suavizado da RIE foi igual ou superior a 1,05 o risco aumenta un 20%, aumento estatisticamente significativo.

O risco relativo de morrer nos grupos cribados cunha “pobre” participación redúcese un 9% pero non de forma significativa. Sen embargo, para os grupos cribados cunha participación superior ou igual ao 50% o risco se reduce nun 29%, neste caso de forma estatisticamente significativa. En ambos casos, en comparación co grupo de mulleres non cribadas.

Variables	**RR**	**IC 95%**		**Valor p**
		Límite inf	**Límite sup**	
Incidencia (RIEs suavizadas)				
Menores de 0,75	0,67	0,62	0,74	<0,001
Entre 0,75 (incluído) e 0,85	0,75	0,69	0,81	<0,001
Entre 0,85 (incluído) e 0,95	1,04	0,98	1,11	0,233
Entre 0,95 (incluído) e 1,05	1,00	-	-	-
Maiores ou iguais a 1,05	1,20	1,13	1,28	<0,001
Cribado				
Ausencia de cribado	1,00	-	-	-
Cribado con participación <50%	0,91	0,82	1,01	0,075
Cribado con participación ≥50%	0,71	0,67	0,75	<0,001

Táboa 1: Axuste do modelo de regresión binomial negativa. Estimacións dos riscos relativos de morrer (RR), os seus intervalos de confianza e o valor p asociados ás variables incidencia e cribado. Os resultados correspondentes ás variables idade, período e cohorte non se amosan.

Á hora de estimar o número de mortes esperadas baixo o suposto de que non se introducise o programa de cribado na poboación, axustouse un modelo similar ao anteriormente descrito no grupo de mulleres non cribadas. Os resultados do axuste amósanse na táboa 2 sen engadir os resultados para as variable idade, período e cohorte.

Variables	**RR**	**IC 95%**		**Valor p**
		Límite inf	**Límite sup**	
Incidencia (RIEs suavizadas)				
Menores de 0,75	0,67	0,61	0,74	<0,001
Entre 0,75 (incluído) e 0,85	0,74	0,67	0,80	<0,001
Entre 0,85 (incluído) e 0,95	1,05	0,98	1,13	0,173
Entre 0,95 (incluído) e 1,05	1,00	-	-	-
Maiores ou iguais a 1,05	1,23	1,15	1,32	<0,001

Táboa 2: Axuste do modelo de regresión binomial negativa no grupo de mulleres non cribadas. Estimación do risco relativo de morrer (RR), o límite inferior e superior do seu intervalo de confianza e o valor p asociado á variable incidencia. Os resultados correspondentes ás variables idade, período e cohorte non se amosan.

A partir deste axuste predíxose o comportamento da mortalidade para todas as mulleres, cribadas e non cribadas, obtendo así as taxas de mortalidade esperadas (figura 2) e número de mortes esperadas (a modo de exemplo, na táboa 3 se presentan os resultados para o período 2007-2012).

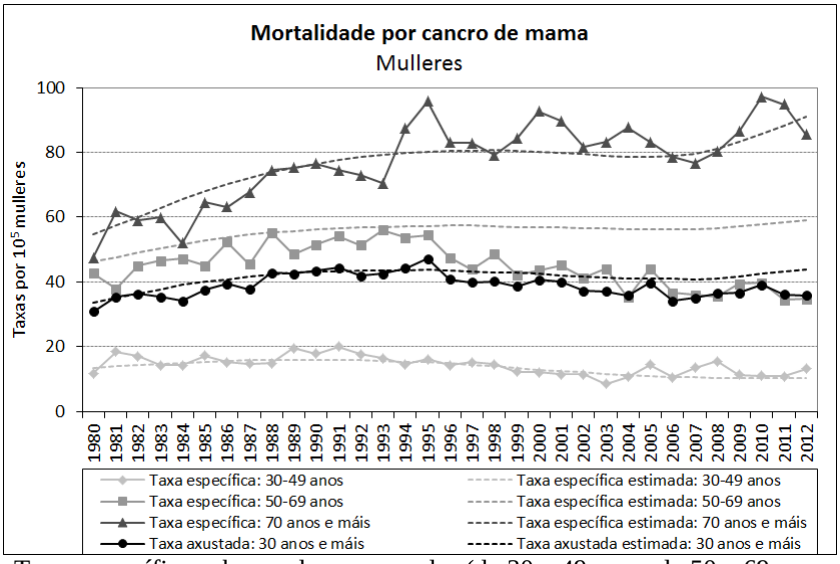


Figura 5. Taxas específicas observadas e esperadas (de 30 a 49 anos, de 50 a 69 anos e de 70 anos e máis) e taxas axustadas observadas e esperadas (poboación estándar: Censo de Galicia de 2001) por 100.000 mulleres para a mortalidade feminina por cancro de mama. As taxas esperadas obtivéronse a partir do axuste do modelo no grupo de mulleres non cribadas.

	30-49 anos	**50-69 anos**	**70 anos e máis**		
	Non cribado	**Cribado≥50%**	**Non cribado**	**Cribado<50%**	**Cribado≥50%**
Def obs	324	770	975	191	303
Def esp	264,8	1.209,1	872,2	196,2	367,9
DA	59,2	-439,1	102,8	-5,2	-64,9
DR	22,4	-36,3	11,8	-2,7	-17,6

Táboa 3: Número de defuncións observadas (def obs) e esperada (def esp) e diferenzas absolutas (DA) e relativas (DR) para o período 2007-2012, por grupo de idade e grupo de cribado.

Con respecto ás taxas de mortalidade, chama a atención o comportamento das taxas específicas no grupo de mulleres de 50 a 69 anos (poboación diana do PGDPCM) xa que, a partir de mediados dos anos 90, a taxa esperada mantense estable con valores lixeiramente por debaixo dos 60 casos por 100.000 mulleres, mentres que as taxas observadas diminúen ata acadar valores entorno aos 35 casos por 100.000.

Con respecto ao número de mortes esperadas, para as mulleres cribadas no grupo de 50 a 69 anos no período 2007-201 a diferenza relativa foi de -36,3% o que representa unha diminución do 36,3% dos casos observados fronte ao que cabería esperar en ausencia de cribado (é dicir, 439 mortes menos das que cabería esperar).

4. CONCLUSIÓNS

Este traballo pon de manifesto unha redución do risco de morrer por cancro de mama nos grupos cribados e do número de mortes observadas nas mulleres de 50 a 69 anos, en comparación coas que cabería esperar en ausencia dun programa de cribado. Este efecto podería estar asociado a un impacto positivo do PGDPCM aínda que non é posible diferencialo totalmente doutros efectos, como poden ser as melloras nos avances terapéuticos, por non dispoñer de datos.

REFERENCIAS

[1] International Agency for Research on Cancer (IARC) (2002). Breast Cancer Screening. 1st ed. Lyon (France): IARC Press.

[2] Castells X, Sala M, Ascunce N, Salas D, Zubizarreta R, Casamitjana M, coordinadores (2007). Descripción del cribado del cáncer en España. Proyecto DESCRIC. Madrid: Plan de Calidad para el Sistema Nacional de Salud. Ministerio de Sanidad y Consumo. Agència d'Avaluació de Tecnologia i Recerca Mèdiques de Cataluña. Informes de Evaluación de Tecnologías Sanitarias, AATRM núm. 2006/01.

[3] Dirección Xeral de Innovación e Xestión da Saúde Pública (2012). Programa galego de deteccion precoz do cancro de mama (PGDPCM). Resultados 1992-2012. Xunta de Galicia. Consellería de Sanidade.

[4] Council of the European Union ()2003. Council Recommendation of 2 December 2003 on Cancer Screening (2003/878/EC). Off J Eur Union, 327, 34-38.

[5] Perry N, Broeders M, de Wolf C, Tornberg S, Holland R, von Karsa L, et al (2006). European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition. In: European Comission, ed.: Office for Official Publications of the European Communities, Luxembourg.

[6] Béatrice Lauby Secretan, Chiara Scoccianti, Dana Loomis, Lamia Benbrahim Tallaa, Véronique Bouvard, Franca Bianchini,and Kurt Straif, for the International Agency for Research on Cancer Handbook Working Group. Breast-Cancer Screening (2015) — Viewpoint of the IARC Working Group. N Engl J Med 372;24.

[7] M G Marmot, D G Altman, D A Cameron, J A Dewar, S G Thompson, M Wilcox (2013)– The Independent UK Panel on Breast Cancer Screening. The benefits and harms of breast cancer screening: an independent review. A report jointly commissioned by Cancer Research UK and the Department of Health (England). October 2012. British Journal of Cancer, 108, 2205-2240.

[8] Carstensen B (1998). Age-period-cohort models for the Lexis diagram. Stat Med, 26, 3018-45.

[9] Armitage P (1998). Age-period-cohort analysis. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons, pp. 82-99.

[10] Sasieni PD (2012). Age-period-cohort models in Stata. Stata Journal, 12, 45-60.

[11] Clayton D, Kaldor J (1987). Empirical Bayes Estimates of Age-standardized Relative Risks for Use in Disease Mapping. Biometrics, 43, 671-81.

[12] Besag J, York J, Mollie A (1991). Bayesian image restoration, with applications in spatial statistics (with discussion). Ann Inst Math, 43, 1-59.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

MODELO LOGÍSTICO MULTINOMIAL: IDENTIFICAÇÃO DE FACTORES ASSOCIADOS À MUDANÇA NA ADESÃO À TERAPÊUTICA ANTI-HIPERTENSORA

Helena Mouriño[1], Sara Rosa[1] e Milene Fernandes[2]

[1] Faculdade de Ciências, Universidade de Lisboa.
[2] ISAMB/IMPSP, Faculdade de Medicina, Universidade de Lisboa.

RESUMO

Introdução: A Hipertensão Arterial (HTA) é uma doença crónica determinada por elevados níveis de pressão sanguínea nas artérias, o que leva o coração a efectuar um esforço adicional para realizar a circulação através dos vasos sanguíneos. Estima-se que a prevalência da HTA na população portuguesa seja de 42,2%, para os adultos com idade compreendida entre os 18 e 90 anos (Polonia *et al*., 2014). Apesar de as doenças cardiovasculares corresponderem à maior causa de morte em Portugal, o controlo da HTA ainda está longe de ser eficaz uma vez que as estimativas apontam para que 44,4% dos hipertensos medicados não alcancem os níveis de pressão arterial considerados normais (Polonia *et al*., 2014). A não-adesão à terapêutica e/ou a inércia terapêutica são apontados como dois dos principais factores de não-controlo da HTA.

Material e Métodos: O presente estudo está inserido no projecto DIMATCH-HTA (Determinants and Impact of Medication Adherence and Therapeutic Change in the Control of Arterial Hypertension among cohorts of immigrants and non-immigrants at the primary care level), conduzido pelo Instituto de Medicina Preventiva e Saúde Pública da Faculdade de Medicina, Ulisboa (PTDC/SAU-ESA/103511/2008). Neste trabalho, pretende-se identificar os factores que influenciam as mudanças de comportamento na adesão à medicação, no que diz respeito à HTA. O estudo longitudinal baseou-se na aplicação de questionários presenciais a doentes hipertensos, com idade compreendida entre os 40 e 80 anos e seguidos nos Cuidados de Saúde Primários da região de Lisboa. Os dados em análise foram recolhidos em dois momentos: Setembro de 2010 (período basal) e Março de 2011 (6 meses depois). O questionário engloba variáveis de natureza diversa, nomeadamente: socio-demográficas; clínicas; avaliação do controlo da HTA e caracterização do acesso aos cuidados de saúde primários. A classificação do paciente como "aderente", ou "não aderente" à medicação para a HTA, baseou-se na aplicação da escala de adesão de Morisky, auto-reportada (Morisky *et al*., 1986). Para identificar os factores que conduzem às mudanças de comportamento entre o período basal e passados 6 meses, procedeu-se ao ajustamento de um modelo de Regressão Logística Multinomial. Este modelo é adequado para descrever situações em que a variável resposta é nominal, com três, ou mais, níveis. Neste estudo, a variável resposta - construída com o objectivo de descrever a dinâmica de adesão, ou não, à medicação para o controlo da HTA nos momentos supra citados - foi categorizada em três níveis: "sempre aderente" (doentes que se mantiveram sempre aderentes); "melhor comportamento de adesão" (doentes que passaram de não aderentes, no basal, a aderentes ao fim de 6 meses) e "mudança/manutenção de não-adesão" (doentes que se mantiveram sempre não aderentes ou que passaram a não aderentes aos 6 meses).

Resultados: A amostra em estudo compreende 285 doentes hipertensos. Uma análise exploratória permitiu verificar que 28,8% dos doentes mantêm-se sempre aderentes à medicação, ao passo que 31,6% mantêm-se sempre não aderentes. Quanto às mudanças de comportamento, 27,1% dos doentes apresentam uma mudança positiva (i.e., passagem de não aderente para aderente), ao passo que para 12,5% dos doentes essa mudança é negativa. Ajustou-se um modelo de Regressão Logística Multinomial aos dados. A selecção das variáveis significativas para descrever o fenómeno em estudo foi realizada a partir do Método *Stepwise*. Com base no modelo estimado, verificou-se que existe uma

probabilidade estatisticamente significativa de os doentes que satisfaçam pelo menos um dos critérios a seguir apresentados tenham uma maior probabilidade não aderirem à medicação, quando comparados com os doentes sempre aderentes: ser de etnia negra (OR=4,25; *p-value*=0,002); não ter ajuda para controlar a HTA (OR=0,40; *p-value*=0,077); não ter recorrido ao farmacêutico (OR=0,38; *p-value*=0,018) ou ao enfermeiro (OR=0,20; *p-value*=0,057), no último ano, para controlar a HTA.

Conclusões: O presente estudo permitiu identificar as principais características dos doentes que conduzem à não-adesão à medicação para o controlo da HTA. Trata-se de uma ferramenta importante para que os profissionais de Saúde possam tomar medidas eficazes no combate à HTA não controlada, com particular atenção para o papel dos profissionais de saúde na promoção da adesão à medicação.

Palavras chave: Estudo DIMATCH. Hipertensão Arterial. Doenças Crónicas. Mudança na Adesão à Terapêutica. Regressão Logística Multinomial.

REFERÊNCIAS

Hilbe, J. M. (2009) Logistic Regression Models. Texts in Statistical Science. CRC Press, Boca Raton.

Hosmer, D. W., Jr., Lemeshow S. e Sturdivant, R. X. (2013) Applied Logistic Regression, Third Edition. John Wiley & Sons, Inc, New Jersey.

Polonia, J., Martins, L., Pinto, F. e Nazare, J. (2014) Prevalence, awareness, treatment and control of hypertension and salt intake in Portugal: changes over a decade. The PHYSA study. Journal of Hpertension, 32:6, 1211-1221.

Morisky D. E., Green L. W. e Levine D. M. (1986) Concurrent and Predictive Validity of a Self-Reported Measure of Medication Adherence and Long-Term Predictive Validity of Blood Pressure Control. Medical Care, 24, 67-74.

WHO (2003) Adherence to Long-Term Therapies. Evidence for action. World Health Organization, Geneva.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

RESTRICTING THE WEIGHTS AND THE JUDGMENT VALUES IN DEA MODELS

Pilar Alberca[1] , Jorge Santos [2]

[1] Departamento de Economía de la Empresa y Contabilidad , UNED, palberca@cee.uned.es
[2] Departamento de Matemática, CIMA, ECT, Universidade de Évora, jmas@uevora.pt

ABSTRACT

This paper analyzes the operational efficiency of the Spanish hotel industry by Data envelopment Analysis (DEA) under a restricted weights modeling framework. The study develops a novel approach to hotel efficiency evaluation that aims to overcome some limitations of conventional DEA models, specifically that several of the units assessed only considers one or some variables instead of all of them. This problem can be solved by applying weights restrictions in order to incorporate judgment values and increase the discriminating potential of the conventional DEA models.

Key words: Efficiency assessment, DEA models, weight restrictions; hotel firms; decision making process.

1. INTRODUCTION

A matter of considerable interest in production theory and economics studies is performance and efficiency assessment. Within the tourism and services industry the main literature (Barros, 2005; Botti, et al., 2009; Chen, 2007; Haugland et al., 2007; Huang et al., 2014) has evaluated the efficiency of firms mainly applying nonparametric frontier methodologies such as DEA (Data envelopment analysis).

DEA frontier modeling is applied in the field of studies on frontier assessment alongside the stochastic frontier models, and both models are appropriate tools in analyzing operational efficiency. However, DEA is a tool of choice when the research has doubts as to which functional form to adopt (Barros and Dieke, 2008). The DEA models have been applied also as the benchmarking tool in operations management to empirically account operational and productive efficiency. DEA also provides a set of improvements and targets that the firms could apply in the first instance. The main advantage of the DEA method is the flexibility and the possibility to identify the "peers" for firms that are not observed to be efficient. However, DEA is based on some restrictive assumptions: the score of efficiency is a relative value to best practice within the particular sample and the efficiency scores are very sensitive to the inputs and outputs specification and size of the sample.

Despite these limitations, DEA is a useful tool for examining the efficiency of manufacturing firms and service companies and this nonparametric frontier analysis offers significant advantages and benefits to increase our understanding about the potential areas of improving performance and efficiency of companies and services providers.

The determination of the weights without the need for a prior knowledge is the main advantage of the nonparametric frontier DEA methodology and its distinctive feature. In the conventional DEA method, weights are unrestricted, however a potential problem when solving conventional DEA models without restrictions on the weights is that some inadequate and non-rational weights can be assigned to inputs or outputs. An absolute flexibility of the weights can cause problems in many studies and some inputs or outputs can be characterized by excessively high and/ or low values despite the decision maker knows that weights belong to some predetermined intervals. In the limit, this can result in some DMUs assigning zero weights to important input measures and effectively ignoring them. Furthermore, the arbitrary values assigned to the weights may be contradicting with prior information of the study and the decision maker.

To address these limitations related to the weights unrestricted in standard DEA models, several researches have applied different techniques (Allen et al. 1997; Dimitrov and Sutton, 2010; Dyson and Thanassoulis, 1988; Roll and Golany, 1993; Takamura and Tone, 2003; Thompson, et al., 1990; Wong and Beasley, 1990). In this study we apply DEA frontier methodology restricting weights values in DEA models to integrate the information of the researcher in the performance and the efficiency evaluation in order to improve the process and the final conclusions analyzing a large sample of Spanish hotel companies .

This rest of the paper is organized as follows. Section 2 presents a brief review of the literature. Section 3 describes the sample and provides the descriptive statistics. Section 4 presents the research design of the study, specifically, the empirical models. Section provides 5 the results of the empirical models and, finally, section 6 discusses the results and presents the overall conclusions.

2. LITERATURE REVIEW

The scientific literature in the field of efficiency and performance evaluation has conducted a wide variety of theoretical and empirical models and applications into several countries and services industries as tourism, banks or franchises, among others.

Previous studies in the tourism industry and hospitality literature assess the efficiency using frontier methodologies (Barros, 2005; Chiang et al. 2004; Haugland et al. 2007; Huang et al., 2014; Manasakis, 2013; Shang et al. 2010; Sanjeev, 2007). Some of these studies have analyzed the activities that are most efficient. For example, Huang et al. (2014) examine the hotel activities applying two-stage methodology in order to assess the efficiency and their determinants. The study analyzes the occupancy activity and catering activity to achieve a better measure of value-added chain and resource allocation in hotel firms.

Several of previous researches analyzes the efficiency with basic DEA model using one stage methodological (Chiang et al. 2004; Sanjeev, 2007; Manasakis, 2013). Others studies apply methods based on several stages and combine conventional DEA model with some complementary analysis. For example Barros and Alves, (2004); Hwang and Chang, (2003) and Barros (2005) apply the Malmquist index measures in order to assess total factor productivity. Other studies combine DEA models with Tobit regression model (Haugland, 2007; Huang et al., 2012) and a bootstrapping procedure (Barros et al. 2011a; Barros et al. 2011b; Shang et al. 2010).

However, the empirical researches in the tourism and hospitality sector often measure the efficiency based on the standard DEA models, and then, they allow the flexibility of the weights for (DMUs) under evaluation. The wide flexibility in assigning the weights in DEA approach can result in indicators of efficiency who do not take account the relative importance of some inputs. In order to overcome this limitation, in this research we apply the DEA model under restricted weight specification. This model is applied to Spanish hotel companies in order to measure operational efficiency. The restricted weight specification enables us to decrease the influence of assigning unrealistic weights in some units and improve the efficiency estimation and to increase the discriminating potential of the conventional DEA model.

The use of the restriction weights and the value judgments in the DEA methodology was applied by several studies in the efficiency assessment and performance evaluation field in order to account the information preferences in the decision maker process. In the DEA methodology framework, the weights restrictions can be divided in two broad categories: relative and absolute weights restrictions. The absolute weights restrictions are typically imposing a range for an individual weight. This approach was developed by Dyson and Thanassoulis (1988) and generalized in Roll, Cook and Golani (1991). Virtual weights restrictions introduced by Wong and Beasley (1990) belong to this category too. The class of relative weights restrictions includes, among others, the assurance region models of type I or II introduced by Thompson, Singleton, Thrall and Smith, (1986) and (1990) respectively, as well as cone ratio DEA models from Charnes, Cooper, Wei and Huang (1989).

Thompson, Singleton, Thrall and Smith, (1986) applied weight restrictions and they compared the results with the standard DEA model. Dyson and Thanassoulis (1988) assess the result of total weight unrestricted in standard DEA models and suggest a possible way of to limit this flexibility. Roll, Cook, and Golany (1991) offer an alternative version of the DEA model imposing boundaries and restricting the weights. The study also proposes methods for defining appropriate bounds. Roll and Golany (1993) develop a conceptual and theoretical framework for the treatment of weights restrictions in DEA

methodology and present alternative methods in order to limit the range of variation and the flexibility of the weights.

Extending this literature, in this paper we assess the efficiency in the hotel companies imposing restrictions on the weights in order to incorporate the preferences of the decision marker in the evaluation process.

DEA is a linear programming method based on the theoretical production process. DEA is a nonparametric frontier approach and a mathematical optimization tool which determines the efficiency of each DMU by maximizing the ratio of a weighted sum of its outputs to a weighted sum of its inputs. For any set of inputs, the value of the DEA scores defines the maximum output producible, and on the other hand, for any outputs, the value of the efficiency estimated by DEA method defines the minimum input producing a given output under all circumstances.

In this study the selection of inputs (resources or costs) and outputs (goods or services) is based according to a generalized Cobb-Douglas function. Therefore, the inputs resources for hotel management include human resources, capital, operating costs, and equipment (Barros, 2005b; Barros et al., 2011). Based on previous researches on the hotel industry, the DEA model is applied using the following input variables: the number of full-time employees to represent manpower, the book value of a property to represent capital investment of hotels and operational cost to represent the cost of inputs. In the empirical framework of the hotel sector, the output is a variable relatively easy to account, for example, as room nights or sales revenue. In this study we use sales as the output measure. Sales are a common variable that is used as the output measure because it represents the value of a firm's achievement of its goals (Barros, 2004; Barros and Alves, 2004; Barros, 2005a; Haugland et al., 2007).

3. SAMPLE AND DESCRIPTIVE STATISTICS

The financial data were collected from the SABI (Sistema de Análisis de Balances Ibéricos) database. The sample includes data on 572 audited hotel firms for the period 2011. Table 1 provides some descriptive statistics for the input and output variables used in this study.

		SALES	CAPITAL	NUMBER OF EMPLOYEES	OPERATIONAL COST
N	Valid	572	572	572	572
Mean		13253	46619	170	2547
Std. Deviation		30297	144227	434	5118
Skewness		11	12	10	7
Std. Error of Skewness		,102	,102	,102	,102
Minimum		18	217	2	5
Maximum		516160	2451857	6170	59017
Percentiles	10	2873	4242	31	258
	25	4512	8173	55	654
	50	7058	18267	86	1215
	75	12945	41535	156	2423
	90	23899	85197	292	5399

Table 1: Main descriptive statistics for the input and output variables.

We conclude that all variables are positively skewed at usual levels of significance, we can confirm this fact by an inspection of figure 1, where we present Boxplots for the 4 variables scaled down, so that they can fit in a single graphic.

From this descriptive statistical analysis we see that there is a great range of variation of the data, especially in the Number of Employees variable, that spreads from a minimum of 2 to a maximum of 6170 so we will use the Variable Returns to Scale model introduced by Banker Charnes and Cooper in 1984 (Banker et al. 1984).

From figure 1 we see that DMU 406 has the biggest values on 3 Variables and ranks 3rd in Operational Cost. Since it is the unit with biggest output, it will be efficient under the BCC model.

Unit 502 ranks first in the input Operational Cost and is among the 3 highest values on the remaining 3 variables.

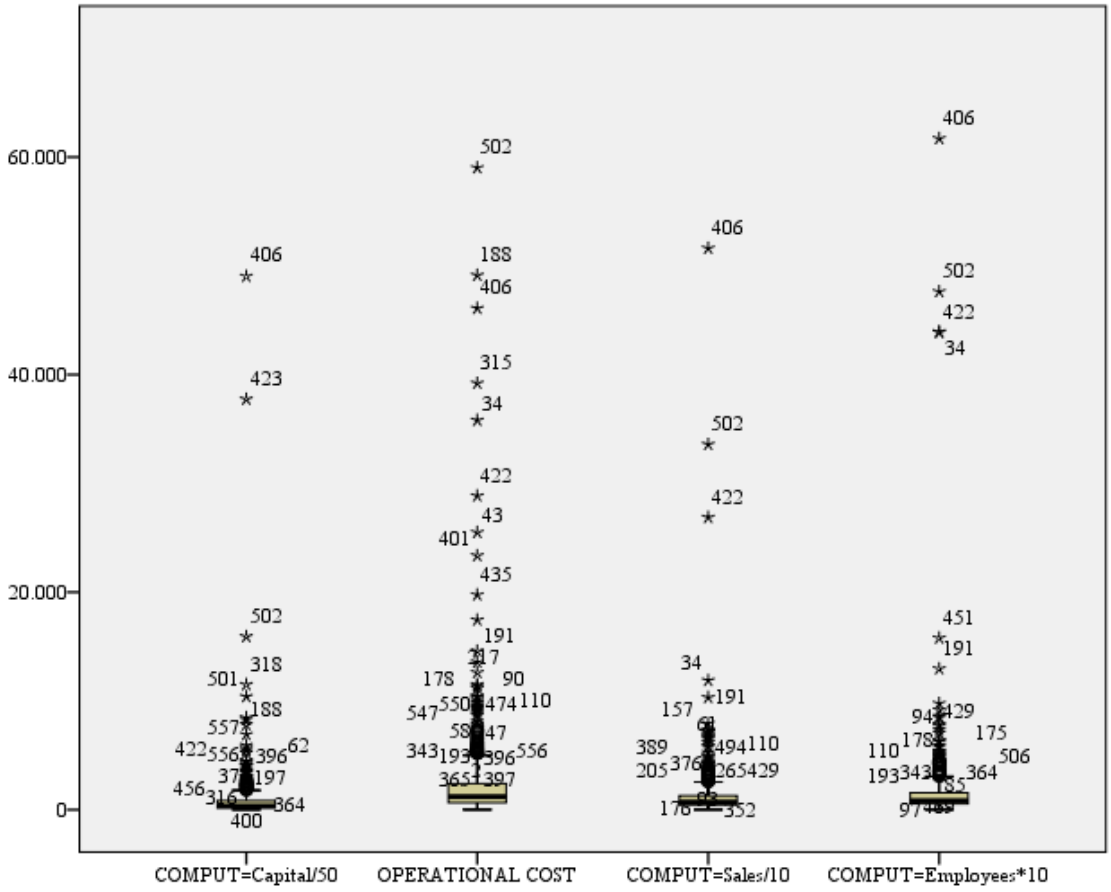


Figure 1 Boxplots for the 4 variables (Employees scaled up; Sales and Capital scaled down)

4. RESEARCH DESIGN AND METHODOLOGY

The DEA methodology uses some linear programmes to construct a piecewise linear production frontier and to compute an efficiency index relative to that frontier. The DEA approach assumes that all firms within a sample have access to the identical technology for transforming a vector of M inputs, denoted by X, into a vector of K outputs, denoted by Y. We assume that technology can be characterised by the technology set, T, defined as follows:

$$T = \left\{ (X, Y) \in \Re_+^M \times \Re_+^K : X \in \Re_+^M \, can.produce.. Y \in \Re_+^k \right\}$$

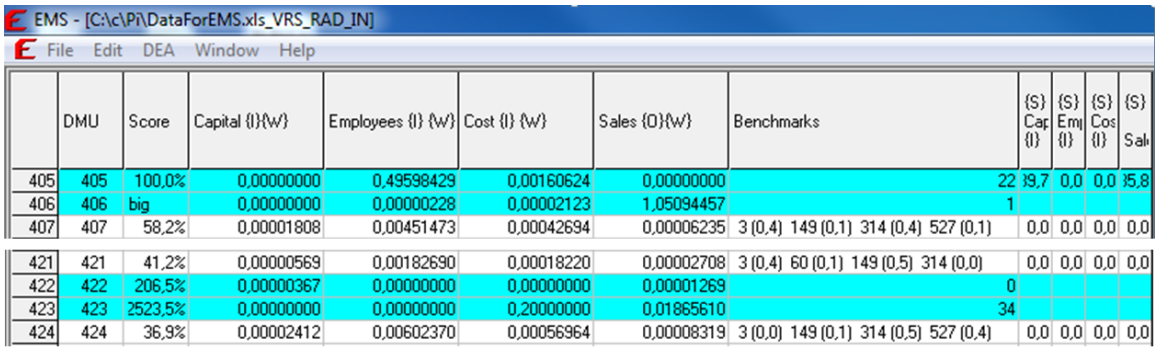
The original DEA model, which originated in a seminal paper from Charnes et al. (1978), evaluates the Debreu-Farrell measure with a hypothesis regarding constant returns to scale (CRS). This original DEA has been extended by Banker, Charnes and Cooper (1984), who proposed variable returns to scale to enable the estimation of pure efficiency and, therefore, the deduction of scale efficiency effects. The variable return efficiency (VRS) score represents pure technical efficiency, a measure of efficiency without scale efficiency. The difference between the overall and pure efficiency measures provides a measurement of the scale efficiency of a productive unit.

The model proposed by Banker et al. (1984) assumes variable returns to scale between inputs and outputs to calculate efficiency. The efficiency score that is obtained with the VRS model provides a score that is at least equal to the score that is obtained using the CRS method. The scale efficiency score is obtained by dividing the CRS efficiency score by the pure technical efficiency VRS score.

5. RESULTS

As explained previously we choose the Variable Returns to Scale model. A careful and detailed analysis of the weights leads us to the conclusion that some DMUs are ignoring some inputs, especially the very efficient ones, specifically DMU 422 only considers Capital while DMU 423 only considers Operational Cost, as can be confirmed by figure 2.

We see that DMU 405 and 406 are ignoring how much capital they spent, so we decided to apply weights restrictions as suggested for the first time by Thompson et al. in 1986.


EMS - [C:\c\Pi\DataForEMS.xls_VRS_RAD_IN]

File Edit DEA Window Help

	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	39,7	0,0	0,0	35,8
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 2: Sample of some results that led us to apply weights restrictions..

To avoid a given DMU "to choose" a rather unbalanced set of weights (as is the case of units 422 and 423 which ignore two inputs it is current practice to place some restrictions on the weights or in the virtual inputs/outputs. This is a usual way to incorporate judgment values and increase the discriminating potential of the model.

In this case since we have a single output we can make a linear multiple regression so that we can get some estimates for the weights.

Coefficients[a]

Model	Unstandardized Coefficients		t	Sig.	99,0% Confidence Interval for B		Collinearity Statistics	
	B	Std. Error			Lower Bound	Upper Bound	Tolerance	VIF
(Constant)	195	376	,51	,604	-777,5	1168,35		
Input Capital	,057	,003	19	,000	0,05	0,06	,621	1,61
Nr of Employes	42	1,24	34	,000	38,74	45,16	,389	2,56
Operational Cost	1,287	,096	13	,000	1,04	1,54	,470	2,12

a. Dependent Variable: OUTPUT (sales)

Table 2 Results for the multiple regression.

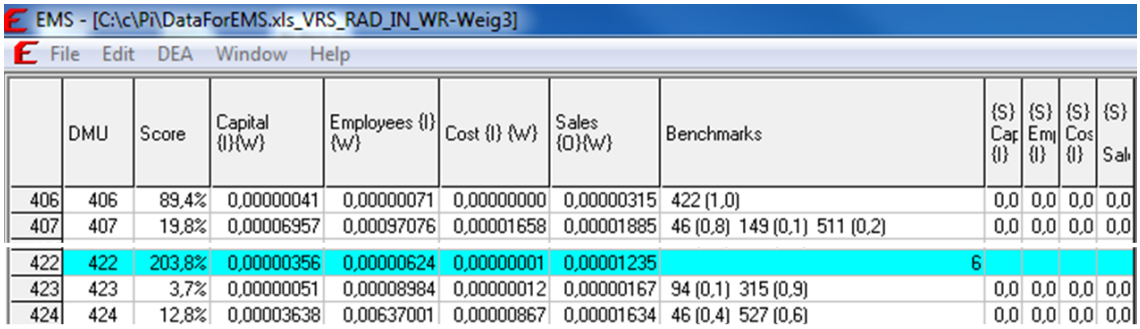
We performed this multiple regression in the SPSS software and got very significative results for all coefficients except for the intercept. Since the intercept is not significative we will not subtract it from the values of the output variable, neither perform another regression without constant, because the results would be almost the same.

This way we established weights restrictions on the basis that the weights would be within the interval [B/10;10B] where B is the respective coefficient taken from table 2.

The new results will lead to lower results of efficiency, having an increased discriminatory power and we see the following: 11 DMUs kept its status of being efficient: 527 511 435 422 315 213 149 125 94 46 34 while the following 21 lost it, changing from efficient to inefficient: 502 423 406 401 321 314 268 244 212 204 194 188 187 181 133 90 65 62 60 43 3.

DMU 149 had an efficiency of 186.4% and was a benchmark for 498 inefficient DMUs and with this last model both figures dropped to 162.7% and 22 respectively.

DMU 46 kept its efficiency of 147,4%, but increased its importance since previously only 29 DMUs were using it as a reference and now this number increased to 464.


EMS - [C:\c\Pi\DataForEMS.xls_VRS_RAD_IN_WR-Weig3]

File Edit DEA Window Help

	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
406	406	89,4%	0,00000041	0,00000071	0,00000000	0,00000315	422 (1,0)	0,0	0,0	0,0	0,0
407	407	19,8%	0,00006957	0,00097076	0,00001658	0,00001885	46 (0,8) 149 (0,1) 511 (0,2)	0,0	0,0	0,0	0,0
422	422	203,8%	0,00000356	0,00000624	0,00000001	0,00001235	6				
423	423	3,7%	0,00000051	0,00008984	0,00000012	0,00000167	94 (0,1) 315 (0,9)	0,0	0,0	0,0	0,0
424	424	12,8%	0,00003638	0,00637001	0,00000867	0,00001634	46 (0,4) 527 (0,6)	0,0	0,0	0,0	0,0

Figure 3: Sample of some results after applying weights restrictions.

When we compare figure 2 with figure 3 it is noticeable that efficiencies with the weight restrictions are lower and we achieved a greater homogeneity of the weights as desired. DMU 406 is not efficient anymore, and its peer group (benchmarks) is DMU 422 who kept its efficient status, despite the aplication of weights restrictions.

6. CONCLUSIONS AND IMPLICATIONS

This paper incorporates the weights restrictions on the conventional DEA methodology. Although restrictions weights involves the formulation of value judgments, the inexistence of restrictions on the weights can result in the exclusion of inputs and relevant factors in the efficiency assessment.

In this study we have demonstrated that this methodology improves the estimation of efficiency compared with the conventional DEA methodology

REFERENCES

Allen, R., Athanassopoulos, A., Dyson, R. G. and Thanassoulis, E. (1997) Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. Annals of Operations Research, 73, 13-34.

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.

Barros, C. P. and Dieke, P. U. (2008) Technical efficiency of African hotels. International Journal of Hospitality Management, 27(3), 438-447.

Barros, C.P. (2004) A stochastic cost frontier in the Portuguese hotel industry. Tourism Economics, 10(2), 177–192.

Barros, C.P. (2005) Evaluating the efficiency of a small hotel chain with Malmquist productivity index. International Journal of Tourism Research, 7(3), 173–184.

Barros, C.P. and Alves, F.P. (2004) Productivity in the tourism industry. International Advances in Economic Research, 10(3), 215–225.

Barros, C.P., Botti, L., Peypoch, N. and Solonandrasana, B. (2011a) Managerial efficiency and hospitality industry: the Portuguese case. Applied Economics 43(22), 2895-2905.

Barros, C.P., Botti, L., Peypoch, N., Robinot, E., Solonandrasana, B. and Assaf, G. (2011b) Performance of French destinations: Tourism attraction perspectives. Tourism Management, 32(1), 141-146.

Botti, L., Briec, W. and Cliquet, G. (2009) Plural forms versus franchise and company-owned systems: a DEA approach of hotel chain performance. Omega, 37(3), 566–578.

Charnes, A., Cooper, W. W., Wei, Q. L. and Huang, Z. M. (1989) Cone ratio data envelopment analysis and multi-objective programming. International Journal of Systems Science, 20(7), 1099-1118.

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.

Chen, C.F. (2007) Applying the stochastic frontier approach to measure hotel managerial efficiency in Taiwan. Tourism Management 28(3), 696–702.

Chiang, W., Tsai H. and Wang L. (2004) A DEA evaluation of Taipei hotels. Annals of Tourism Research, 31(3), 712-715.

Dimitrov, S. and Sutton, W. (2010) Promoting symmetric weight selection in data envelopment analysis: A penalty function approach. European Journal of Operational Research, 200(1), 281-288.

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

Haugland, S.A., Myrtveit, I., and Nygaard, A. (2007). Market orientation and performance in the service industry: A data envelopment analysis. Journal of Business Research, 60(11), 1191-1197.

Huang, C. W., Ho, F. N. and Chiu, Y. H. (2014) Measurement of tourist hotels′ productive efficiency, occupancy, and catering service effectiveness using a modified two-stage DEA model in Taiwan. Omega, 48, 49-59.

Huang, Y., Mesak, H., Hsu, M.K. and Qu, H. (2012) Dynamic efficiency assessment of the Chinese hotel industry. Journal of Business Research, 65(1), 59-67.

Hwang, S. and Chang, T. (2003) Using data envelopment analysis to measure hotel managerial efficiency change in Taiwan. Tourism Management, 24(4), 357–369.

Manasakis, C., Apostolakis, A. and Datseris G. (2013) Using Data Envelopment Analysis to measure hotel efficiency in Crete. International Journal of Contemporary Hospitality Management, 25(4), 510-535.

Roll, Y. and Golany, B. (1993) Alternate methods of treating factor weights in DEA. Omega, 21(1), 99-109.

Roll, Y., Cook, W. D. and Golany, B. (1991) Controlling factor weights in data envelopment analysis. IIE transactions, 23(1), 2-9.

Shang, J.K., Wang, F.C. and Hung, W.T. (2010) A stochastic DEA study of hotel efficiency. Applied economics, 42(19), 2505-2518.

Sanjeev, G.M. (2007) Measuring efficiency of the hotel and restaurant sector: The case of India. International Journal of Contemporary Hospitality Management, 19(5), 378-87.

Takamura, Y. and Tone, K. (2003) A comparative site evaluation study for relocating Japanese government agencies out of Tokyo. Socio-Economic Planning Sciences, 37(2), 85-102.

Thompson, R. G., Singleton Jr, F. D., Thrall, R. M. and Smith, B. A. (1986) Comparative site evaluations for locating a high-energy physics lab in Texas. Interfaces, 16(6), 35-49.

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.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

A METHOD FOR PROCESSING PERCEPTUAL DIALECTOLOGY DATA

Laura Calaza Díaz[1], Soraya Suárez Quintas[2], Rosa M. Crujeiras [1], Alberto Rodríguez Casal[1], Xulio Sousa[2] e José Ramón Ríos Viqueira[3]

[1]Departamento de Estatística e Investigación Operativa. Universidade de Santiago de Compostela.
[2]Instituto da Lingua Galega. Universidade de Santiago de Compostela.
[3]COGRADE. CITIUS. Universidade de Santiago de Compostela.

ABSTRACT

This work presents a contribution to the question of how to develop data processing (both technically & statistically) in perceptual dialectology (PD) studies. This discipline focuses on the study of non-linguists perceptions of dialectal variation. Such tasks require a large amount of information whose treatment can be improved if suitable tools from Geographic Information Systems (GIS) for the treatment of geographic information are combined with powerful statistical data analysis methods. In order to carry out this work, simulated data, corresponding to a case study from PD research from Galicia, has been used to demonstrate the procedure.

Keywords phrases: Baddeley's distance, dialectometry, perceptual dialectology, QGIS.

1. MOTIVATION

The *Tecnoloxías e Análise dos Datos Lingüísticos* (Technologies and Analysis of Linguistic Data) network[1] was created to encourage collaborative work among linguistic, statistical and computational scientific fields. The interdisciplinary work among these areas arises from the firm desire to promote iniciatives to strengthen knowledge transfer and research results.

In particular, during the last decade special efforts have been made in the field of Perceptual Dialectology (PD), a discipline regarded as a subset of folk linguistics and related to dialectology and sociolinguistics. PD research consists of an analysis of non-linguists knowledge, beliefs and attitudes about dialectal variation. This kind of research has aroused the interest of many dialectologists worldwide. However, there is still no detailed analysis of the perceptions of non-linguists concerning dialectal variation in Galicia. Therefore, our study, which intends to discover and analyse Galician speakers beliefs and opinions regarding dialectal variation, aims to fill a lacuna in Galician language studies. Specifically, our PD survey attempts to:

- Discover if people are aware of and recognize regional variations of the Galician language.
- Identify factors which influence geographical varieties of Galician language recognition.
- Assess in which way peoples perceptions correspond to the geo-linguistic varieties traditionally recognised in Galician studies (Fernández Rei, 1990; Sousa, 2006).

The main goals of this work are therefore twofold: the identification and determination of geographical positions of different Galician dialects, and of subjective judgments regarding the different features each dialectal variety has.

[1]TecAnDaLi (`http://ilg.usc.es/tecandali/`)

This work was partly funded by TecAnDaLi (which is co-funding by the European Regional Development Fund (FEDER) under the Galician Operational Programme 2007-2013).

Such tasks require information collection about respondents in a defined geographic region. We must therefore be able to produce a large database, which collects, among other information, respondents' perceptions maps. This means that data from PD studies must be processed in a computerized way. In addition, this type of studies requires intensive field work which in turn requires that information be gathered in a cost-efficient, practical and portable manner.

The main aim of this communication is to present techniques which allow for the automated collection of PD data and to introduce statistical tools in the PD field which enable the realization of proper statistical analysis.

2. METHODS

2.1. *Theoretical foundation*

PD is a linguistic discipline that focuses on the study of how speakers perceive variation in language. Unlike traditional dialectology, which is almost exclusively concerned with identifying and analysing the variables and variants that define the geographical varieties of a language, this area of study focuses on analysing the way in which people perceive geo-linguistic variation. What is therefore interesting about PD is its focus on non-linguists' insights into language (folk linguistics). Studies within this framework reverse traditional roles, bringing to the foreground speakers knowledge about language variation in the spatial dimension.

PD studies began in the early twentieth century in Holland, where the first dialect maps based on speakers' geolinguistic perceptions appeared. Until the mid-twentieth century, this country and Japan played the main role in studies within this field. However, it was not until the 1980s that this discipline was developed. At that time, Dennis Preston (1989) reviewed and modernized the study of speakers perceptions on geolinguistic variation, redefining the field and adapting it to modern linguistics. Since Preston, who is regarded today as the most representative figure of perceptual dialectology, many studies have emerged which are inspired by some of the five techniques developed by him (Preston, 1999) for PD studies:

- Draw-a-Map: informants have to draw borders in a blank map, identifying the locations where they believe different dialects exist.
- Degree of Difference: speakers are asked to rate the similarity or difference between two dialectal regions.
- Correct and Pleasant: informants rate regions according to how correct or pleasant they think the variants spoken there are.
- Dialect identification: informants listen to recorded speech samples from a given dialect continuum and must assign each voice to a specific territorial area.
- Qualitative data: respondents must answer a series of open questions related to the identifying features of the different dialects they recognized, connected with the characteristics that they attribute to the speakers of each one of those dialects.

In relation to these five methodological possibilities, there are at least two basic lines of PD work. One of the alternatives focuses on a draw-a-map task for dialectal identification in order to detect places where people recognize different perceptual varieties (geographical perspective), whilst the other approach considers qualitative and quantitative descriptive analysis (avaliative perspective). Traditionally, they are treated separately. Therefore, the innovative aspect of our study lies in the combination of the methods offered by Preston, with the adjustments employed by Montgomery (2007, 2010), placing at its core a perceptive test to study Dialect identification by using Draw-a-Map. In addition, quantitative analysis will be supported by recent statistical methods with a geometrical flavour. This procedure enables the diagnosis and inference of PD data beyond a simple descriptive analysis based on numbers.

In that context, we must emphasize that the knowledge of speakers' (non-linguists') opinions on dialectal variation is of utmost importance both for the study of variation and linguistic change as for language planning. On the one hand, the research into attitudes and perceptions can help us achieve a better understanding of language variation, change, maintenance and decay. On the other, it helps to understand certain attitudes and behaviours and promotes the implementation of language planning adapted to the needs of a particular speech community.

In a situation like that of Galician, characterized by the contact of two languages with unequal status, the importance of discovering speakers opinions, beliefs and perceptions in relation to the language variety they speak and those spoken by others is, perhaps, even greater. In this sense, it is essential to supplement the lack of work carried out in this field so as to advance in knowledge and contribute to halting and even pushing back unfavourable dynamics for the language.

2.2. *Technical support*

The following points are not aiming at completeness, but rather provide an overview of technical support for PD. Firstly, data collection is presented, with the implementation of an app for this task. Secondly, we show how PD data can be treated and exploited by employing tools provided by a Geographic Information System (GIS). Finally, we illustrate the procedure followed in order to perform a proper statistical analysis for such data.

- **Data collection**
  An offine application was built to allow automated collection of data. The sample below illustrates its simple design (Figure 1) and how it edits information.

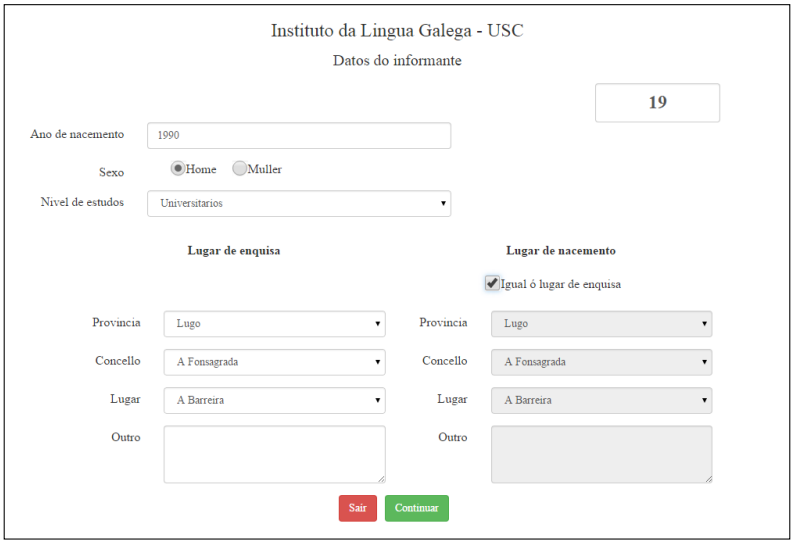


Figure 1: Front view of the data collection application.

For the purpose of our study, three main sections were included in the perceptive test:

1. A part to collect some sociodemographic variables.
2. A draw-a-map task to identify respondents' perceptions.
   The informants were presented with a variety of oral stimulation and asked for their perceptual judgments. Seven different audio tracks (six recordings of dialect speech and one of standard Galician) were played to each participant. In that application, they had a map of Galicia on which the seven biggest cities were marked (an example of a map is shown in Figure 2), and thus they were asked to assign each voice to a specific geographical area.
3. A questionnaire with long answers that must be processed for linguistics' experts, with questions regarding:

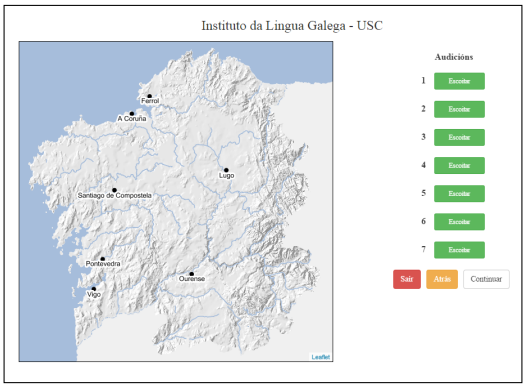


Figure 2: Map of Galicia with the seven biggest cities marked in the PD application.

- Their knowledge about dialect variation in the Galician language area.
- Their perceptions about how correct and pleasing the language varieties they identify are.
- The degree of difference or similarity between the various geographical varieties they recognize and their own variety.
- The identifying features of each of these dialects and the characteristics they attribute to their speakers.

This application allows all PD data to be stored in a relational database at the PostgreSQL (`http://www.postgresql.org/`). The storage mechanism was designed to link all the information within a Spatial Reference System. In particular, its outstanding contribution is made by the draw-a-map task (Figure 3), which allows this selection to be automatically translated into a polygon. It can therefore store the information given by respondents in a geometric pattern.

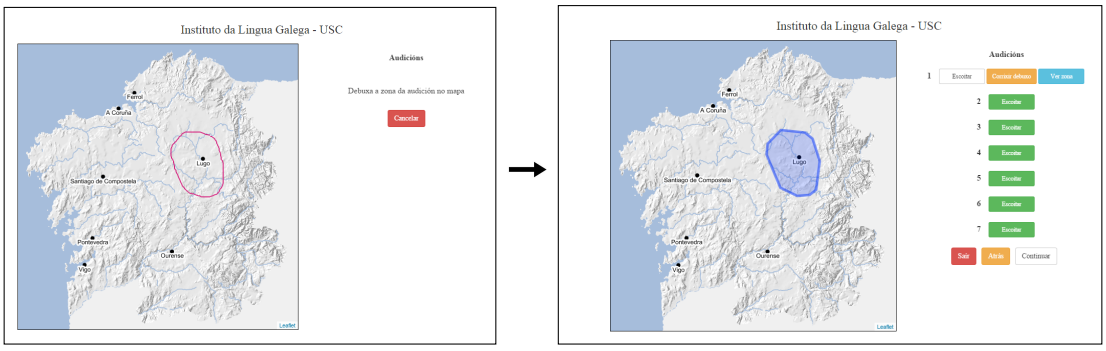


Figure 3: Draw a map example in the PD application.

- **Geographic information treatment**
  Recent research has focused on improving perceptual labels visualization (Cukor-Avila et al. 2012; Mongomery and Stoeckle, 2013). In line with these ideas, we use a Geographic Information System (GIS) for PD visualization. In particular, we use QGIS, an easy-to-use open source GIS application, ideal for data viewing, editing, and producing maps. For this reason, a clear useful approach is related to PD work in order to display geographical recognition of dialect areas. Up to this point, the method employed for processing data enables an aggregation of the draw-a-map by scanning and then incorporating it into a map as layers. However, one of the advantages of using QGIS is that it can be connected to a PostgreSQL/PostGIS database and allows us to view spatially any table containing a geometry column. As a result, we can simplify the accessibility of information.

- **Data analysis**
  A focus on perceptual data visualization means that there has been little sophisticated statistical analysis framework developed in this field. This examination has been so far limited to a qualitative analysis of open-ended answers, or to descriptive analysis of degrees of difference or even correctness and pleasantness of speech. This technical gap has reduced the analytical potential of the data supplied by respondents. Therefore, our aim is to carry out this study by taking advantage of the full potential of information in order to extend our knowledge of PD. Thus, we must distinguish two key points to be treated in our study.

  On the one hand, as mentioned previously, we must realize that we obtain information about locations and perceptual regions in a reference space. This geographical data must be properly processed. That is, it must be used to direct comparison with data from traditional dialectology and dialectometric studies, and therefore to identify which factors influence the recognition of dialectal varieties.

  In this way, for our propose we consider maps as images, and identify locations by pixels. Any region could be represented as a composition of pixels; that is, displayed as an image. Therefore, in order to compare the respondents perceptions with geo-linguistic varieties or establish differences between points and regions, the criterion used was a distance measured to assess the diferences between two images (Baddeley, 1992). Let X be a pixel raster (Galician map), with $A, B \subseteq X$ (two identified locations or regions). Then the Baddeley distance, Δ_b, is then defined as

$$\Delta_b(A,B) = \left[\frac{1}{n(X)} \sum_{x \in X} |d^*(x,A) - d^*(x,B)|^p \right]^{1/p},$$

  here $d^*(x,A) = \min\{d(x,A), c\} = \min\{\inf[d(x,a), a \in A], c\}$, where $d(x,x')$ is a metric defining 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.

  Hence, it is possible to create new variables which lead us to study the relation between locations and auditions' regions or selections, like measuring:

  – Distance between two points. For instance, between audition place and birthplace.
  – Distance between the audition place and respondents' selection.
  – Discrepances between respondents selection and dialectal varieties traditionally recognised in Galician studies.

  Once all the information was gathered, and apart from performing descriptive analysis and data visualization, we were also interested in making inferences regarding PD. In this way, having defined variables quantifying distances, we study which factors influenced in varieties recognition by using regression analysis.

  All statistical analysis were performed using `R` 3.1.3. (R development Core Team, 2015)[2]. PD treatment was conducted using `RPostgreSQL`, `rgeos`, `spatstat`, `maptools`, `rasterVis`, `MASS` packages.

3. METHOD ILLUSTRATION

[2]R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

In order to illustrate the overall process, we show an application of this methodology using simulated PD data. Six audition places, which represent different dialectal varieties in Galicia, were incorporated into the database in order of recording: Mazaricos, A Guarda, San Ramón, Xinzo de Limia, A Fonsagrada, and O Barco. A total of nine fictitious informants were included in order to exemplify the process (given that there is no interest on this data description, we only focus on the treatment overview).

- **QGIS** [3]
  Focusing on data visualization, we wish to give a basic outline here on what QGIS can be used for in relation to PD studies. As mentioned before, respondents selections were stored as a polygon, with the disadvantage that they can be off the map. Figure 4 shows a registered drawn map from our fictitious respondents.

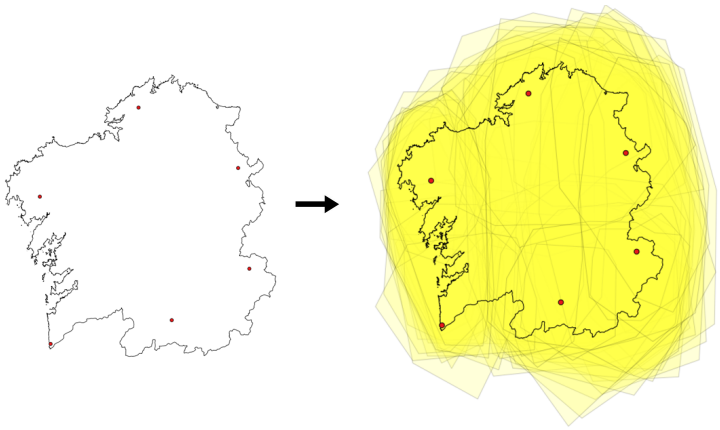


Figure 4: Map of Galicia including audition locations/ Map of Galicia including audition locations and respondents' selections.

This can be easily corrected by using a SQL editor which allows for the intersection between these polygons and the one that defines the Galician geometry.

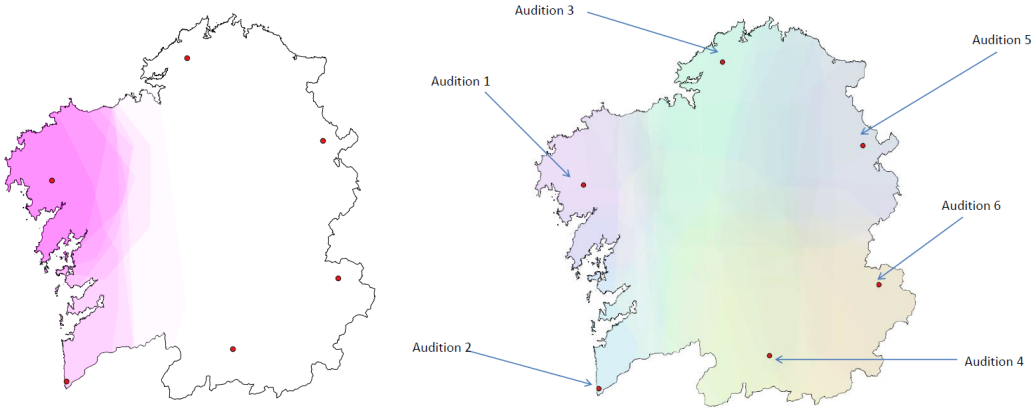
In order to obtain preliminary ideas of respondents' knowledge of dialectal varieties, respondents perceptions can be aggregated for each audition and a colour scale established according to selection frequencies; the color intensities show different degrees of agreement (Figure 5(a)). We can even merge all these representations in order to visualize the dialectal recognition of our respondents, with each color corresponding to each auditions answers 5(b)).

We must be aware that respondents' answers could overlap with different regions, so the dialectal varieties determined by linguists could be split or joined in different regions as in the example below (Figure 6). For instance, let the red area be the traditional linguistic variety for audition 3, the blue area one be the respondent answer to audition 1, and the green area, audition 3.

This enables us to configure new dialectal regions according to speakers opinions and perceptions, which can differ from linguistics studies, as can be seen in Figure 5(b)), which shows perceptual areas of our simulated PD study in Galicia, and in which we clearly differentiate a total of six perceptual dialect varieties.

- **R project**
  Although this can also be done by using `R project` this is not an inmediate process as with QGIS. First of all, we need to configure a database connection from `R` to PostgresSQL, by using RPostgreSQL package:

[3] QGIS Development Team (2015). QGIS Geographic Information System. Open Source Geospatial Foundation Project. `http://qgis.osgeo.org`.



(a) Frequencies for a single audition (b) Frequencies colormap for all auditions

Figure 5: Frecuencies maps elaborated with respondents' answers and represented using QGIS.

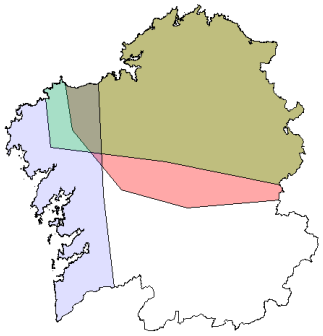


Figure 6: An example of how respondents' answers for audition 1 and 3 (blue and green respectively) overlap a dialectal variety determined by linguists for audition 3 (red).

```
require(RPostgreSQL)
m <- dbDriver("PostgreSQL")
con <- dbConnect(m,user="user_name",password="*****",dbname="database_name")
```

In this way, it is neccesary to use SQL language to extract information from the different tables in our database. An example of how to obtain information is shown below:

```
sql.galicia <- "SELECT ST_AsText(st_transform(geo,25829)) AS ShapeWKT from
+ geo_galicia;"
rs = dbSendQuery(con,sql.galicia);rs
df = fetch(rs,n=-1)
geo_galicia=readWKT(df)
```

By using the `rasterVis` library we can also represent frecuencies maps as shown in Figure 7.

In addition, we can perform statistical analysis in order to study which factors influence the geographical varieties of Galician language recognition. As mentioned before, we can use Baddeley's distance to measure the discrepancies between points and/or polygons. We therefore notice if respondents are able to place and identify dialectal varieties by measuring distances. This led us to define new variables of interest:

– Dependent variable: the distance between respondents' selected region and the true region of dialectal varieties (called "theoretical distance").

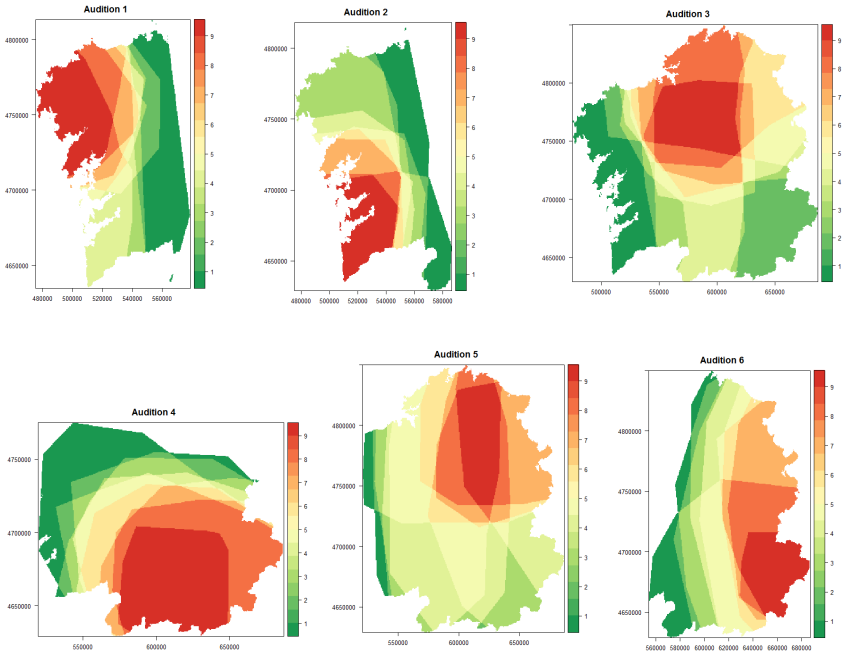


Figure 7: Frequencies maps for different auditions obtained with R project.

– Explanatory variables: distance between birthplace and audition place, distance between birthplace and survey place, distance between survey place and audition place.

These are managed together with other several variables: gender, age range, studies, etc. Under these circumstances, a regression model can be formulated in order to study the determining elements in the recognition of varieties.

The following case provides an example of the results obtained for our simulated PD study. In particular, we focus on third-region variety recognition. Firstly, the true variety region (in this case, considered as one of the polygons issued by respondents) and the selected regions by respondents are represented in Figure 8, on the left and right, respectively.

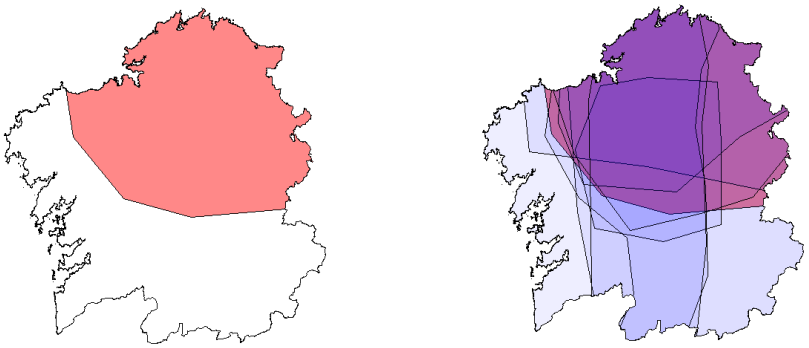


Figure 8: Dialectal region for the third audition / Respondents' selections for the third audition.

Once regions and sociodemographic information are stored in computerized data, we can directly access to obtain:

– Dependent variable: "theoretical distance" (measured in kilometres) - a total of nine distance values. The polygon selected as the true region, the second respondants selection, had a distance equal to zero.

– Explanatory variables: gender, age, educational level and some locations, such as birthplace. To deal with this type of information, for example, we can quantify the distance between audition place and the birthplace, measured in kilometres (see Figure 9).

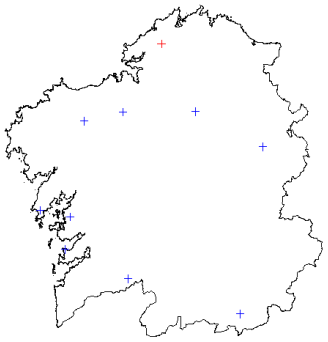


Figure 9: Locations of the audition place (red) and respondents' birthplace (blue).

Thus, we can merge data of interest from the third audition and create our base to work which to work, as shown in Table 1.

Table 1: Database for the third audition in our ficticious PD study.

id	theoretical distance (km)	gender	age	educational level	birthplace (km)
1	35.574	h	45	primary	52.015
2	0.000	m	43	primary	72.774
3	6.303	m	41	secondary	96.135
4	30.342	h	48	secondary	151.062
5	19.042	h	53	primary	137.039
6	9.340	m	39	secondary	50.424
7	27.842	h	51	secondary	130.229
8	2.229	m	36	universitary	187.196
9	30.121	h	55	secondary	157.935

After defining the variables and the data set, a multiple regression analysis can be performed in order to study which factors are influencing the dialectal varieties recognition. This can be written as:

$$Y = X\beta + \epsilon,$$

where X includes both categorical and continuous variables and ϵ is the error term. For our particular case,

$$Theoretical\ distance = \beta_0 + gender\beta_1 + age\beta_2 + educational\ level\beta_3 + birthplace\beta_4 + \epsilon.$$

Model variables have been selected by the Akaike Information Criterion (AIC) using a backward procedure. Final model regression includes gender, educational level and birthplace as explanatory variables, results obtained are shown in Table 2.

We can observe that the average of the theoretical distance, without considering explanatory variables, is equal to 40.71 (km). In addition, our results showed that being a woman and born far from the audition place decrease the dialectal identification in 31.24 and 0.14 points respectively. However, a higher level of education increase the third audition recognition.

The previous results are obtained with simulated PD data. However, over a period of three months, from July to September 2015, the fieldwork to obtain real data will be performed and subsequently analyzed with these tools.

Table 2: Final regression model.

Variable	Coefficients	p-value
Intercept	40.7055	< 0.05
gender:woman	-31.2349	< 0.05
educational level:secondary	8.8054	0.03
educational level:universitary	18.6868	0.05
birthplace	-0.1385	0.03
Adjusted R-squared: 0.9464		

REFERENCES

Baddeley, A. J. (1992) An error metric for binary images. In Robust Computer Vision, W. Förstner and S. Ruwiedel (Eds.), Karlsruhe, Wichmann, pp. 59–78.

Cukor-Avila, P., Jeon, L., Rector P. C., Tiwari, C. and Shelton, Z. (2012) "Texas - Its Like a Whole Nuther Country": Mapping Texans' Perceptions of Dialect Variation in the Lone Star State. Proceedings of the Twentieth Annual Symposium About Language and Society–Austin. Texas Linguistics Forum, 55, pp. 10–19.

Fernández Rei, F. (1990) Dialectoloxía da lingua galega. Vigo: Edicións Xerais de Galicia.

Montgomery, C. (2007) Northern English Dialects: A perceptual approach. PhD thesis, University of Sheffield.

Montgomery, C. (2011) Starburst Charts: Methods for investigating the geographical perception of and attitudes towards speech samples. In Studies in Variation, Contacts and Change in English (eVARIENG) Volume 7.
`http://www.helsinki.fi/varieng/journal/volumes/07/montgomery/index.html`

Montgomery, C. and Stoeckle, P. (2013) Geographic information systems and perceptual dialectology: a method for processing draw-a-map data. Journal of Linguistic geography, 1, 52–85.

Preston, D. R. (1989) Perceptual Dialectology: Nonlinguists' Views of Areal Linguistic. Dordrecht: Foris.

Preston, D. R. (1999) Introduction. In Preston, D. R. (ed.): Handbook of Perceptual Dialectology, vol.1: xxiii-xxxix. Amsterdam: John Benjamins Publishing Company.

Sousa, X. (2006) Aproximación á análise dialectométrica das variedades xeolingüísticas galegas: un estudo comparativo, in M. C. Rolão Bernardo & H. Mateus Montenegro (ed.), Actas do I Encontro de Estudos Dialectológicos, Ponta Delgada: Instituto Cultural de Ponta Delgada, 345-362.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

MODELIZACIÓN ONOMÁSTICA

María José Ginzo Villamayor[1], Rosa M. Crujeiras[1]

[1]Departamento de Estatística e Investigación Operativa, Universidade de Santiago de Compostela

RESUMO

Os apelidos poden ser utilizados como unha fonte de información para caracterizar a poboación dunha rexión, dado que a análise dos patróns que se observan na distribución dos apelidos reflicte aspectos importantes dos movementos poboacionais. As investigacións desenvolvidas no contexto de estudo, ata a data, non teñen en conta a dimensión espacial e espazo–temporal da evolución dos apelidos; por iso, este traballo céntrase na introdución de métodos estatísticos para o procesamento de datos e o modelado en xeolingüística, especificamente, nos apelidos de Galicia.

Palabras e frases chave: Besag-York-Mollié, Inferencia Bayesiana, INLA, Onomástica

1. INTRODUCIÓN

O obxectivo principal deste traballo é o modelado espacial e espazo-temporal de patróns de apelidos en Galicia. Fixando rexións administrativas, como por exemplo, concellos, pode facerse uso de métodos espaciais e espazo–temporais para a análise de datos de conteo que permitan modelar o patrón subxacente á evolución dos apelidos. Estes métodos serán útiles para caracterizar patróns de evolución dos apelidos formulados mediante modelos xerárquicos. Para o axuste de modelos xerárquicos neste contexto faise uso da metodoloxía INLA (Integrated Nested Laplace Approximation) proposta por Rue et al. (2009) que se aplica aos datos dos apelidos en Galicia, proporcionados polo Instituto de Estatística de Galicia (censo de 2011, http://www.ige.eu/).

Neste traballo analízase o patrón espacial, por concellos, dos apelidos "Crujeiras", "Ginzo" e "Rodríguez" en Galicia para o ano 2011. Ademáis, faise unha análise espazo–temporal para o apelido "Ginzo". Para este análise usánse os datos do Padrón continuo de poboación dende o ano 2010 ata 2015.

1.2 DATOS

Os datos dos que se dispón son os do censo de tódolos Concellos de Galicia para o ano 2011. Neste traballo estudaranse tres apelidos de forma espacial: "Crujeiras", "Ginzo" e "Rodríguez", só cando as persoas o levan no primeiro apelido. Na Figura 1 móstranse as localizacións xeográficas dos tres apelidos en estudo.

- **Crujeiras**[1]**:** 488 persoas, localízase principalmente na comarca do Barbanza concretamente no concello de Ribeira. Crujeiras é a castelanización de Cruxeiras, e a maior parte dos seus portadores proceden do lugar Cruxeiras concello de Ribeira. Pola súa parte, é unha variante fonética de Curuxeiras, plural de curuxeira 'lugar onde aniñan e habitan as curuxas' e como 'aldea ou casal situado nun lugar esgrevio e penedoso', de orixe etimolóxica discutida. Está en relación cos apelidos Curuxeira, Coruxeira, Cruxeira, que tamén están castelanizados na súa maior parte. Con diferentes formas e grafías, tanto a forma singular como a forma en plural están documentadas como complemento onomástico desde comezos do século XV.

[1]**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.

- **Ginzo**[1]**:** 130 persoas, localízase principalmente na comarca da Mariña Oriental máis precisamente na Pontenova.

  Ginzo é a castelanización de Xinzo, topónimo de orixe prerromana (med. Genitio) que se repite en 12 lugares de Galicia, amais 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 Casarego (Asturias).

- **Rodríguez**[1]**:** 118209 persoas. No ano 2011 era o apelido máis común, repartido por toda Galicia.

  Rodríguez é o máis frecuente dos apelidos galegos, presente en tódolos concellos de Galicia, con case 240.000 ocorrencias se temos en conta primeira ou segunda posición no apelido. É de tipo patronímico con sufixo "-ez", que era o procedemento unsado na Idade Media para significar 'fillo de', porque daquela os apelidos non eran hereditarios, e cambiaban de xeración en xeración. A alta frecuencia deste apelido vén dada pola do nome de que procede, Rodrigo, un dos nomes de orixe xermánica máis espallados na Idade Media, pola existencia dun santo (mártir en Córdoba no século X) e dalgúns personaxes históricos relevantes, como o último dos reis godos ou o heroe castelán Roy Díaz de Vivar, o Cide. Documéntase nos textos medievais desde o século XI.

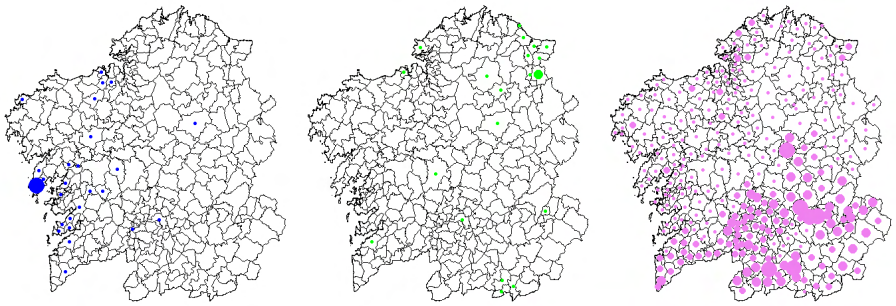


Figura 1: De esquerda a dereita: localizacións en Galicia dos apelidos Crujeiras, Ginzo e Rodríguez. O tamaño do punto é proporcional o número de persoas que levan ese apelido ponderado pola poboación do concello. Para que se viran claramente os puntos, o que se fixo foi dividir en 4 intervalos cada apelido en función do cociente número de persoas e poboación e asignóuselle un tamaño de punto.

No período 2000–2015 o número medio de persoas que levan o apelido Ginzo de forma indistinta no primeiro ou no segundo lugar é de 261, cun comportamento bastante uniforme no período. Salientar que nos anos 2000 e 2015 son nos que menos persoas levan o apelido (252) e pola contra no ano 2008 é o ano que máis se rexistran (267). Un resumo por provincias ao longo do tempo pode consultarse na Figura 2.

2. METODOLOXÍA

Nas últimas décadas, a dispoñibilidade de datos espaciais e espazo–temporais aumentou de forma considerable, principalmente debido ao avance das ferramentas computacionais que permiten recoller os datos en tempo real.

A análise da distribución xeográfica e temporal dos apelidos permite estudar a variabilidade espacial e temporal da estrutura das poboacións humanas. Deste xeito, os apelidos poden empregarse como como fonte de información das características da poboación. Ademais, a análise de patróns de apelidos proporciona información dinámica dos movementos de poboación. En Ginzo et al. (2013) móstrase como a través do estudo dos apelidos mediante medidas de isonimia combinadas con ferramentas de análise clúster se poden obter mapas de apelidos que reflicten o proceso

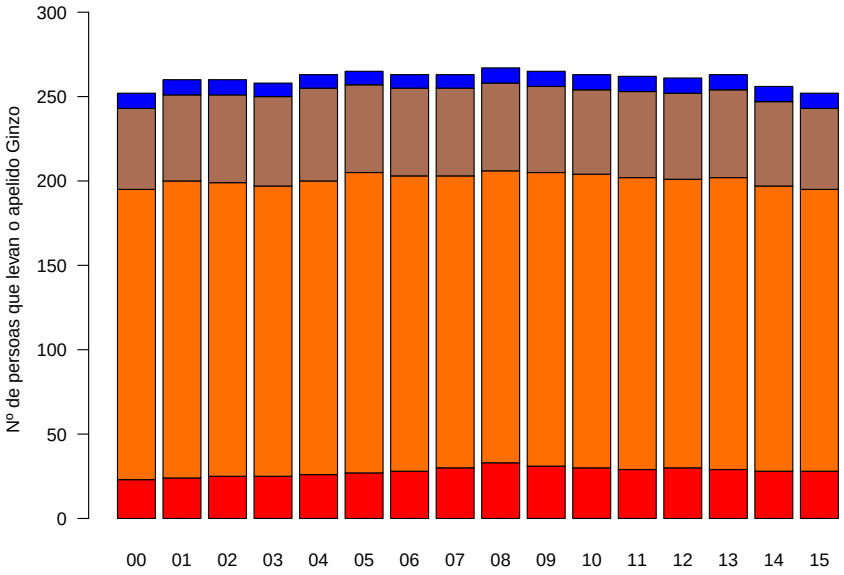


Figura 2: Distribución provincial do apelido Ginzo no período 2000–2015, en cor vermella representa as persoas da provincia da Coruña, en laranxa as da provincia de Lugo, en marrón as de Ourense e en azul as de Pontevedra.

de urbanización das zonas rurais as urbanas, entre outros. Preséntanse a continuación os dous modelos estudados, espacial e espazo–temporal, empregados neste traballo.

2.1 Modelización espacial

Para analizar o patrón espacial dos apelidos en Galicia, considerando o efectos da covariable número de habitantes por concellos, axustouse o modelo proposto por Besag et al. (1991) (modelo BYM), adaptado a este contexto. Un dos supostos deste modelo é que o log-risco[2] pódese descompoñer como suma dunha compoñente espacial estruturada e un erro aleatorio, pero tamén se pode incluír o efecto suave dalgunha covariable. Formalmente, o modelo BYM formúlase do seguinte xeito:

Denótase por Y_i = o número de persoas con ese apelido no concello i, para cada $i = 1, \ldots, n$

Este proceso de conteo será modelado a través dun modelo Poisson–LogNormal (ver Banerjee et al. (2004) páx. 162). É dicir,

$$Y_i|\eta_i \sim \text{Pois}(E_i \exp(\eta_i)),$$

onde E_i é a poboación en risco (estimación da poboación que leva o apelido en estudo), η_i (os riscos log–relativos) un predictor lineal e as variables Y_i condicionalmente en η_i son independentes. O campo latente η_i terase en conta para modelar a estrutura subxacente e recoller a variabilidade espacial. Unha formulación sinxela vén dada por:

$$\eta_i = \mu + f_s(s_i) + f_u(s_i),$$

onde s_i é o centroide de cada concello e f_s e f_u denotan os efectos espaciais estruturado e o non estruturado, respectivamente. Para f_s, imporase un campo aleatorio Gaussiano de Markov (GMRF) intrínseco (Rue e Held (2005)). Para f_u, considerase o proceso de ruido branco que representa a "sobredispersión" que poden presentar os concellos:

[2]Estes modelos empréganse no ámbito da epidemioloxía. Aquí log–risco enténdese como a poboación susceptible de posuír o apelido en estudo, en palabras epidemiolóxicas, o risco de "padecer" un determinado apelido.

- Denotando por $z(s_i) \equiv f_s(s_i), i = 1, \cdots, n$, Z é un GMRF, (véxase a Sección 2.2 no libro de Rue e Held (2005) para unha definición precisa de GMRF). Os $z(s_i)$, $z(s_j)$ con $i \neq j$ son dependentes con estrutura de Markov e seguen unha distribución $N(0, \tau_2)$.
- Por outra banda, denótase por $w(s_i) \equiv f_u(s_i), i = 1, \cdots, n$, W é ruido branco. Os $w(s_i), i = 1, \cdots, n$ son independentes con distribución $N(0, \tau_1)$.

2.2 Modelización espazo–temporal

Investigando só o patrón espacial dos apelidos non nos permite concluír nada sobre outra das compoñentes de variación, a temporal, que poder ser igualmente de interese. O modelo anterior pode estenderse facilmente ao caso espazo–temporal incluíndo o tempo. A variable resposta para un concello i será:

$$Y_{it} \; = \text{ o número de persoas con ese apelido no concello } i \text{ no tempo } t$$

que será observada nos n concellos e en T instantes do tempo. O modelo espacial anterior esténdese ao permitir a compoñente temporal quedando:

$$Y_{it}|\eta_i \sim \text{Pois}(E_{it} \exp(\eta_{it})),$$

A formulación que seguirá neste caso o campo latente é:

$$\eta_{it} = \mu + f_s(s_i) + f_u(s_i) + f_T(t).$$

con $t = 1, \ldots, T$, onde en $f_T(t)$ se especifica a estrutura temporal. Dita estrutura pode corresponder a unha compoñente lineal no tempo, a unha compoñente suave ou ben a un proceso que inclúa correlación temporal.

2.3 Inferencia Bayesiana con INLA

O axuste dos modelos realizouse empregando o método baseado na aproximación por integradores de Laplace aniñados INLA[3]), proposto por Rue et al. (2009).

INLA propociona unha ferramenta rápida e útil para axustar modelos gaussianos latentes (os procesos que rixen f_s, f_u e f_T teñen distribucións Gaussianas), incluíndo modelos con estrutura temporal ou espacial nun contexto Bayesiano. Neste traballo farase uso deste metodoloxía para axustar un modelo aos datos de tres apelidos en Galicia, Crujeiras, Ginzo e Rodríguez e detallarase cómo se fixo o axuste con `R-INLA`[4] (dispoñible en www.r-inla.org).

Os obxectivos da inferencia Bayesiana son as distribucións marxinais a posteriori para cada compoñente do vector de parámetros. O procedemento INLA calcula a aproximación numérica das distribucións a posteriori que sexan de interese, baseados no método de aproximación de Laplace.

Para levar a cabo a inferencia Bayesiana hai que especificar os hiperparámetros do efecto espacial a priori e da parte aleatoria. Os hiperparámetros son a precisión do τ_1 no modelo `iid` e a precisión τ_2 do modelo `besag`, como $\theta = (\log \tau_1, \log \tau_2)$. Sobre estes parámetros considéranse unhas distribucións a priori, log–gamma neste caso, con parámetros iniciais (1,0.00005).

```
> prior.spat=c(1,0.00005)
> hyper.spat=list(prec=list(param=prior.spat))
> prior.iid=c(1,0.00005)
> hyper.iid=list(prec=list(param=prior.iid))
```

A análise de sensibilidade do modelo á eleción do parámetro das distribucións a priori pode facerse en base o criterio Criterio de Información de Desviación (DIC). Este criterio baséase na desviación Bayesiana (véxase Speigelhalter et al. (1998)) e que é unha xeneralización dos Criterio de Información de Akaike (AIC) e do Criterio de Información Bayesiano (BIC). O criterio DIC

[3] Bayesian computing with INLA!, (2009) URL: http://www.r-inla.org/.

[4] `R-INLA` é un paquete de `R` que implementa a aproximación da inferencia Bayesiana usando INLA (Martino e Rue, 2010a).

emprégase, especialmente, sobre distribucións a posteriori obtidas mediante métodos Markov chain Monte Carlo (MCMC).

Ademais é un criterio que proporciona bos resultados para modelos xerárquicos como os aplicados neste traballo. Os modelos que teñen un valor DIC máis pequeno son os preferidos. O DIC penaliza tanto polo axuste como pola complexidade do modelo. O axuste soe ser mellor ao introducir un número maior de parámetros ao modelo, pero isto compénsase no criterio DIC cunha penalización segundo o número de parámetros buscando un equilibrio entre a bondade do axuste e a complexidade do modelo. Para a súa construción débese definir a desviación:

$$D(\lambda) = -2\ln\left(p\left(Y|\lambda\right)\right)$$

onde Y representa a variable resposta, λ os parámetros específicos de cada modelo. (Os parámetros que aparecen no BYM son μ, τ_1 e τ_2).

Así mesmo pode obterse a desviación media a posteriori como: $\bar{D} = E\left[D(\lambda)\right]$. Por outra banda a desviación Bayesiana das medias a posteriori é $\hat{D} = D\left[E(\lambda)\right]$. Deste xeito temos que o criterio DIC ven definido por:

$$DIC = 2\bar{D} - \hat{D}$$

3. RESULTADOS

3.1 Resultados da modelización espacial con INLA

Para o ano 2011 en Galicia hai 488 persoas que levan o apelido Crujeiras, 130 o apelido Ginzo e 118209 o apelido Rodríguez. Fíxose un axuste dacordo ao exposto na Sección 2.1 tomando os hiperparámetros a priori que emprega por defecto en `inla`, xa que o emprego de outros non melloran os resultados. A especificación da fórmula para executar o modelo en R–INLA foi a seguinte:

```
> formula1 = NUMERO ~ f(apexx315.struct,model="besag",graph="nc.adj",
adjust.for.con.comp = TRUE) +f(ID2,model="iid")
```

- Na función `f` especifica o tipo da función tanto para a parte estruturada como a non estruturada. Para a f_u a opción por defecto é `iid`, significa neste caso que a variable aleatoria `ID2` é independente e con distribución Gaussiana.
- A variable `NUMERO` indica o número de persoas que levan o apelido en estudo, (Crujeiras, Ginzo ou Rodríguez), no concello i.
- A variables `apexx315.struct` e `ID2` indican o concello i, $i \in 1, \ldots, 315$.

O grafo de Galicia non é un conxunto conexo, senón que está formado por dúas compoñentes conexas: a Illa de Arousa e os 314 restantes concellos integrados na Península Ibérica. Isto hai que indicarllo ao modelo co parámetro `adjust.for.con.comp = TRUE`. Para executar o algoritmo INLA empregase a función `inla`:

```
> result1  =  inla(formula1,family="poisson",data=df.APE,E=df.APE$pobpad01,
control.compute=list(dic=TRUE))
```

A función `inla` devolve un obxeto da clase `inla`, que contén unha serie de elementos que poden ser explorados, como por exemplo na Táboa 1 móstrase o DIC obtidos para os tres modelos así como información da distribución a posteriori dos coeficientes dos modelos estimados (media, desviación típica e cuantiles).

Na Figura 3 móstrase o patrón espacial dos apelidos Crujeiras, Ginzo e Rodríguez tendo en conta o número de habitantes por concello, representa a media a posteriori do efecto estruturado e na Figura 4 a parte aleatoria non estruturada.

3.2 Resultados da modelización espazo–temporal con INLA

	DIC		media	desviación típica	0.025cuantil	0.5cuantil	0.975cuantil
Crujeiras	215.39	Intercepto	-13.32	0.69	-14.81	-13.24	-12.19
Ginzo	181.18	Intercepto	-13.63	0.78	-15.29	-13.51	-12.40
Rodríguez	2816.88	Intercepto	-3.16	0.01	-3.17	-3.16	-3.15

Táboa 1: DIC e resultados da distribución a posteriori para a media do proceso espacial estruturado.

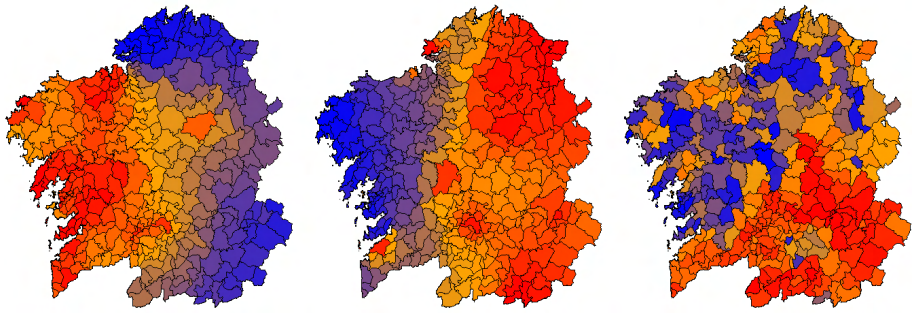


Figura 3: Media a posteriori da parte estruturada do campo latente η_i. De esquerda a dereita: para os apelidos Crujeiras, Ginzo e Rodríguez dos datos do censo 2011.

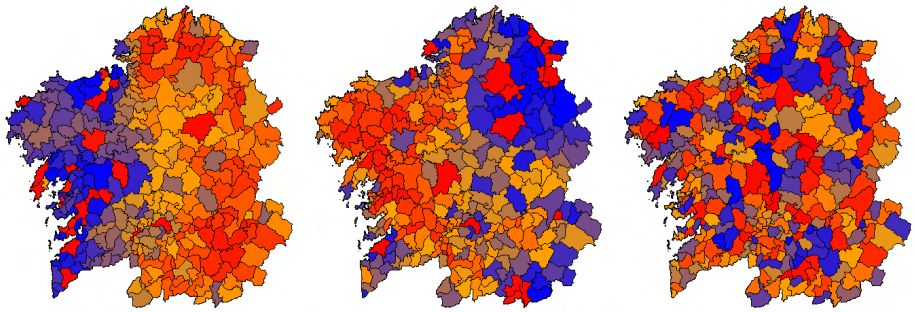


Figura 4: Media a posteriori do efecto non estruturado do campo latente η_i. De esquerda a dereita: para os pelidos Crujeiras, Ginzo e Rodríguez dos datos do censo 2011.

Neste apartado fíxose un axuste sinxelo de acordo ao exposto na Sección 3.2. Eliximos como parámetro para o efecto temporal o `prior.iid=c(1,1)` e axustamos o modelo ao segundo conxunto de datos.

A especificación da fórmula para executar o modelo en R–INLA ten o seguinte formato:

```
> formula.par <- NUMERO ~ 1 + f(ID2,model="bym",graph="nc.adj",
adjust.for.con.comp = TRUE, constr=TRUE) + f(ID21,anoNum,model="iid",
hyper=hyper.iid, constr=TRUE) + anoNum
```

- A variable `NUMERO` indica o número de persoas que levan o apelido Ginzo no concello i no ano t.
- As variables `ID2` e `ID21` indican o concello i, $i \in 1, \ldots, 315$.
- `anoNum` representa o ano t, $i \in 1, \ldots, 16$.

Na primeira parte do modelo especifícase o modelo BYM como o modelo espacial descrito na Sección 2.1, mentres que `anoNum` identifica o efecto temporal global (estimado como un efecto fixo) e `f(ID21,anoNum,model=''iid'', hyper=hyper.iid, constr=TRUE)` é a tendencia diferencial, é dicir, a interacción entre espazo e o tempo, modelado a través da distribución `iid`. Para executar o algoritmo INLA mediante a función `inla` emprégase a seguiente expresión:

```
> result1  <- inla(formula.par,family="poisson",data=misDatos,E=poblacion*0.2,
                  control.predictor=list(compute=TRUE),
                  control.compute=list(dic=TRUE,cpo=TRUE), verbose=TRUE)
```

Neste caso emprégase como "poboación en risco" o 20 % da poboación total do concello, facendo unha estimación naqueles concellos que o apelido Ginzo está máis presente. Na Táboa 2 tense un resumo para os coeficientes dos efectos fixos do modelo estimado (media, desviación típica e cuantiles). O DIC neste caso é 2264.18. Tanto o `intercepto` como o coeficiente asociado á variable ano, que recolle a compoñente temporal, son significativos.

	media	desviación típica	0.025cuantil	0.5cuantil	0.975cuantil
Intercepto	-15.10	1.29	-17.68	-14.87	-12.96
Ano	-0.18	0.06	-0.31	-0.18	-0.08

Táboa 2: Estimación dos coeficientes do modelo espazo–temporal axustado para o apelido Ginzo.

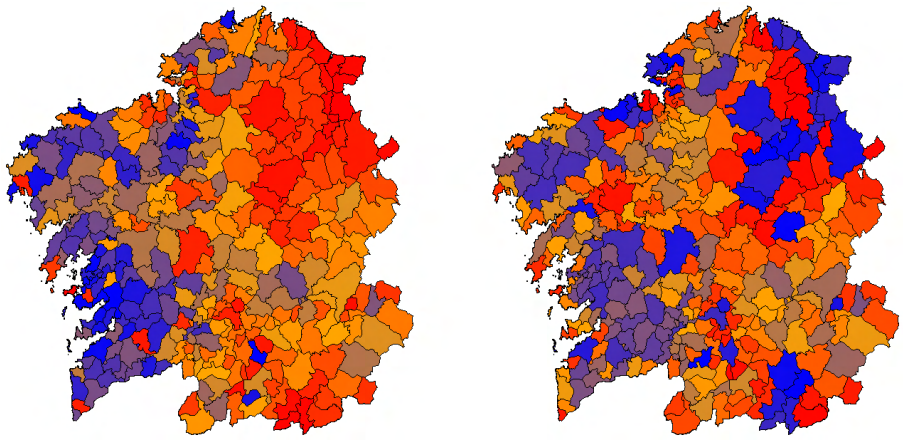


Figura 5: Esquerda: Media a posteriori da parte non estruturada do campo latente η_i para o axuste do modelo espazo–temporal.

4. CONCLUSIÓNS

A conclusión que se extrae deste estudo unha vez aplicado o modelo espazo–temporal fronte ao modelo espacial é que os mapas das distribucións dos apelidos son similares. Quizás poida dicirse que nas veciñanzas desas 261 persoas que levan o apelido Ginzo non hai moito movemento nin no espazo nin ao longo do tempo. Seguramente se tivéramos máis anos poderíamos obter resultados máis interesantes, xa que tendo en conta a idade das persoas e facendo cortes temporais como en Ginzo et al. (2013) mediante técnicas de análise clúster pódese detectar o proceso de urbanización en Galicia ou mesmo o mapa das dioceses de Galicia (antes do proceso urbanización o mapa dos apelidos de Galicia distribuíase da mesma forma que o mapa das dióceses).

AGRADECEMENTOS

María José Ginzo e Rosa M. Crujeiras agradecen o apoio dos proxectos 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[5], aos investigadores Xulio Sousa Fernández[6] e Ana Isabel Boullón Agrelo[6], e o Instituto Galego de Estatística por ceder os datos para esta investigación.

[5]Páxina web de TecAnDali http://ilg.usc.es/tecandali/.
[6]Instituto da Lingua Galega (USC).

REFERENCIAS

Banerjee S., Carlin, B. E., Gelfand, A. (2004) Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC Monographs on Statistics & Applied Probability

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.

Blangiardo, M. e Cameletti, M. (2015) Spatial and Spatio-temporal Bayesian Models with R – INLA. John Wiley & Sons, Ltd

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).

Rue, H., e Held, L. (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall, CRC Press, London.

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.

Spiegelhalter, D., Best, N. G., Carlin, B. (1998) Bayesian deviance, the efective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Biostatistics, University of Minnesota.

Comunicacións en forma de póster

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Os titulados universitarios no Sistema Universitario de Galicia (SUG) traballan en Galicia nos dous primeiros anos posteriores á súa titulación?

Mª Carmen Fernández Montes[1], José Eduardo López Pereira[1], Mª Patrocinio Morrondo Pelayo[1,] Pedro Faraldo Roca[2], Rosa María Crujeiras Casais[2], Adela Martínez Calvo[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]

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

RESUMO

Desde a creación da Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG) no ano 2001, vense analizando a situación laboral dos titulados universitarios no Sistema Universitario de Galicia (SUG), de forma periódica e continua a través dos estudos de inserción laboral desenvolvidos pola ACSUG. O obxectivo principal destes estudos é obter información da percepción dos titulados sobre a formación e a universidade, as características do proceso de procura de emprego, a situación do mercado de traballo dos titulados no SUG e as súas condicións laborais.

Este traballo recolle unha análise da integración no mercado laboral en Galicia dos titulados universitarios no SUG, centrándose principalmente no lugar onde atopan traballo os titulados correspondentes aos seis últimos estudos de inserción laboral realizados pola ACSUG.

Palabras e frases chave: inserción laboral, titulados universitarios, lugar de traballo, tempo en atopar o primeiro emprego.

1. INTRODUCIÓN

A consolidación da periodicidade anual destes estudos ao longo dos anos garante a obtención de históricos de resultados permitindo analizar a evolución no tempo da situación laboral dos titulados universitarios. Entre outros factores, analízase a porcentaxe de titulados que traballan en algo relacionado coa titulación estudada, preparan oposicións e/ou estudan, lugar de traballo, tipoloxía dos contratos, nivel retributivo e tempo en atopar emprego relacionado coa súa titulación. Neste traballo, centrámonos na análise do lugar de traballo dos titulados analizando un total de 29.844 rexistros.

Na Sección 2 preséntase un resumo dos datos técnicos dos estudos realizados sobre as cohortes de titulados no 2005-2006 ata o 2010-2011 (véxase Referencias). O resumo dos principais resultados ocupa a Sección 3. Na Sección 4 achéganse unhas conclusións xerais.

2. DATOS TÉCNICOS DOS ESTUDOS

A poboación obxecto de estudo está constituída polos titulados no SUG durante os cursos académicos comprendidos entre os cursos académicos 2005-2006 ata 2010-2011 (EIL0506 ata EIL1011, onde as siglas EIL designan "Estudo de Inserción Laboral"), analizando un total de 29.844 rexistros. 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	**EIL0506**	**EIL0607**	**EIL0708**	**EIL0809**	**EIL0910**	**EIL1011**
Poboación	10.280	8.689	9.216	8.846	8.922	9.289
Mostra	5.362	4.569	4.830	4.965	5.004	5.114
%	52,16%	52,58%	52,41%	56,13%	56,09%	55,05%
Traballo de campo	maio 2008	maio/xuño 2009	maio/nov. 2010	outubro 2011	decembro 2012	decembro 2013

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. RESULTADOS

Para contextualizar o lugar de traballo dos titulados do SUG, analizamos a porcentaxe de titulados que están traballando no momento de realizar a enquisa: 59,74% (EIL1011) traballa en algo relacionado coa súa titulación, na Táboa 2 obsérvase un aumento no número de titulados que traballan. Por outro lado, un 24,11% (EIL1011) continúan estudos e un 14,11% (EIL1011) preparan oposicións.

	EIL0506	**EIL0607**	**EIL0708**	**EIL0809**	**EIL0910**	**EIL1011**
Si	63,60%	62,75%	58,12%	55,80%	55,37%	59,74%
Si, pero non relacionado	7,47%	3,37%	3,81%	3,50%	2,95%	3,59%
Non	28,93%	33,89%	38,07%	40,71%	41,68%	36,67%

Táboa 2: Porcentaxe de titulados que traballan (no momento de realizar a enquisa). No caso de que traballen, indícase a porcentaxe que ocupa un posto non relacionado coa titulación estudada. Resultados para o total do SUG.

A maioría dos titulados desenvolven o seu traballo en Galicia como se pode observar na Figura 1, obtendo para case todos os estudos unha porcentaxe próxima ao 80,0%. Con respecto aos que traballan noutra comunidade autónoma de España representan menos do 15% e menos dun 1% desempeñan a súa labor no resto do mundo. As provincias de A Coruña e Pontevedra, ver Táboa 3, concentran a maioría de titulados con valores arredor do 42% e o 25% respectivamente, fronte a Lugo e Ourense que roldan valores entre o 9%.

Centrándonos nos resultados acadados no último estudo EIL1011, o 82,52% dos titulados do SUG que traballa, faino en Galicia, co seguinte reparto por provincias: o 41,47% na provincia da Coruña, o 24,28% na de Pontevedra, un 8,77% en Lugo e o restante 8,00% en Ourense. Un 13,55% dos titulados traballa noutra provincia española e tan só un 3,93% saíu do país, case todos eles (3,15%) con destino a outro país europeo. Na Figura 1 achégase unha comparativa destes datos ao longo dos distintos estudos.

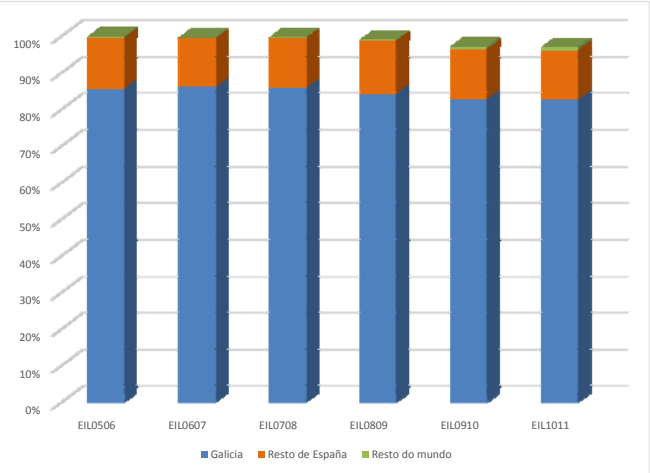


Figura 1: Distribución dos titulados segundo o lugar de traballo: Galicia, resto de España, resto do mundo. Comparativa dos seis últimos estudos: EIL0506 ata EIL1011.

	EIL0506	EIL0607	EIL0708	EIL0809	EIL0910	EIL1011
A Coruña	41,18%	44,07%	43,65%	40,16%	40,59%	41,47%
Lugo	8,72%	8,99%	8,70%	11,05%	9,69%	8,77%
Ourense	8,88%	8,49%	9,11%	8,16%	8,68%	8,00%
Pontevedra	26,76%	24,46%	24,40%	24,69%	23,60%	24,28%

Táboa 3: Distribución dos titulados segundo o lugar de traballo, considerando as provincias de Galicia. Resultados para o total do SUG.

Amósase na Figura 2 o lugar de traballo para os titulados no SUG, por ramas de coñecemento. Os titulados que en máis alta porcentaxe traballan en Galicia son os de Ciencias Sociais e Xurídicas III (91,30%), cun reparto por provincias similar ao do SUG. Os titulados que máis saen de Galicia para traballar son os de Ciencias (18,39% resto de España, 5,27% resto de Europa), os de Enxeñaría e Arquitectura (18,32% resto de España, 6,12% resto de Europa) e os de Ciencias Sociais e Xurídicas IV (18,71% resto de España e 9,81% resto de Europa).

Un caso diferente é o dos titulados en Artes e Humanidades II, onde a porcentaxe de titulados que traballan en Galicia é pouco inferior á do SUG (77,45%) e traballan noutra provincia española moitos menos ca no SUG (7,30%). Pola contra, un 11,35% traballa no resto de Europa e un 3,90% no resto do mundo (0,78% no total do SUG).

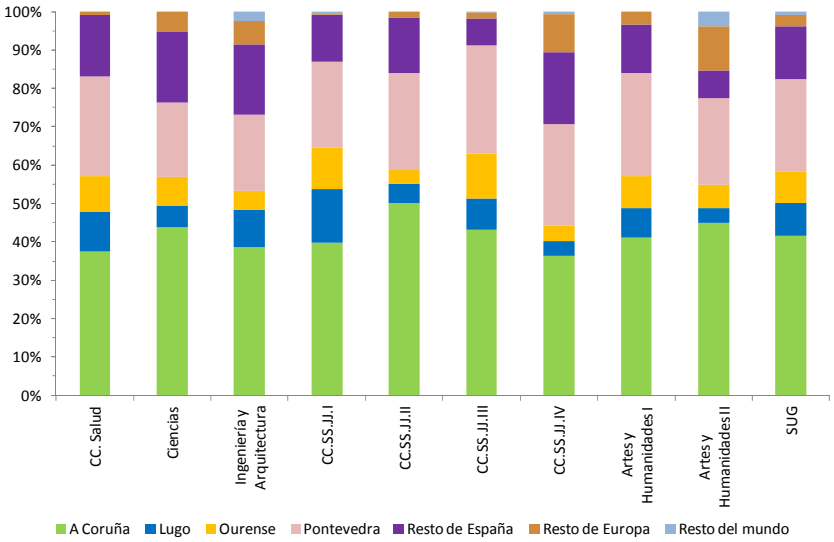


Figura 2: Distribución dos titulados segundo o lugar de traballo. Resultados por rama de coñecemento para os titulados no curso académico 2010-2011 (EIL1011).

Con respecto ao grao de satisfacción dos titulados sobre a situación xeográfica do seu posto de traballo pódese observar na Figura 3 que un 78,25% dos titulados está moi ou bastante satisfeito co lugar xeográfico do seu posto de traballo, mentres que un 13,46% está pouco ou nada satisfeito. Este grao de satisfacción vénse mantendo bastante estable nos estudos dende o EIL0607.

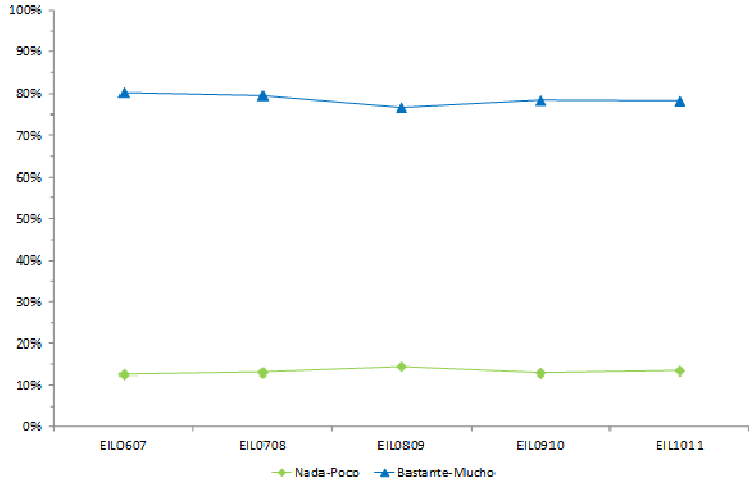


Figura 3: Satisfacción coa zona xeográfica de traballo. Resultados para o total do SUG. Comparativa con estudos anteriores.

4. CONCLUSIÓNS

Os titulados do SUG desempeñan o seu traballo maioritariamente en Galicia. En particular un 82,52% dos titulados enquisados no EIL1011 exercían o seu traballo dentro da comunidade autónoma, co seguinte reparto por provincias: A Coruña (41,47%), Pontevedra (24,28%), Lugo (8,77%) e Ourense (8,00%). Respecto ás ramas de coñecemento, no último estudo 2010-2011, os titulados que en máis alta porcentaxe traballan en Galicia son os de Ciencias Sociais e Xurídicas III (91,30%), cun reparto por provincias similar ao do SUG.

REFERENCIAS

Faraldo, P., Crujeiras, R.M., Lombardía, M.J., Matías, J.M., Naya, S., Fernández, B.M., Carreira V., Abalde, J.E., Fernández, M.C. (2009). Estudo da inserción laboral dos titulados no Sistema Universitario de Galicia 2005-2006. Axencia para a Calidade do Sistema Universitario de Galicia.

Faraldo, P., Crujeiras, R.M., Lombardía, M.J., Naya, S., Matías, J.M., Carreira V., Fernández, B.M., Pereira, J.E., Fernández, M.C. (2010). Estudo da inserción laboral dos titulados no Sistema Universitario de Galicia 2006-2007. Axencia para a Calidade do Sistema Universitario de Galicia.

Faraldo, P., Crujeiras, R.M., Lombardía, M.J., Naya, S., Matías, J.M., Carreira V., Fernández, B.M., Martínez, A., Morrondo, M.P., Pereira, J.E., Fernández, M.C., Sierra M.D., Mallo, M., Baña, J.M. (2011). Estudo da inserción laboral dos titulados no Sistema Universitario de Galicia 2007-2008. 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., Velón, L. (2013). Estudo da inserción laboral dos titulados no Sistema Universitario de Galicia 2008-2009. 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. (2014). Estudo da inserción laboral dos

titulados no Sistema Universitario de Galicia 2009-2010. 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 2010-2011. Axencia para a Calidade do Sistema Universitario de Galicia.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

A inserción laboral dos titulados en Másteres no Sistema Universitario de Galicia 2007-2008, 2008-2009 e 2009-2010

Mª Carmen Fernández Montes[1], José Eduardo López Pereira[1], Mª Patrocinio Morrondo Pelayo[1] , Jose Carlos de Miguel Domínguez[2], David Rodríguez González[2], Beatriz Valcárcel Aguiar[2], Alejandro Vecino Aguirre[2], Luís Pedro Pereira Andrade[3], Manuel Guisado González[4]

1 Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG)
2 Departamento de Economía Cuantitativa. Universidade de Santiago de Compostela
3 Departamento de Economía Aplicada II. Universidade da Coruña
4 Departamento de Organización de Empresas e Marketing. Universidade de Vigo

RESUMO

No ano 2013, a Axencia para a Calidade do Sistema Universitario de Galicia (ACSUG) inicia un novo proxecto onde amplía o campo de estudo da inserción laboral dos titulados universitarios no Sistema Universitario de Galicia (SUG), aos titulados en másteres. Entre os obxectivos principais deste proxecto están obter información sobre aspectos organizativos e desenvolvemento dos másteres, así como coñecer a motivación principal para realizar estudos de máster e análise sobre a experiencia laboral que tiveron estes últimos anos, antes e despois da realización do máster.

Este traballo analiza a motivación principal pola que os titulados universitarios cursaron un máster, a situación laboral posterior a realización do máster, así como a utilidade do mesmo en relación a traxectoria laboral experimentada e a satisfacción dos titulados coa súa realización.

Palabras e frases chave: másteres, inserción laboral, titulados, motivación, situación laboral, utilidade dos másteres

1. INTRODUCIÓN

Coa estrutura dos estudos universitarios actuais: grao, máster e doutoramento, froito da harmonización dos sistemas universitarios no proceso de adaptación ao Espazo Europeo de Educación Superior (EEES), destácase a relevancia acadada polos estudos de máster tanto no ámbito universitario como no ámbito laboral. Por un lado, as tendencias existentes no mercado laboral están directamente relacionadas coas universitarias, sendo o máster un factor decisivo moitas veces para a consecución dun emprego, e por outro lado, nalgúns casos, formando parte da etapa formativa previa aos programas de doutoramento.

2. DATOS TÉCNICOS DOS ESTUDOS

A poboación obxecto de estudo está constituída polos titulados en másteres no SUG correspondentes aos cursos académicos 2007-2008, 2008-2009 e 2009-2010[i]. A enquisa realizouse entre maio e xuño de 2013, polo que para as distintas cohortes de titulados en másteres transcorreron 5, 4 e 3 anos entre a finalización do máster ata o momento de realización da enquisa.

	EILMásteres0708		**EILMásteres0809**		**EILMásteres0910**		**TOTAL**	
	Poboación	**Mostra**	**Poboación**	**Mostra**	**Poboación**	**Mostra**	**Poboación**	**Mostra**
Artes e Humanidades	26	23	47	37	96	69	169	129
Ciencias	80	58	116	92	223	174	419	324
Ciencias da Saúde	169	132	200	156	184	150	553	438
Ciencias Sociais e Xurídicas	86	62	219	168	731	562	1036	792
Enxeñaría e Arquitectura	95	73	244	180	399	306	740	559
SUG	**456**	**348**	**826**	**633**	**1633**	**1261**	**2915**	**2242**

Táboa 1: Tamaño da poboación e mostra dos titulados en másteres no SUG nos cursos académicos 2007-2008, 2008-2009 e 2009-2010 por rama de coñecemento.

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 sitúase no +/-0,99%, 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 traballo de campo foi realizado pola empresa Instituto Sondaxe, S.L. desde o 2 de maio ata o 12 de xuño de 2013.

2. RESULTADOS

Motivación para realizar un máster: O principal motivo que leva os titulados en másteres a realizar estes estudos é complementar a formación académica para obter unha maior especialización no mercado de traballo, ampliar as saídas laborais e alcanzar un maior desenvolvemento profesional, cunhas porcentaxes de 72,4% (EILMásteres0708), 75,5% (EILMásteres0809) e 58,8% (EILMásteres0910).

	EILMásteres0708	EILMásteres0809	EILMásteres0910
Complementar a formación académica (investigación)	30,7%	25,8%	29,0%
Complementar a formación académica (traballo)	72,4%	75,5%	58,8%
Detectar necesidades formativas na experiencia laboral anterior	5,2%	5,4%	4,6%
Era necesario ou obrigatorio para o futuro emprego	4,0%	3,8%	18,7%
Outros	4,0%	3,0%	1,5%

Táboa 2: Motivos para realizar o máster. Resultados para o total do SUG.

Situación laboral no momento de realizar a enquisa: Un 83,6% (EILMásteres0708), 78,8% (EILMásteres0809) e un 73,7% (EILMásteres0910) dos titulados en másteres no SUG traballaba no momento da realización da enquisa (maio e xuño de 2013). Analizando a situación dos titulados que non traballan e tampouco buscan emprego, destacan as porcentaxes de titulados que realizan outros estudos 2,8% (EILMásteres0708), 2,8% (EILMásteres0809) e 4,2% (EILMásteres0910), principalmente de doutoramento ou de grao.

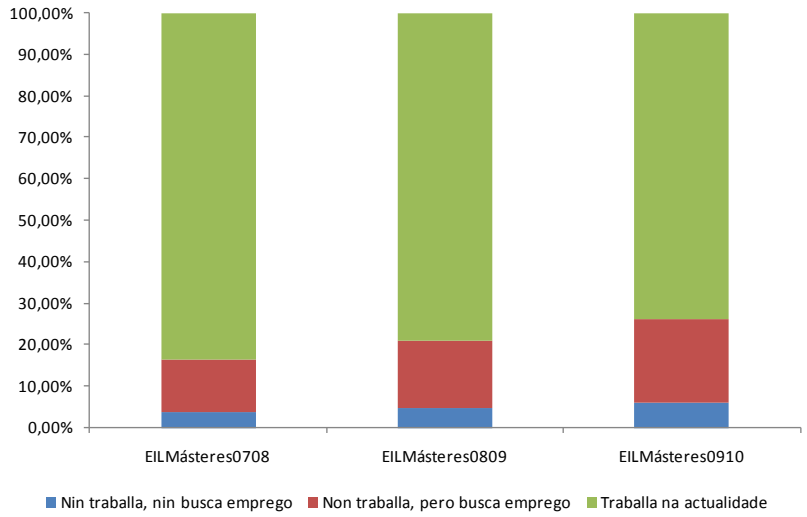


Figura 1: Situación laboral actual no momento de realizar a enquisa. Resultados para o total do SUG.

Valoración da utilidade do máster: Case a metade dos titulados en másteres no SUG valoran como bastante ou moi útil (nunha escala desde 1 ata 5) a realización do máster en función da súa traxectoria

laboral posterior (tanto para os que traballaron con posterioridade como aos que non), 47,4% (EILMásteres0708), 42,6% (EILMásteres0809) e 42,1% (EILMásteres0910).

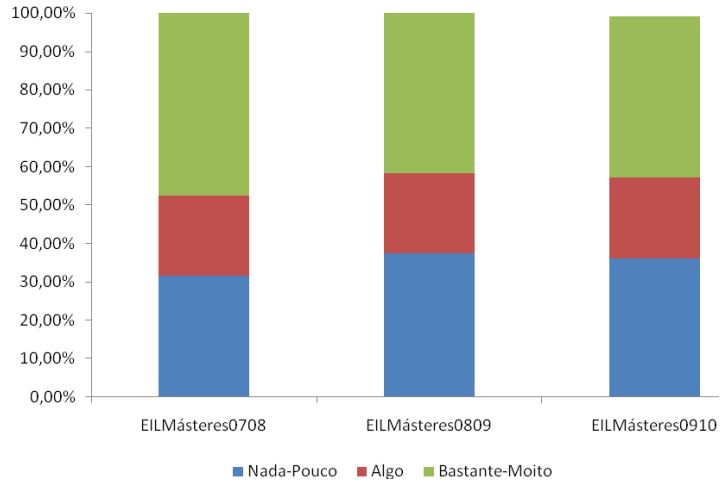


Figura 2: Valoración da utilidade do máster en función da traxectoria laboral posterior. Resultados para o total do SUG.

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.

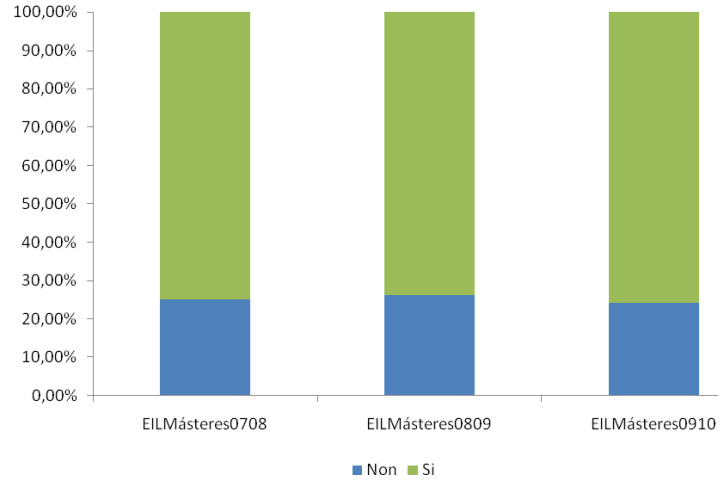


Figura 3: Volverían realizar o máster. Resultados para o total do SUG.

3. CONCLUSIÓNS

Os titulados universitarios en másteres queren complementar a súa formación académica para obter unha maior especialización no mercado de traballo, traballan na súa maioría despois de cursar o máster, consideran que foron de utilidade en función da súa traxectoria laboral posterior e volverían a cursalo.

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.

1 Os resultados obtidos para as distintas cohortes de titulados designaranse como EILMásteres0708 (resultados acadados para os titulados en másteres no curso académico 2007-2008), EILMásteres0809 (resultados acadados para os titulados en másteres no curso académico 2008-2009) e EILMásteres0910 (resultados acadados para os titulados en másteres no curso académico 2009-2010).

XII *Congreso Galego de Estatística e Investigación de Operacións*
Pazo de San Marcos-Deputación de Lugo, 24-25-26 de outubro de 2015

ANÁLISIS DE LA VISIÓN DE LA PRENSA VIGUESA SOBRE LA EXCLUSIÓN SOCIAL EN LA PRIMERA FASE DE LA CRISIS (2009-2010)

Concepción GómezLópez[1], Pilar García Soidán[1]

[1]Facultad de Ciencias Sociales y de la Comunicación, Universidad de Vigo

RESUMEN

La historia parece confirmar el carácter cíclico de la economía, con etapas de bonanza y recesión, si bien difícilmente podría haberse previsto que la burbuja inmobiliaria y la desaceleración económica sobrevenida en España en el año 2008 diese paso a una descomunal crisis económica, financiera y fiscal, que colocó al país al borde del rescate. De hecho, esta crisis galopante causó gran impacto en el mercado laboral y provocó despidos, expedientes de regulación de empleo, cierres, quiebras o recortes en los sistemas de protección social, que hicieron tambalear los cimientos del Estado de Bienestar que floreció tras la Transición.

Las entidades sociales han denunciado desde un primer momento el impacto que la crisis ha tenido en la ciudadanía, a partir de las demandas que reciben y las necesidades que deben atender. Los últimos informes alertan de que actualmente más de 12 millones de españoles (672.000 gallegos) viven en riesgo de pobreza. Particularmente en ciudades como Vigo, los efectos de la crisis comenzaron a notarse desde sus inicios, como pone de manifiesto el incremento en casi un 50% del número de desempleados en esta ciudad en los dos primeros años y las graves consecuencias sociales derivadas de esa situación.

Los medios de comunicación reflejan o construyen la realidad que ofrecen a la audiencia. Salvo raras excepciones, gozan de credibilidad social y generan corrientes de opinión. El objetivo de esta investigación es analizar el tratamiento periodístico de esa realidad construida en torno a la pobreza y la exclusión social que afecta a muchas personas en Vigo, como consecuencia de la crisis. Se trata de comprobar en qué aspectos se han interesado los medios y si la visión que nos han presentado concuerda o no con la realidad. Particularmente, interesa determinar si los medios de comunicación cumplen una función social (educar y servir para el libre intercambio de opiniones) y/o si sus fines son específicamente periodísticos (informar y orientar), inherentes a su razón de ser.

Palabras y frases clave: Pobreza, exclusión, crisis, desempleo, prensa.

1. INTRODUCCIÓN

La desaceleración económica sobrevenida en España en el año 2008 causó gran impacto en el mercado laboral ocasionando despidos, expedientes de regulación de empleo, cierres y quiebras de empresas y negocios familiares, así como recortes en los sistemas de protección social que articulan las distintas administraciones.

Los datos del informe elaborado por el Valedor del Ciudadano de Vigo, Luis Espada Recarey, desvelan que entre septiembre de 2008 (momento de la irrupción de la crisis y diciembre de 2010 el número de desempleados en la ciudad olívica pasó de 19.969 a 29.156. La Encuesta de Población Activa (EPA) del cuarto trimestre de 2009 cifra en 158.900 el número de personas activas en Vigo, dato superior a los 152.600 registrados un año después. En ese período de tiempo, la población ocupada en Vigo se vio reducida en un 7,4% al pasar su número de 133.500 a 123.600. En este período de tiempo, la población ocupada en Vigo se redujo un 7,4% al pasar de 133.500 a 123.600 personas. En el cuarto trimestre de

2010, la tasa de ocupación en Vigo en el sector de población de 16 o más años es del 48,1% superando así algunas de las ciudades gallegas como Ferrol (43,8%), Ourense (47,5%) o Pontevedra (46,6%), pero por debajo de Santiago (56,4%), Lugo (52,9%) o A Coruña (48,8%). Por otro lado, la pobreza y la exclusión social sobrevenida como consecuencia de la crisis actual y el desempleo son, como temas informativos, uno de los asuntos que forman parte o deberían formar parte del denominado periodismo social.

En 1948, el politólogo norteamericano Harold Lasswell señaló tres funciones clásicas de los medios de comunicación de masas: vigilancia, correlación y transmisión de la cultura. Respecto a la primera función, la de vigilancia, Lasswell considera que los medios actúan como centinelas en catástrofes y desgracias sobrevenidas como los desastres naturales, crisis económicas, guerras inminentes. Es en estos casos en los que los medios ejercen una función de vigilancia social y de control de los poderes públicos y gubernamentales. En cuanto a la función de correlación, se refiere a que los medios de comunicación, además de informar, ofrecen explicaciones e interpretaciones propias de la noticia, que llegan al ciudadano a través de los editoriales, artículos de opinión, debates, tertulias y programas de opinión. La tercera función señalada por Lasswell es la educativa o función de transmisión del legado cultural social por la cual, los medios transmiten información sobre la sociedad misma, su historia, normas, valores, aciertos y desaciertos.

Tradicionalmente según se enseña en las Facultades de Periodismo y Comunicación en España, las funciones de los medios de comunicación de masas son las de informar, formar y entretener. Martínez Albertos atribuye a los medios fines comerciales (entretenimiento), fines específicamente periodísticos (informar y orientar a los lectores) y fines sociales (educar a los públicos y servir de instrumentos para la libre comunicación de las opiniones.

Sobre la última de las funciones, la función social, el profesor Martínez Albertos señala que "la función social de los medios periodísticos es hacer posible que los ciudadanos vean satisfactoriamente atendido su derecho individual y colectivo a una información técnicamente correcta.

La naturaleza de estos temas exige a los periodistas que le otorguen un tratamiento profundo y responsable y una contextualización que facilite su comprensión y más concienciación en la opinión pública.

En todo caso, el periodismo responsable responde siempre a una serie de principios y valores recogidos en códigos éticos o deontológicos[1] , en declaraciones profesionales e incluso institucionales. Entre éstos principios o códigos de conducta podemos citar los de verdad, precisión, independencia, equidad, imparcialidad y responsabilidad.

En España, la Federación de Asociaciones de la Prensa de España (FAPE) establece en su Código Deontológico de la profesión periodística, (aprobado en la Asamblea Extraordinaria de Sevilla en 1993), es el documento de deontología periodística más importante para los profesionales de la información que ejercen su actividad en nuestro país. En este código se recogen las normas y principios éticos que guían la labor profesional de sus asociados en aras al respeto de los derechos fundamentales. En su artículo siete dice que: "El periodista extremará su celo profesional en el respeto a los derechos de los más débiles y los discriminados. Por ello, debe mantener una especial sensibilidad en los casos de informaciones u opiniones de contenido eventualmente discriminatorio o susceptibles de incitar a la violencia o a prácticas humanas degradantes".

A la construcción de la realidad que hacen los medios se suman también otros factores. La feroz competencia por ganar más lectores y la preferencia por noticias impactantes relegan a un segundo plano los temas sociales porque se considera que son cuestiones que enganchan menos al lector o venden menos. Este estudio pretende averiguar si los medios prestan o no escasa atención a este tipo de problemas sociales y/o realizan un tratamiento superficial dando escasa relevancia a estas cuestiones cuando publican algo, a no ser que se trate de un tema sensacionalista, que despierte morbosidad, que genere audiencias o venda portadas.

Así, pues, este trabajo de investigación pretende identificar el discurso que transmiten los medios de comunicación más influyentes en Vigo sobre la pobreza y la exclusión social sobrevenida como consecuencia de la crisis económica de principios del siglo XXI. En concreto, se analizan las piezas informativas aparecidas en tres medios de comunicación que se corresponden con el formato de prensa escrita (Faro de Vigo, La Voz de Galicia y Atlántico Diario).

El ámbito geográfico de investigación se circunscribe con carácter exclusivo a la ciudad de Vigo, municipio de España perteneciente a la provincia de Pontevedra, situada en la comunidad autónoma de Galicia, al noroeste de España. La elección de la ciudad olívica para realizar el trabajo que proponemos

[1]La palabra castellana deontología procede, como tantas otras en nuestro idioma, de la lengua griega. En concreto de δεον, que es el participio neutro del verbo impersonal δει, que significa "lo obligatorio, lo justo, lo adecuado, lo conveniente, el deber". El otro vocablo griego es λογος, que significa "tratado, investigación, estudio, ciencia". La deontología, en su definición etimológica o nominal, representa la generalidad de la perspectiva ética, refiriéndola a los deberes, reglas y normas de una profesión.

no es casual ya que Vigo reúne una serie de requisitos que la convierten en lugar idóneo para realizar este estudio. Es el municipio más poblado de Galicia y el decimocuarto de España, con 294.997 habitantes empadronados en el año 2014, de los cuales 200.832 lo estaban en la ciudad propiamente dicha y los restantes 93.971 habitantes se distribuyen en 16 parroquias periurbanas y una parroquia rural. Vigo ha sido también tradicionalmente la ciudad más industrializada de Galicia con sectores pujantes como la industria pesquera, la automoción y la construcción naval.

Respecto al ámbito temporal de la investigación, la fase empírica se abarca el periodo comprendido entre octubre 2009 y marzo de 2010; es decir, cuando los daños colaterales de crisis eran ya más que evidentes. Los tres medios escogidos para realizar la investigación (Faro de Vigo, Atlántico Diario y La Voz de Galicia) son los tres periódicos más influyentes de la ciudad en número de lectores y en repercusión mediática. En la selección de los mismos también se ha tenido presente que los archivos de prensa escrita pueden consultarse fácilmente y trabajar con ellos sin restricción de tiempo, mediante el servicio de prensa digitalizada en Internet o los ejemplares de Faro y Atlántico disponibles en la Biblioteca Central de Vigo. No es tan sencillo, sin embargo, el acceso a los archivos sonoros de las emisoras de radio o a las imágenes de las televisiones locales. Hay que solicitar permisos y ceñirse a los tiempos que marquen. Además, los medios audiovisuales no suelen guardar archivos relacionados con este tipo de informaciones y, por otra parte, centran principalmente su atención en personajes del ámbito político o económico.

De acuerdo a la problemática planteada, y en base a las inquietudes y a los intereses explicados hasta el momento, se establece como objetivo general de esta investigación es estudiar el discurso que han difundido los tres medios seleccionados sobre la pobreza y exclusión social sobrevenida como consecuencia de la crisis económica. El ámbito de temporal de este estudio, la fase empírica abarca el período comprendido entre octubre de 2009 y marzo de 2010Además se trata de comprobar en este trabajo si, mediante las noticias publicadas estos medios se han ceñido o no al cumplimiento de los fines específicamente periodísticos (informar y orientar a la opinión pública, educar y servir para el libre intercambio de opiniones) o si por el contrario se hacen escaso eco de estos temas porque no encajan en los intereses particulares de las empresas periodísticas que gestionan los diarios seleccionados.

2. MATERIAL Y MÉTODOS

Para desarrollar este estudio, se ha elaborado una base de datos a partir de las noticias recogidas en los tres medios citados (Faro de Vigo, La Voz de Galicia y Atlántico Diario) en el período considerado (desde octubre de 2009 a marzo de 2010. En principio se pensó recopilar todos los titulares relacionados con la temática tratada, pero un primer estudio piloto puso de manifiesto el tamaño excesivo de esta muestra, por lo que resultaría poco manejable a la hora de plantear la selección de variables y la realización del análisis. Por ello se decidió acotar la recogida de datos, utilizando un método de muestreo sistemático, cuyo resultado final dio lugar a la recopilación de 116 noticias sobre la pobreza y exclusión social motivada por la crisis y el desempleo. La tabla siguiente recoge las principales características de esta investigación.

Estudio:	'Pobreza y exclusión social sobrevenida durante la crisis económica y su tratamiento en la prensa viguesa en el período 2009-2012'.
Período:	Seleccionado por muestreo sistemático. Comprende 6 semanas (5-11 de octubre de 2009, 2-8 de noviembre de 2009, 21-27 de diciembre de 2009, 4-10 de enero de 2010, 22-28 de febrero de 2019, 15-21 de marzo de 2010).
Diarios:	• La Voz de Galicia, edición Vigo • Faro de Vigo • Atlántico Diario.
Muestra:	116 noticias

Tabla 1: Ficha técnica del estudio

Con este material, se elaboró una base de datos de distintas variables, a través de las cuales se trató de recoger la información principal de cada noticia, esencial para su clasificación y tratamiento estadístico posterior. Entre otras, se consideraron las siguientes variables: fecha de publicación, extensión, sección, género periodístico, inclusión de foto y/o gráfico, tema y tono de la noticia, protagonista y fuente de la información. Se diseñó una aplicación en formato Excel para la grabación de datos, que luego fueron depurados y analizados utilizando el mismo programa. En cuanto a la información recopilada, se llevó a cabo un análisis exploratorio gráfico, cuyos principales resultados se muestran en la sección siguiente.

3. RESULTADOS Y DISCUSIÓN

Comenzando por considerar el diario en el que se han publicado las informaciones seleccionadas, la Figura 1 recoge la distribuci´n de la muestra por medio.

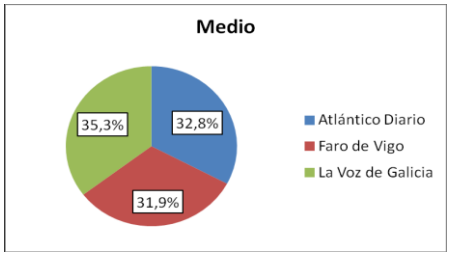


Figura 1. Distribución de la muestra por medio

Los resultados obtenidos permiten concluir que el número de piezas insertadas en los tres diarios es bastante equilibrado, con diferencias porcentuales entre ellos que no superan el 3,4%. La Voz es el rotativo que más noticias publicó en el periodo analizado y Faro de Vigo, el que menos cabida ha dado a estos temas en sus páginas.

En el gráfico siguiente, se presentan los datos relativos al día en que se ha publicado la información, donde se ha diferenciado entre día laborable (sem), sábado o domingo (finsem) o día festivo.

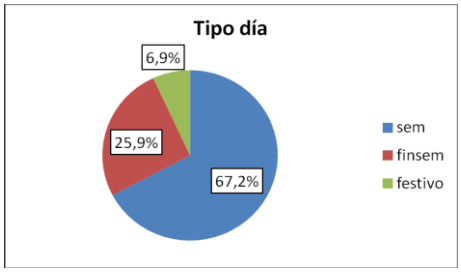


Figura 2: Distribución de muestra por día de publicación

Se observa en al Figura 2 que los temas analizados tiene mayor repercusión en días laborales (67,2%)que en fines de semana (25,9%). Los festivos son los días menos prolíficos en la publicación de noticias relacionadas por la pobreza y la exclusión (6,9%), si bien debemos tener presente que en el período analizado solo se computaron 3 festivos. De ahí que podría tratarse de temas recuerrentes para surtir al medio de información en días de escasa actividad mediática.

Otro aspecto que se ha contemplado en este estudio es el espacio que ocupan las noticias dentro del medio, para lo que se ha utilizado una escala porcentual, de manera que el 100% se corresponde con una página entera y las restantes extensiones se presentan de forma proporcional.

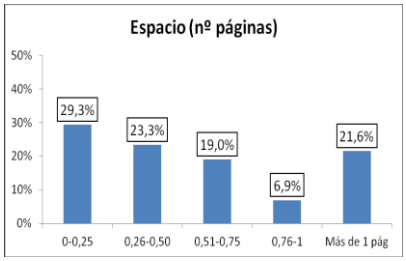


Figura3: Distribución de la muestra por espacio

Los resultados representados en la Figura 3 indican que aproximadamente la mitad de las informaciones (52,6%) abarcan a lo sumo media página, aunque en el otro extremo existe un porcentaje destacado /21,6%) cuya extensión es superior a una página. En el priemr grupo se concentran princpalmente notas breves que dan cuenta de un hecho o acontecimiento sin profundizar en él y sin

ofrecer voz del protagonista, mientras que las oticias más amplias suelen aparecer en cuadernillos o suplementos especiales o incluso en publi-reportajes institucionales.

La relevancia que los medios otorgan a este tipo de informaciones se refleja también en el número de portadas que les conceden. Los datos obtenidos a este respecto se resumen en el gráfico que se muestra a continuación.

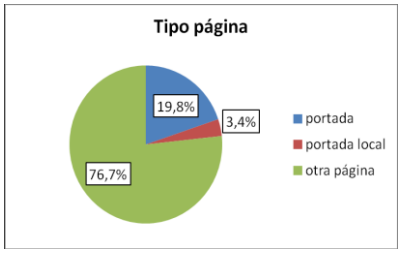


Figura 4: Distribucion de la muestra por tipo de página

Durante el período analizado, los titulares relacionados con la pobreza y la exclusión, el paro y la crisis solo fueron portada del diario en el 19,8% de los casos, o de ediciones locales en el 3,4%.Esto permite concluir la escasa importancia o relevancia que otorgan los diarios vigueses a estas temáticas. Otra variable considerada en esta investigación recoge la sección del periódico en la que han salido publicadas las noticias incluidas en la muestra, cuya distribución se presenta en la Figura 5.

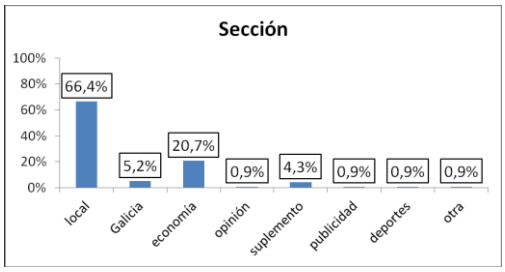


Figura 5: Distribución de la muestra por sección

Se desprende del gráfico anterior que la mayor parte de la información sobre pobreza y exclusión se incluye en la sección local (66,4%), seguida a distancia por la sección de economía (20,7%) o la didicada a Galicia (5,2%) lo cual es bastante llamativo, teniendo en cuenta que se trata de la ciudad que siempre ha sido cnsiderada como el motor económico de la comunidaa autónoma. Destaca también el reducido peso que tienen este tipo de noticias en editoriales y columnas de opinión (0,9%) lo que demuestra el escaso interés que despiertan en los tres medios analizados, ya que cuando realmente les preocupa un tema y quieren hacer eco sobre el , se recurre típicamente a artículos de opinión.

Los diferentes textos analizados sitúan la información en distintos lugares de referencia, tal y como se resume en la Figura 6.

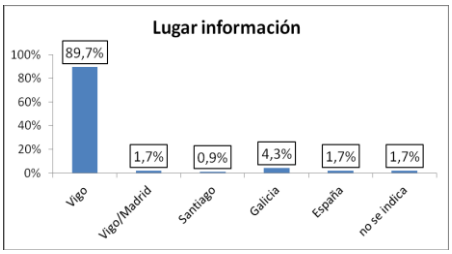


Figura 6: Distribución de la muestra por lugar de referencia de la información

La mayoría de las publicaciones se corresponden con información vinculada a Vigo (89,7%) y, por tanto, en el tratamiento de la pobreza y la exclusión social otorgan el mayor protagonismo a esta ciudad. Un porcentaje muy reducido de informaciones toman como referencia la comunidad gallega (4,3%) o España en general (1,7%), para abordar temáticas de tipo administrativo o relacionadas con empresas y sectores que trascienden el ámbito local.

La Figura 7 arroja conclusiones interesantes sobre la distribución de los titulares analizados por género informativo, desde el pequeño breve o reseña, pasando por el chiste, lanoticia, el reportaje o la entrevita y los géneros de opinión (editorial y artículo).

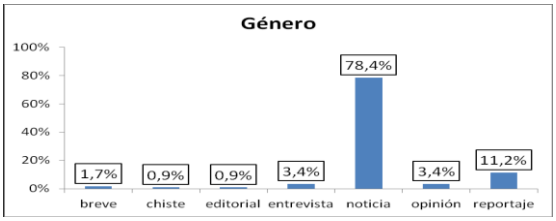


Figura7: Distribución de la muestra por género informativo

Así, cabe destacar que los rotativos vigueses han situado la pobreza como un tema cada vez más destacado en sus agendas, pero en la cobertura informativa se echa en falta más rigor y una mayor implicación del periodista. De hecho, el género que prevalece es la noticia (78,4% de los casos) pero apenas se utilizan las reseñas breves (1,7%) o los artículos de opnión (3,4%). En escasas ocasiones se recurre a dos de los géneros de mayor análisis, el reportaje (11,2%) o la entrevista (3,4%). De ahí que los lectores reciben el "que" de la noticia pero no el "porqué", lo que indica resistencia o poca implicación del periodista y /o del medio para profundizar en estos temas.

La fuente de información presenta gran variedad en el conjunto de noticias consideradas, como se muestra en la Figura 8.

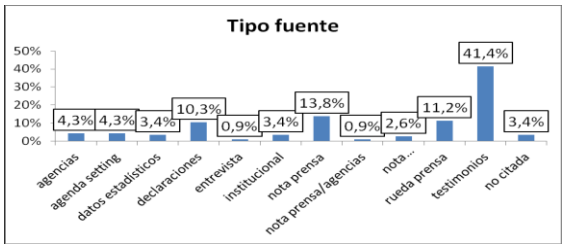


Figura 8: Distribución de muestra por fuente de información

Del gráfico anterior se deduce que la información proviene principalmente del testimonio directo (41,4%), bien del protagonista de la información o de sus representantes (asociaciones sindicatos, ONG, políticos e instituciones) Le siguen a gran distancia otras fuentes como la nota de prensa (13,08%), la rueda de prensa (11,2%) o las declaraciones realizadas en actos previamente convocados (10,3%). En menor medida los diarios se nutren de los teletipos de agencia (4,3%) que presentan un estilo más aséptico, de la agenda setting de los políticos (4,3%), y de los datos estadísticos (3,4%) que se facilitan en las notas y comunicados de prensa institucionales. Otras veces no se cita la fuente (3,4%) y, excepcionalmente se mezclan datos recogidos de notas de prensa/ y a la entrevista (0,9%). Esto indica que aunque no se profundiza en los temas con géneros de opinión o de mayor análisis como el reportaje o la entrevista, a los medios les interesa en cierta medida ofrecer información puntual y trabajada sobre estas cuestiones a sus lectores.

El número de fuentes de información, consultadas por el periodista para elaborar las piezas informativas, ha sido también una de las variables analizadas en esta investigación.

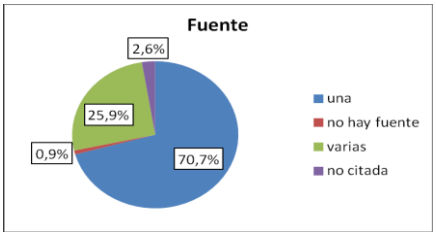


Figura 9: Distribución de la muestra por número de fuentes

La Figura 9 muestra que en la mayor parte de los casos se consulta una sola fuente (70,7%), lo que indica que el periodista no suele contrastar la información que recibe. Si el testimonio es la fuente dominante, como se desprende de la Figura 8, la noticia que se deriva del mismo se limita a publicarlo sin verificar u ofrecer otro punto de vista del hecho narrado. Por otra parte, en aquellos casos en los que se utiliza la nota de prensa, el teletipo o, incluso el dato estadístico, tampoco se contrasta la información con otras fuentes alternativas. Es muy probable que la celeridad y el trabajo a contrarreloj al que se ve sometido el periodista en su día a día no le deje tiempo suficiente para realizar una labor más precisa de consulta. En un porcentaje muy reducido de casos no se cita la fuente consultada (2,6%) o no hay fuente (0,9%).

Tambien se hace referencia en este estudio a los protagonistas de la información, con el objeto de establecer en qué medida tienen presencia en las noticias las personas o los sectores más afectados por la crisis, así como las entidades y organizaciones que trabajan con ellos para mejorar su situación.

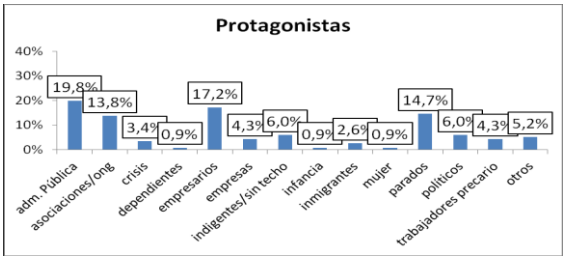


Figura 10: Distribución de la muestra en función de sus protagonistas.

En las noticias sobre pobreza, exclusión, crisis y desempleo, que publican los medios de comunicación, predomina el discurso que interesa difundir a las administraciones públicas (19,8%) o a los empresarios de sectores aquejados por la crisis (17,2%). En menor medida, obtienen cierta visibilidad las asociaciones y organizaciones que trabajan con los colectivos afectados por la pobreza y la exclusión (13,8%) y los propios desempleados (14,7%). El reflejo social que tienen los grupos empresariales a través de los medios es positivo y logra que los lectores se solidaricen con su situación. No obstante, es una visión sesgada de la realidad porque apenas se ofrece la otra cara de la moneda, la de los trabajadores de esos sectores cuya situación, en muchos casos es precaria, como se observa en el gran número de noticias que hacen referencia a cierres de empresas, expedientes de regulación o movilizacion y concentraciones de protesta protagonizada por trabajadres en situación precaria o que ven reducidos su derechos laborales. El reflejo social que tienen estos colectivos en Vigo no es nunca positivo, sino reivindicativo en unos casos o aséptico en otros, como ser verá más adelante al estudiar el tono de las informaciones publicadas. En cualquier caso, la existencia de de entidades que ejercen de portavoces les permite visibilizar o argumentar públicamente sus necesidades, aunque el espacio que se les otorgue sea escaso o no se busque implicar en ello a las instituciones que pueden dar un giro a la realidad que sufren. En el siguiente gráfico se presenta la distribución de la muestra de acuerdo con la iniciativa de la que parte su publicación, es decir, si es decisión del periodista buscar información sobre el tema, o si ésta le llega por otras vías (rueda de prensa, nota de prensa, teletipo de agencia, etc).

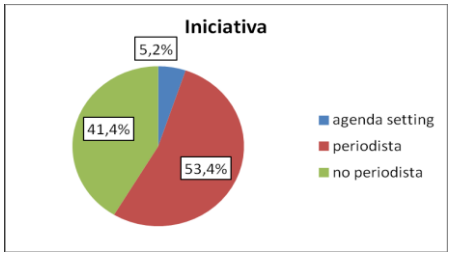


Figura 11: Distribución de la muestra en función de la iniciativa de la que parte

Se desprende de estos resultados que los profesionales de la información muestran interés por las noticias relacionadas con la pobreza, la exclusión, la crisis y el desempleo. De hecho, más de la mitad de las piezas publicadas son iniciativa de estos profesionales (el 53,4%). Sin embargo, se observa también que la agenda de los políticos de las distintas administraciones marca cada vez más los temas que publican los periódicos, puesto que el 41,4% de las noticias proceden de la agenda de estos medios. Por

ello, el periodista debe ingeniárselas para hacer más rápido su trabajo y aprovechar las comparecencias públicas para formular preguntas sobre temas que forman parte de la agenda-setting de los medios.

La Figura 12 recoge la clasificación de las noticias por tema de información, de la que se concluye que hay temas recurrentes o primarios y otros quedan en segundo plano.

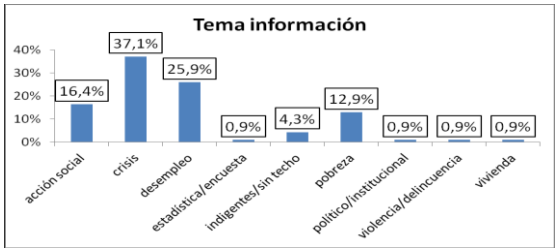


Figura12: Distribución de la muestra por el tema de la información.

Se observa que la temática que prevalece por encima de todas se relaciona con la crisis (37,1%) seguido por las noticias sobre desempleo (25,9%), lo que denota que son las que más interesan a los medios. Menor repercusión tiene la información sobre acción social (16,4%), cuestiones relacionadas con la pobreza (12,9%) o la realidad de los indigentes o personas sin techo (4,3%). Son excepcionales lo titulares relativos a la vivienda (0,9%), la violencia o la delincuencia (,0,9%) o los datos estadísticos sobre sectores o actividades económicas (0,9%). Estos resultados nos indican que los medios intentan reflejar que la crisis y el paro son reales, si bien las informaciones se quedan ahí, sin profundizar en las consecuencias de esa realidad o sin aportar soluciones o intentar la implicación de quien puede aportarlas. Nuevamente la funcion social que se presupone a los medios queda en entredicho. Pudiera pensarse como en caso de las fuentes de la información que, la premura les impide realizar con más precisión su trabajo.

Otro aspecto que se ha trabajado en este estudio es el discurso o tono empleado por el periodista en la información publicada sobre la pobreza y la exclusión social.

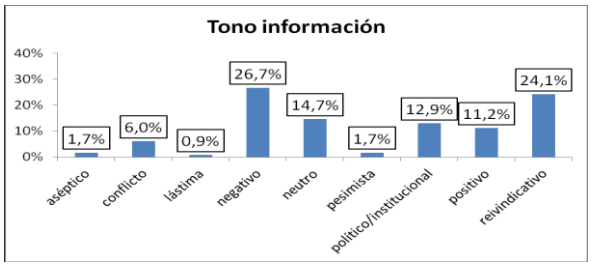


Figura 13: Distribución de la meustra por el tono de la información

Del gráfico anterior se puede concluir que en este tipo de noticias predomina en el discurso el tono negativo (26,7%), seguido del tono reivindicativo (24,1%) lo que indica que el periodista se solidariza o muestra cierta empatía con las personas afectadas por estas situaciones. El tono político institucional ocupa el tercer lugar (12,9%) y predomina en aquellas informaciones que proceden de las administraciones y que mantienen un tratamiento más aséptico o neutral.

La Figura 14 presenta la distribución de los textos analizados de acuerdo con el carácter pasivo o activo que se otorga al sujeto de la misma, con idea de reflejar el grado de visibilidad o presencia de los protagonistas de esta realidad social en las noticias que se refieren a ellos.

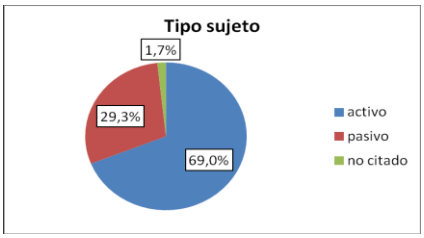


Figura 14: Distribución de la muestra según el sujeto de la información

En la mayoría de los casos aparece o se cita a un protagonista (69,0%), mientras que en el 29,3% no se visibiliza sujeto alguno o se da la información sin su testimonio o concurrencia. En un porcentaje muy reducido no se cita ningún tipo de sujeto que protagoniza la información (1,7%).

En las noticias seleccionadas, los textos se suelen acompañar de fotografías y/o de otros elementos gráficos, como se observa en el gráfico siguiente.

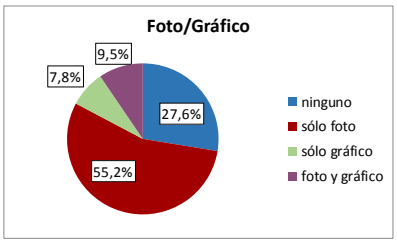


Figura 15: Distribución de la muestra según presencia de fotografías y/o elementos gráficos.

Aunque la imagen es importante para ilustrar la información, en el 35,3% de las noticias sobre pobreza y exclusión no aparece ninguna fotografía. Sin embargo, el 55,2% de las noticias sí se acompañan de imágenes y en el 9,5% incluso aparece además algún otro elemento gráfico. La infografía forma parte del 17,3% de las noticias, para reflejar en general datos estadísticos que pueden ayudar a contextualizar el tema aunque, en ocasiones, no se profundiza en su significado.

4. CONCLUSIONES

En este estudio se ha tratado de obtener una primera aproximación al discurso que difunden tres medios de comunicación de gran difusión en Vigo (Faro de Vigo, Atlántico Diario y La Voz de Galicia), para analizar el tratamiento que hacen de la pobreza y la exclusión social sobrevenida como consecuencia de la crisis económica. Los datos obtenidos indican que estos diarios sitúan estas problemáticas como temas cada vez más destacados en sus agendas. Sin embargo en la cobertura informativa se echa en falta más profundidad en el contenido, que ponga de manifiesto la realidad de cada situación, las razones que la han generado e incluso la posible propuesta de soluciones. La falta de implicación se observa también en la ausencia de artículos de opinión y editoriales que aborden esta temática, prevaleciendo en general la noticia sobre otros géneros periodísticos más analíticos, como la entrevista o el reportaje de carácter humano. Básicamente se recurre a la presentación de titulares, complementados con el breve testimonio de alguno de los protagonistas, cuando han estado presentes en actos públicos, concentraciones, manifestaciones o ruedas de prensa que formen parte de la agenda política o institucional del día. No obstante, no suele haber seguimiento de los temas para localizar fuentes alternativas que acompañen a esos testimonios o que incidan en esas problemáticas en los días posteriores.

El tono negativo y el reivindicativo son los que más destacan en el tratamiento de esta información, lo que denota solidaridad de los periodistas con la población afectada por la crisis, aunque también en otros casos se limitan a la redacción de textos con un lenguaje más propio del ámbito político-institucional. En la presentación de estas noticias es habitual la presencia de gráficos acompañando a los textos informativos u otros elementos gráficos, si bien estos últimos suelen aportar datos estadísticos sobre los que se entra en profundidad. Analizando la temática tratada, el estudio recoge muchas piezas que hablan de la crisis y del desempleo pero es mucho menor el volumen de informaciones que ponen el enfoque en los protagonistas que viven el drama de estas situaciones: las personas pobres, sin techo o las que son deshauciadas de sus viviendas. En este sentido cabe demandar más responsabilidad social a los medios para que, ejerciendo su función de servicio público, realicen paralelamente una contextualización más profunda y responsable de esta problemática social y no se limiten solamente a sensibilizar al lector y a lograr mayor concienciación y empatía hacia este tipo de situaciones, tan frecuentes en el contexto actual. Hablar abiertamente de las causas de la pobreza y exclusión, de sus consecuencias, denunciar injusticias y proponer soluciones deberían ser prácticas habituales en los medios, particularmente cuando se abordan problemas que afectan a muchas personas, que sufren directamente las consecuencias del paro y de la crisis y de las que rara vez se escucha su voz.

Así pues, del estudio del discurso utilizado por los tres periódicos más influyentes en Vigo sobre la pobreza y la exclusión social sobrevenida por la crisis, se desprenden ciertas carencias respecto al

cumplimiento de la función social y de los fines específicamente periodísticos (informar y orientar a la opinión pública, educar y servir para el libre intercambio de opiniones) que son inherentes a los medios de comunicación. Se echa en falta más rigor y una mayor implicación acorde con la función social y cabe demandar más especialización a los medios en el tratamiento de esta información que eviten caer en errores como la superficialidad en el manejo de este tipo de temáticas sociales.

REFERENCIAS

Bentolila, S. y Jansen, M. (2012) La reforma laboral de 2012: Una primera evaluación. Apuntes FEDEA-Laboral, 14.

Carrascosa Puertas, L. (2014) Como la prensa nos cuenta la crisis. Editorial Arquitectura Viva AV.

Conde-Ruiz, J. I., Felgueroso, F. y García-Pérez, J. I. (2011) Reforma Laboral 2010: Una primera evaluación y propuestas de mejora. Revista de economía aplicada, 19, 147-182.

Contreras, B. (2012) Factores intervinientes en los procesos de exclusión de las personas sin hogar. En Cáceres, M. D. y Lucas, A. (Ed.): Crisis y cambios en la sociedad contemporánea. Comunicación y problemas sociales. Editorial Fragua, Madrid.

Contreras B. y Alguacil A. (2012) Pobreza y exclusión social en la Comunidad de Madrid. En Lucas, A., Cortés, C. y Cáceres, M. D. (Ed.): La Realidad Social de Madrid. Una perspectiva sociológica, Editorial Fragua, Madrid.

Dearing, J. W. y Rogers, E. (1996) Agenda-setting Vol. 6. Sage Publications.

Espada Recarey, L. (2011) Pobreza y exclusión social: perspectiva comparada y estudio específico del municipio de Vigo. Grupo SR.

Lázaro González, I. E. (2012) Pobreza y exclusión social de la Infancia. Universidad Pontificia de Comillas.

Laparra, M. (2008) Exclusión social en España: un espacio diverso y disperso en intensa transformación (Vol. 24). Cáritas Española.

Laparra, M., Navarro, M. L. y Eransus, B. P. (2010) El primer impacto de la crisis en la cohesión social en España (Vol. 32). Cáritas Española.

Martínez Albertos, J. L (1992) Curso general de redacción periodística. Madrid, Paraninfo.

Minujin, A., Delamónica, E. y Davidziuk, A. (2006) Pobreza infantil: conceptos, medición y recomendaciones de políticas públicas. Facultad Latinoamericana de Ciencias Sociales, Sede Académica Costa Rica.

Roba, O., Cerutti, A., Canetti, A., Parafita, D. y Duarte D. (2013) Pobreza infantil, desigualdad y políticas sociales en el Uruguay. En Minujin, A. y Llobet, V. (Ed.): Estrategias regionales de medición de pobreza en niñez. Secretaría Nacional de Niñez, Adolescencia y Familia Ministerio de Desarrollo Social. Presidencia de la Nación Argentina

Spicker, P., Álvarez Leguizamón, S. y Gordon, D. (2009) Definiciones de pobreza: doce grupos de significados, dentro de pobreza. Un glosario internacional. Buenos Aires, Colección CLACSO-CROP.

Torres Gutiérrez, F. J. y García Leiva, F. (2012) La espacialidad urbana del fenómeno de las personas sin hogar. Apreciaciones a partir del caso de Sevilla. La población en clave territorial: procesos, estructuras y perspectivas de análisis. Actas XIII congreso de la población española. Ministerio de Economía y Competitividad, Dirección General de Investigación y Gestión del Plan Nacional de I+D+i.

Uribe, J. y Alonso, S. (2010) Personas en situación de sin hogar en Barcelona: Perfiles, estado de salud y atención sanitaria. Introducción. Labor Hospitalaria: Organización y Pastoral de la Salud, 295, 8-14.

La internacionalización como factor de supervivencia de las empresas del área de Vigo

Manuel Meijide Vecino, Antonio Vaamonde Liste, Ricardo Luaces Pazos, Patricio Sánchez Bello.

Universidad de Vigo

RESUMO

La vida de las empresas, y su extinción, han sido estudiadas con distintos enfoques en la literatura especializada reciente, y son muchos y diversos los elementos que deben considerarse para la comprensión de las causas de longevidad y supervivencia. En este trabajo se analiza la edad y algunos factores asociados a la supervivencia de las empresas del área de influencia de Vigo, mediante técnicas estadísticas como el método de Kaplan-Meier, regresion de Cox, regresión robusta, y razón de riesgo. Los resultados obtenidos destacan la influencia de la internacionalización de las empresas como factor de protección frente a las dificultades generadas por la crisis.

Palabras e frases chave: firm survival, firm growth, firm failure, life duration

1. INTRODUCIÓN

Se utiliza una muestra aleatoria de 928 empresas del área de influencia de Vigo, con datos de las cuentas anuales depositadas en el registro mercantil, obtenidas mediante la colaboración de la Base de Datos Ardán, gestionada por el Consorcio de la Zona Franca de Vigo. Se dispone de 231 variables: balance, cuenta de resultados, distintos ratios, año de constitución, sector de actividad, y otras características de las empresas. Toda la información se refiere al período 2008-2012. Para los cálculos se utiliza el programa estadístico R.

2. RESULTADOS

La variable principal en este estudio es la edad de la empresa, expresada en años. Algunos estadísticos descriptivos nos permitirán una primera aproximación:

MEDIA	18.13
DESVIACIÓN TÍPICA	9.89
RANGO INTERCUARTÍLICO	10
MÍNIMO	3
PRIMER CUARTIL Q1	12
MEDIANA	16
TERCER CUARTIL Q3	22
MÁXIMO	112
TAMAÑO DE MUESTRA	928

Tabla 1: Edad de las empresas de Vigo

Un histograma facilita la apreciación de las características más relevantes:

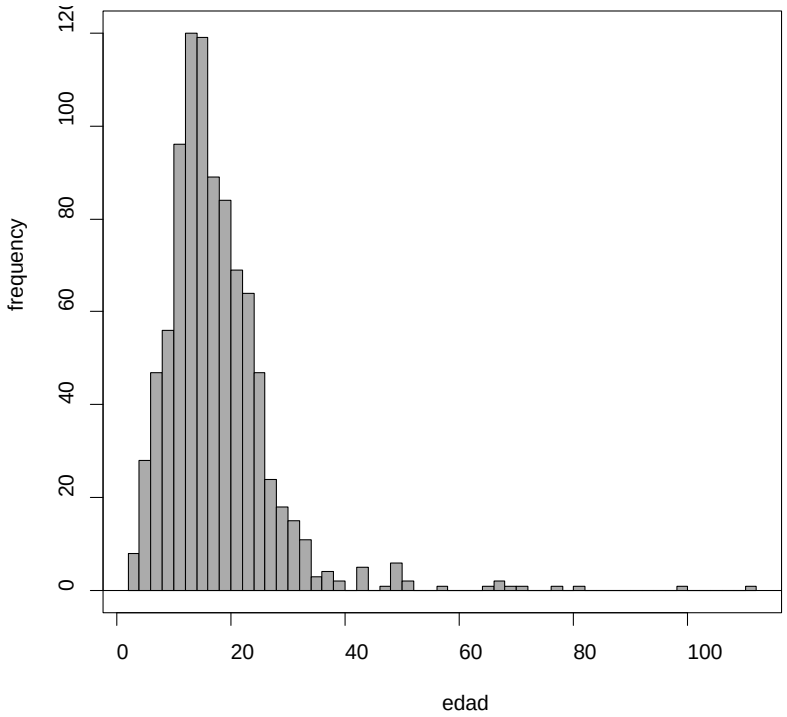


Figura 1: Edad de las empresas de Vigo.

Como se puede observar la práctica totalidad de las empresas tienen actualmente edades comprendidas entre 0 y 40 años (la mitad no supera los 16), con algunas notables excepciones: una o dos empresas en la muestra alcanzan el siglo de vida.

Mediante el método de Kaplan-Meier, y con los datos de los años recientes (2011-2012), construimos la curva de supervivencia, que muestra para cada valor de edad la proporción (probabilidad estimada) de supervivencia. Añadimos a la curva un factor explicativo: el carácter de empresa exportadora o internacionalizada:

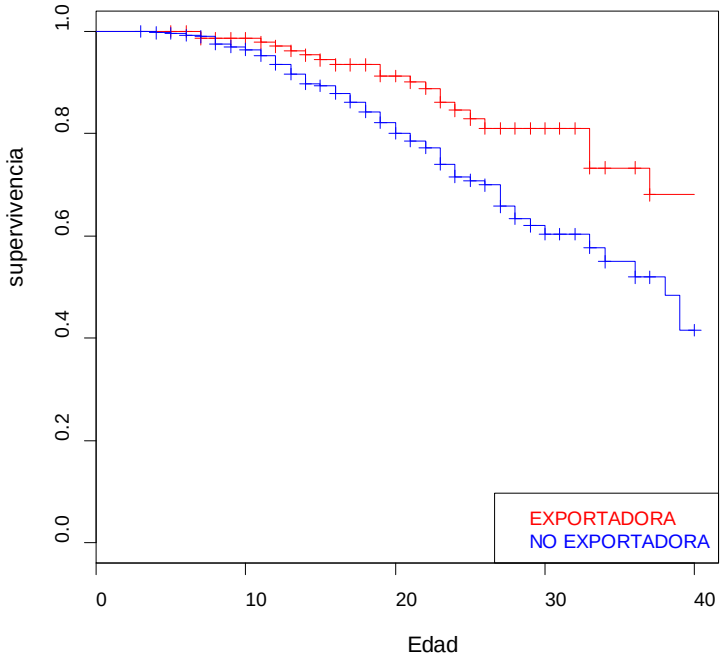


Figura 2: Curva de supervivencia de las empresas del área de Vigo

La curva correspondiente a la empresa exportadora está sistemáticamente por encima de la correspondiente a la empresa no exportadora, mostrando para todos los valores de edad una mayor proporción de empresas supervivientes: la vida de las empresas exportadoras es más larga, y el riesgo de desaparición menor.

Las tablas de vida, con las cuales se construye el gráfico anterior, muestran esa diferencia, indicando para cada edad el porcentaje o probabilidad estimada de supervivencia:

EMPRESA EXPORTADORA

EDAD	EN RIESGO	DESAPARECIDAS	SUPERVIVENCIA
0	149	0	1.000
10	138	2	0.986
20	79	8	0.913
30	29	7	0.810
40	12	3	0.681
50	7	1	0.624

EMPRESA NO_EXPORTADORA

EDAD	EN RIESGO	DESAPARECIDAS	SUPERVIVENCIA
0	779	0	1.000
10	684	26	0.965
20	238	78	0.801
30	38	27	0.604
40	12	6	0.415
50	8	0	0.415

Tabla 2. Tablas de vida: empresa exportadora y no exportadora

Con los datos recientes, a título de ejemplo, se aprecia en la tabla 2 que a los 40 años de vida la supervivencia de las empresas exportadoras es del 68,1%, mientras que entre las no exportadoras es solamente el 41,5%

Aplicamos a los datos el modelo de regresión de Cox, o modelo de riesgos proporcionales:

```
                              coef  exp(coef)  se(coef)    z       p
exportadora[No exportadora] 0.6335  1.8842    0.2293   2.763  0.00573
```

El coeficiente de la regresión correspondiente a la empresa no exportadora es 0,6335, positivo, lo que indica que el riesgo de desaparición de la empresa aumenta entre las empresas no exportadoras. La razón de riesgo (HR, Hazard Ratio), exponencial del coeficiente anterior, es 1,88: el riesgo de desaparición se multiplica por 1,88 si la empresa es no exportadora con respecto a la que exporta. El valor p = 0.0057, menor que el nivel de significación habitual (0,05) indica que esa diferencia entre ambos grupos de empresas es estadísticamente significativa. Un gráfico adicional permite valorar mejor esa relación:

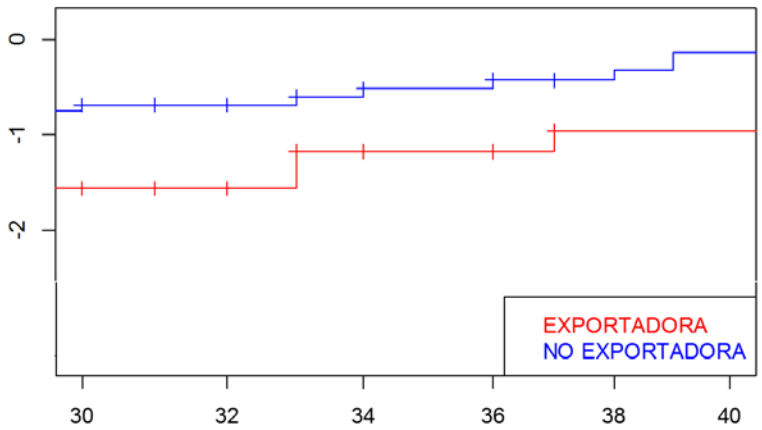


Figura 3: Tasas de riesgo (escala log-log) en función de la edad

Se representa en color rojo la tasa de riesgo de las empresas exportadoras, menor sistemáticamente que la de las empresas no exportadoras representada en color azul. La escala es doblemente logarítmica, ya que de ese modo los puntos deben estar aproximadamente sobre una recta si los riesgos son proporcionales a lo largo del tiempo (edad de la empresa); si se comportan de igual modo en ambos grupos (exportadoras y no exportadoras) las rectas serán además paralelas; ambas características se cumplen razonablemente. El riesgo de desaparición de una empresa aumenta con su edad, pero para cualquier edad es sistemáticamente mayor (casi dos veces mayor) entre las empresas que no exportan.

Aplicamos un test específico para valorar la hipótesis de riesgos proporcionales requerida por el modelo de regresión de Cox:

```
> cox.zph(cox1)
                                  rho    chisq     p
exportadora[T.No exportadora]  -0.109   1.85    0.173
```

El coeficiente de correlación rho entre el tiempo de supervivencia transformado y los residuos de la regresión es prácticamente cero, y el valor P indica que se puede aceptar la hipótesis de riesgos proporcionales. Un gráfico adicional, que representa el valor del coeficiente estimado de la regresión a lo largo del tiempo, muestra que se mantiene aproximadamente constante a lo largo del rango útil de la edad de la empresa. Las dos líneas punteadas indican los límites del intervalo de confianza, es decir la fiabilidad de la estimación.

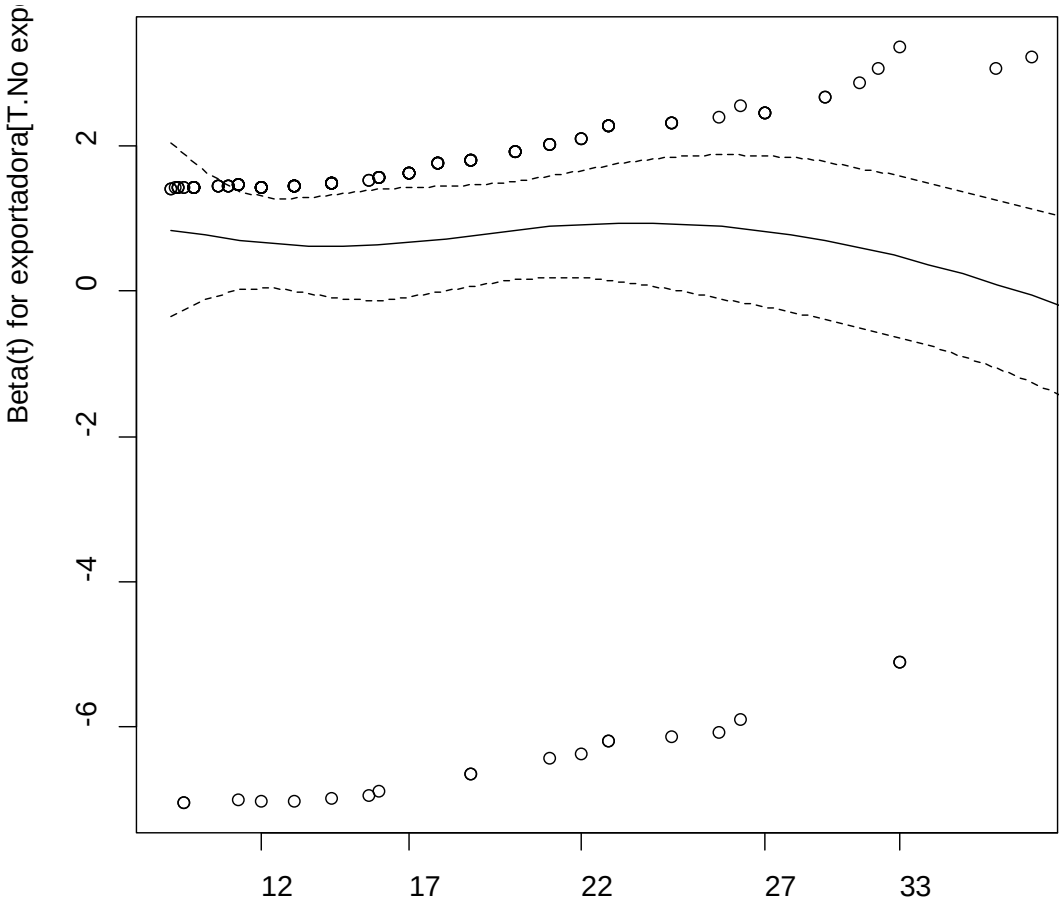


Figura 4: Coeficiente de la regresión de Cox en función de la edad de la empresa

Utilizamos adicionalmente otros dos métodos de estimación: máxima verosimilitud parcial, y un método robusto (Bednarski, 1993; 2012); este último permite reducir el efecto de los valores atípicos sobre la regresión:

```
> cox <- coxr(os~exportadora); cox

Partial likelihood estimator
                                 coef  exp(coef) se(coef)     p
exportadora[T.No exportadora]  0.657    1.93     0.229    0.00414
Wald test=8.22 on 1 df, p=0.00414


Robust estimator
                                coef    exp(coef) se(coef)    p
exportadora[T.No exportadora]  0.806      2.24    0.292   0.00574
Extended Wald test=7.63 on 1 df, p=0.00574
```

Los resultados de la regresión robusta muestran una razón de riesgos (2,24) significativa, ligeramente más alta que con el método anterior. El test de Wald indica también que el modelo de regresión es significativo.

3. CONCLUSIONES

La metodología estadística del análisis de supervivencia aplicada a las empresas permite mejorar nuestro conocimiento sobre los procesos de desarrollo y riesgo de desaparición, de especial interés en tiempos de crisis. Para las empresas de Vigo y su área de influencia, una muestra grande permite medir el efecto sobre ese riesgo de factores como la internacionalización: las empresas que no exportan tienen un riesgo doble de extinción que las que lo hacen, y ese riesgo añadido parece ser independiente de la edad de la empresa: cualquiera que sea esta edad, la internacionalización multiplica por dos la probabilidad de sobrevivir.

REFERENCIAS

Bednarski, T. (1989). On sensitivity of Cox's estimator. Statistics and Decisions. 7, 215–228.

Bednarski, T. (1993). Robust estimation in Cox's regression model. Scandinavian Journal of Statistics. Vol. 20, 213–225.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. Biometrika 69, 553-566.

Link, C. L. (1984). Confidence intervals for the survival function using Cox's proportional hazards model with covariates. Biometrics 40, 601-610.

Therneau T and Grambsch P (2000), Modeling Survival Data: Extending the Cox Model, Springer-Verlag.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

METODOLOGÍAS PARA LA AUTOEVALUACIÓN, GESTIÓN Y ANÁLISIS DEL PERFIL DE RIESGO BASADO EN ORSA

Segura Gisbert Jorge

[1] Universidad de Córdoba

Muñoz Murgui Francisco

[2] Universidat de València

RESUMEN

La aprobación de la Directiva Europea 2009/138 de Solvencia II supone un cambio sustancial en la revisión de las vigentes normas de supervisión de las compañías aseguradoras adoptando un enfoque basado en el riesgo. Bajo este nuevo paradigma denominado Solvencia II, se repasan de forma minuciosa las normas de valoración de las entidades reflejando una mejor medición y gestión de todos los riesgos. El eje básico de Solvencia II es el establecimiento del binomio riesgo-requerimientos de capital, incluyendo una adecuada evaluación del mismo como garante de protección al asegurado.

Como trabajo previo a la transposición definitiva de la Directiva, EIOPA emite unas Directrices de implementación cuyo ámbito de aplicación se centra en el marco de gestión y la evaluación prospectiva de los riesgos. En este sentido, dicha evaluación representa la visión propia de la empresa de su negocio asegurador, su perfil de riesgo, el cálculo y determinación de los requerimientos de capital así como de otros medios requeridos para afrontar tales riesgos. La entidad debe decidir por sí misma cómo realizar tal evaluación a la vista de la naturaleza, la escala y la complejidad de los riesgos inherentes a su actividad. Los métodos empleados pueden incluir desde sencillas pruebas de resistencia hasta modelos económicos de capital más o menos sofisticados.

La evaluación de la importancia de la desviación del perfil de riesgo con respecto a las hipótesis subyacentes al cálculo del SCR garantiza que la entidad aseguradora comprende dichas hipótesis y estudie si son adecuadas. De este modo, la empresa deberá comparar las hipótesis utilizadas con su propia comprensión del perfil de riesgo. Este proceso debe evitar que la empresa simplemente confíe en que los requisitos reglamentarios de capital son adecuados para su negocio. En este sentido, no se deberá limitar a un simple informe sino que ha de implementarse de tal manera que el resultado de la evaluación constituya un elemento fundamental en el proceso de adopción de decisiones de la entidad.

En este trabajo, nos centramos en diversas metodologías para la autoevaluación y gestión del riesgo basado en ORSA para el SCR de Primas y Reservas.

Palabras y frases clave: Solvencia II, ORSA, Gestión basada en el riesgo, SCR, EIOPA

Referencias

Alonso González, P y Albarrán Lozano, I (2008). Análisis del riesgo en seguros en el marco de Solvencia II: Técnicas estadísticas avanzadas Montercarlo y Bootstrapping. Fundación Mapfre.

Blanchard, R. The Lates ton ORSA: Many jurisdictions throughout the world are trying to define ORSA and the coming requirement for all insurers to have an ORSA. The Actuarial Review. The newsletter of The Casualty Actuarial Society Volume 39, number 3, August 2012.

Directiva 2009/138/CE del Parlamento Europeo y del Consejo, de 25 de Noviembre de 2009, sobre el seguro de vida, el acceso a la actividad de seguro y de reaseguro y su ejercicio (Solvencia II).

Durán Santomil, P y Otero González, L (2010). El análisis financiero dinámico como herramienta para el desarrollo de modelos internos en el marco de Solvencia II. Fundación Mapfre.

Gisler, A (2009). The Insurance Risk in the SST and in Solvency II: Modeling and parameter estimation. ASTIN Colloquium. Helsinki.

Pozuelo de Gracia, E (2007). El Fair Value de las Provisiones Técnicas de los seguros de vida. Fundación Mapfre.

Savelli, N y Clemente, G.P (2009). Modeling aggregate non-life underwriting: Standard formula vs Internal Model. Giornale del L´ Institute degli Attuari, 72. 295-333.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

REDUCCIÓN DE LA DEMORA DIAGNÓSTICA TRAS LA INTRODUCCIÓN DE LA VÍA RÁPIDA DE ACCESO A LA COLONOSCOPIA

Beatriz López-Calviño[1], Salvador Pita-Fernández[1] , Paloma González-Santamaría[2] , Rocío Seijo-Bestilleiro[1], Sonia Pértega-Díaz[1] , Teresa Seoane-Pillado[1]

[1] Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario A Coruña. (CHUAC). [2] Centro de Saúde Meicende. A Coruña

RESUMO

Análisis de datos longitudinales de la demora diagnóstica (interalo de tiempo desde 1º síntomas-diagnóstico) aplicando la metodología de las series de tiempo interrumpido con regresión segmentada para estudiar si la introducción de la vía rápida de acceso a la colonoscopia en la Atención Primaria (A.P), modifica la demora diagnóstica (intervalo 1º síntomas – diagnóstico) de los pacientes diagnosticados de cáncer colorrectal.

Palabras e frases chave: colorectal neoplasms, delay, longitudinal analysis, segmented regression

1. INTRODUCIÓN Y OBJETIVOS

El cáncer colorrectal se diagnostica habitualmente por manifestaciones clínicas, como resultado de un programa de cribaje o como hallazgo casual. Entre el inicio de la enfermedad y el diagnóstico o el tratamiento de la misma transcurre un intervalo de tiempo variable que se conoce como demora. La demora diagnóstica puede verse afectada por las características de la enfermedad, del paciente y del sistema sanitario. Desde Junio del 2010 se ha iniciado un programa que permite tanto al médico de Atención Primaria como de Atención Especializada solicitar la prueba de colonoscopia en aquellos pacientes que cumplan criterios de inclusión, para acortar el tiempo hasta el diagnóstico.
Objetivo: Estudiar el efecto de la introducción de la vía rápida de acceso a la colonoscopia ante la sospecha clínica de cáncer colorrectal sobre la dermora diagnóstica.

2. MATERIAL Y MÉTODOS

Diseño: Estudio observacional con seguimiento prospectivo en el Complexo Hospitalario Universitario A Coruña (CHUAC) desde Enero 2007 hasta Diciembre 2012 (En Junio de 2010 se introduce la vía rápida de acceso a la colonoscopia en Atención Primaria).
Criterios inclusión: Casos incidentes de cáncer colorrectal con diagnóstico histológico(CIE-9ª:153 154), con consentimiento informado y aprobación del comité de ética.
Criterios de exclusión: casos prevalentes/recurrentes, múltiples, atendidos en hospitales privados.
Tamaño muestral: n=1891 que permiten estudiar las características clínicas y la demora diagnóstica con una precisión de ±2,26% y un 95% de seguridad.
Mediciones: Variables demográficas, demora terapéutica (intervalo de tiempo síntomas-tratamiento), demora diagnóstica (intervalo de tiempo síntomas-diagnóstico), demora del paciente(intervalo de tiempo síntomas-1ºcontacto médico Atención Primaria(AP) ó Atención Especializada(AE)), demora del sistema de salud (intervalo de tiempo médico–diagnóstico).
Metodología para el cálculo de las demoras:

1. Cálculo fecha 1ª consulta con el médico por síntomas de ccr:
    1. Se consideró como 1ª fecha la del médico de primaria
    2. En los casos que la diferencia (fecha médico AE-fecha médico AP)<0, es decir, los pacientes acudieron primero a especializada antes que a primaria, se consideró como primera fecha la del médico de AE. Dicha situación ocurrió en 67 casos, de ellos el 55,2% fue a urgencias hospitalarias.

3. En los casos donde no figure la fecha del médico de AP, se considera la fecha del 1º contacto con el médico de AE.

2. Se consideró como 1ª fecha de inicio de síntomas la que figura en la historia de AP
    1. En los pacientes acudieron primero a especializada antes que a primaria, se consideró como primera fecha la del médico de AE (n=67 casos).
    2. En los casos donde no figure la fecha de inicio de síntomas de la historia de AP, se considera la fecha que figuraba en la historia de AE.
    3. En los casos que no figure fecha de inicio de síntomas ni en la historia clínica de AP ni de AE, se completará con la fecha que figura en la entrevista.
3. Demora del paciente: se calcula como la diferencia (fecha 1ª consulta con el médico por síntomas de CCR-fecha de inicio de síntomas). No se considerarán demoras PI de más de 2 años (≥730 días).
4. Demora del sistema de salud: Se calcula como la diferencia (fecha 1º informe de AP-fecha 1ª consulta con el médico por síntomas de CCR).No se considerarán demoras HSI de más de 2 años.
5. Demora diagnóstica: Se calcula como la suma de la demora del paciente y la de la administración. En los casos que no figure la fecha del médico, se calcula como (Fecha 1º informe de Anatomía Patológica – Fecha inicio síntomas).
6. Demora terapéutica (desde el diagnóstico hasta el inicio del tratamiento)
    1. Cálculo de la demora quirúrgica: diferencia (fecha cirugía-fecha 1º informe AP)

En 258 casos se detectó que la diferencia (fecha cirugía-fecha de 1º informe AP) < 0, es decir, fueron intervenidos directamente y el informe AP se obtuvo a posteriori. En dichos casos, se considera que esta demora fue de 0 días, al ser tratados quirúrgicamente en el mismo momento que se diagnosticaban.

    2. Calculo de la demora hasta inicio quimioterapia: diferencia (fecha inicio Quimioterapia-fecha 1º informe Anatomía Patológica)
    3. Calculo de la demora hasta inicio radioterapia: diferencia (fecha inicio Radioterapia-fecha 1º informe AP)
    4. Si la diferencia desde cirugía hasta inicio quimioterapia<0, es decir, iniciaron la quimioterapia antes de operarse, se considera como demora terapéutica del diagnóstico hasta el inicio de la quimioterapia.
    5. Si la diferencia desde cirugía hasta inicio radioterapia<0, es decir, iniciaron la radioterapia antes de operarse, se considera como demora terapéutica del diagnóstico hasta el inicio de la radioterapia.
7. Demora desde inicio de síntomas hasta el tratamiento: Se calcula como la suma de la demora diagnóstica + demora terapéutica.

Análisis estadístico: Descriptivo. Comparación de medias de las demoras según año de diagnóstico.
Para evaluar el impacto antes y después de la introducción de la vía rápida de acceso a la colonoscopia de Atención Primaria en Junio del 2010, sobre las diferentes demoras durante el período de estudio, se realizará un análisis de series de tiempo interrrumpido con regresión segmentada que permite estimar los efectos de la intervención.

Tras la observación de una medida en intervalos regulares en el tiempo pre y post intervencion (introducción de la vía rápida), una observación por momento, se utiliza la regresión segmentada para estimar la evolución de la variable dependiente antes de la intervención, para estimar la tendencia después de la intervención, para comprobar los cambios en la variable dependiente (observaciones) antes y después de la intervención y para comprobar los cambios en la pendiente de la tendencia de la intervención pre y post

Se especifica el siguiente modelo de regresión lineal para estimar la media de tiempo y la tendencia en la demora antes y después de la introducción de la vía rápida.

$$Y_t = \beta_0 + \beta_1 * \text{tiempo}_t + \beta_2 * \text{intervención}_t + \beta_3 * \text{tiempo después de la intervención}_t + e_t,$$

Siendo,

Y_t, número de observaciones/mes

tiempo_t, tiempo (meses) desde el inicio de periodo de observación

intervención_t , variable indicadora de antes y después de la intervención, codificándola como: "0" antes de la intervención y como "1" después de la intervención

$\text{tiempo después de la intervención}_t$, número de meses tras la intervención, codificándola como: "0" antes de la intervención y el valor de la demora después de la intervención

En este modelo, β_0, estima el nivel basal de los resultados de interés a tiempo cero

β_1, estima el cambio en la sensibilidad antes de la intervención (i.e la tendencia basal)

β_2, estima el cambio en el nivel de sensibilidad inmediatamente después de la intervención (i.e, desde el final del segmento anterior)

$\beta_1+\beta_3$, es la pendiente tras la intervención

e_t (error), representa la variabilidad aleatoria no explicada por el modelo

Aspectos ético-legales: Aprobación CEIC de Galicia (2009/110) y se garantiza el R.D 15/1999

3. RESULTADOS

Edad media al diagnóstico fue 69,8±11,3años, siendo hombres(60,3%). 63,1% fueron tumor de colon. El 92,2% fue adenocarcinoma, siendo el estadio más frecuente T3(58,9%), N0(49,4%) y M0(67,9%).

La demora desde el inicio de síntomas hasta el inicio del tratamiento fue 172,5±149(Mediana=125días). Siendo la demora síntomas-diagnóstico 144,0±146(Mediana=90días); la de síntomas-1ºcontacto médico AP/AE 71,0±112(Mediana=30días) y la demora médico-diagnóstico de 89,5±118(Mediana=42 días).

Se observa una reducción significativa de todos los intervalos de tiempo del proceso diagnóstico, descendiendo la demora diagnóstica un 3,38% por año. (Figura 1)

Se realizan modelos de regresión segmentada para determinar el efecto de la vía rápida sobre las demoras. Objetivándose como la vía rápida tiene un efecto significativo (p=0,014) en la reducción de la demora síntomas-diagnóstico (Tabla 1) . Dicho reducción significativa se observa fundamentalmente en el intervalo 1º contacto con el médico-diagnóstico. (Tabla 2)

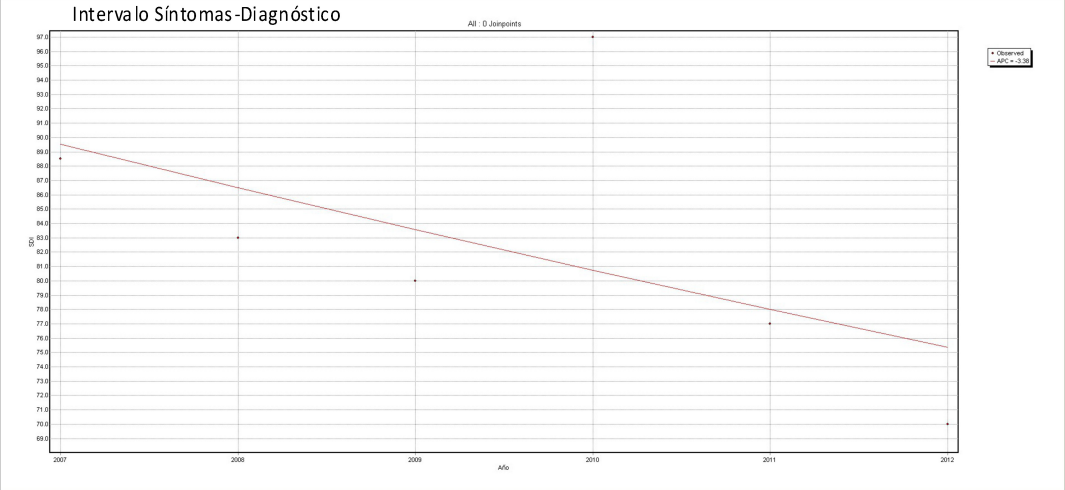


Figura 1: Evolución de la demora diagnóstica durante el periodo de estudio.

Tabla 1: Regresión de series interrumpidas con regresión segmentada para determinar si la introducción de la vía rápida de acceso a la colonoscopia modifica la demora síntomas-diagnóstico.

Variables	B	E.T.	t	p
Tiempo	-0,003	0,012	-0,252	0,801
Introducción vía rápida (antes vs. después)	30,280	12,291	2,464	0,014
Tiempo desde la introduccion de la vía rápida	-0,079	0,024	-3,223	0,001

Constante	127,441	7,695	16,561	<0,001

Tabla 2: Regresión de series interrumpidas con regresión segmentada para determinar si la introducción de la vía rápida de acceso a la colonoscopia modifica el intervalo de tiempo 1º médico-diagnóstico.

Variables	B	E.T.	t	p
Tiempo	0,000	0,007	0,017	0,986
Introducción ertapenem (antes vs. después)	2,821	7,256	0,389	0,697
Tiempo desde la introduccion de la vía rápida	-0,030	0,014	0,019	0,038
Constante	65,869	4,562	14,439	<0,001

4 CONCLUSIÓNS

La introducción de la vía rápida de acceso a la colonoscopia ante la sospecha clínica de cáncer colorrectal muestra una reducción significativa de la demora diagnóstica.

REFERENCIAS

Esteva et al.: Factors related with symptom duration until diagnosis and treatment of symptomatic colorectal cancer. BMC Cancer 2013 13:87.

Wagner AK, Soumerai SB, Zhang F and Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. Journal of Clinical Pharmacy and Therapeutics 2002; 27:209-309.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

DEMORA DIAGNÓSTICA Y PRONÓSTICO EN PACIENTES CON CÁNCER COLORRECTAL

Pértega Díaz S[1], Pita Fernández S[1], González Saez L[2], López Calviño B[1], Seoane Pillado T[1], Rodríguez Camacho E[3], González Santamaría P[4]

[1] Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario A Coruña (CHUAC)-Instituto de Investigación Biomédica A Coruña (INIBIC)
[2] Servicio de Cirugía. Complexo Hospitalario Universitario A Coruña (CHUAC)
[3] Servicio de Programas Poboacionais de Cribado. Dirección Xeral de Innovación e Xestión da Saúde Pública. Consellería de Sanidade, Xunta de Galicia.
[4] Centro de Salud de Meicende, Xerencia de Xestión Integrada A Coruña

RESUMEN

Se realiza un estudio de cohortes retrospectivo en n=942 casos incidentes de cáncer colorrectal para determinar la asociación entre la demora diagnóstica y la supervivencia. La mediana de la demora fue 3,4 meses (3,7 recto y 3,2 colon), variando según los síntomas presentados. Los resultados muestran como el intervalo primeros síntomas-diagnóstico no se asocia con la supervivencia en pacientes con cáncer de colon. En pacientes con cáncer de recto, la supervivencia es más baja en pacientes con demoras más cortas

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

1. INTRODUCIÓN

Existe controversia en relación con la posible asociación de la demora diagnóstica y la supervivencia de pacientes con cáncer colorrectal. Otros autores han reportado una asociación no lineal entre la demora y el riesgo de muerte.

El objetivo de este estudio ha sido determinar la asociación entre el intervalo primeros síntomas-diagnóstico y la supervivencia de pacientes con cáncer colorrectal.

2. PACIENTES Y MÉTODOS

Se realizó un estudio de cohortes retrospectivo de n=942 casos incidentes de cáncer colorrectal.El intervalo primeros síntomas-diagnóstico se calculó como el tiempo transcurrido entre los primeros síntomas atribuíbles al cáncer (según constaba en la historia clínica) y la confirmación anatomopatológica del diagnóstico.

Se analizaron los tumores de colon y recto de forma separada. Se empleó un modelo de regresión de Cox para estimar el valor de hazard ratio asociado a mortalidad a los 5 años tras el diagnóstico, en función de la demora, tras ajustar por edad y género. El intervalo primeros síntomas-diagnóstico se modelizó de dos formas: como una variable categórica en función de los cuartiles de demora, o como una variable cuantitativa mediante splines penalizados.

3. RESULTADOS

La mediana de la demora diagnóstica fue de 3,4 meses, significativamente mayor en tumores localizados en el recto que en tumores localizados en el colon (3,7 vs. 3,2 meses; $p<0,001$). La demora no se asoció con el estadío al diagnóstico (Estadio I=3,6 meses, Estadio II-III=3,4 meses, Estadio IV=3,2 meses; p=0,728).

Tras ajustar por edad y género, en tumores de recto, los pacientes con demora en el primer cuartil registraron una peor supervivencia que el resto (p=0,003), mientras que en pacientes con tumores en el colon no se observó una asociación entre la demora y la supervivencia (p=0,282).

El análisis basado en splines objetivó, en tumores de recto, cómo el riesgo de mortalidad a 5 años se incrementa progresivamente a media que disminuye la demora por debajo de la mediana (3,7 meses), y disminuye a medida que la demora aumenta hasta aproximadamente 8 meses. En tumores de colon, no se objetivó una asociación significativa entre la demora y la supervivencia.

4. CONCLUSIONES

El tiempo transcurrido entre los síntomas de la enfermedad y la confirmación del diagnóstico no se asocia con la supervivencia en pacientes con cáncer de colon. En pacientes con cáncer de recto, la supervivencia es más baja en pacientes con demoras más cortas.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

DESARROLLO DE UN SCORE DE PREDICCIÓN DE MORTALIDAD EN PACIENTES VIH CON NEUMONÍA POR *PNEUMOCYSTIS JIROVECII*

Rodríguez Osorio I[1], Mena A[1], Pértega S[2], Meijide H[1], Castro-Iglesias A[1], Pedreira J[1], Poveda E[1], López Calviño B[2], Seoane Pillado MT[2]

[1] Grupo de Virología Clínica. Instituto de Investigación Biomédica A Coruña (INIBIC)
[2] Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario A Coruña (CHUAC) - Instituto de Investigación Biomédica A Coruña (INIBIC)

RESUMEN

Se realiza un estudio de cohortes retrospectivo para desarrollar un score predictivo de mortalidad por Neumonía por *Pneumocystis jirovecii* en pacientes VIH. Se desarrolla un nomograma, una ecuación de predicción de mortalidad y un score de predicción basado en puntos mediante el uso de regresión logística múltiple y splines cúbicos retringidos. Se asociaron a mayor mortalidad la edad, el uso reciente de drogas, el gradiente alveolo-arterial de oxígeno y los valores de hemoglogina, creatinina y albúmina al ingreso.

Palabras clave: Virus de la inmunodeficiencia humana; Neumonía por *Pneumocystis jirovecii*; Mortalidad; Spline; Nomograma.

1. INTRODUCIÓN

La neumonía por *Pneumocystis jirovecii* (NPJ) continúa siendo la enfermedad definitoria de SIDA más frecuente. Aunque su incidencia ha disminuido, sigue siendo un problema clínico importante en pacientes VIH.

El objetivo de este estudio fue determinar la mortalidad por NPJ en pacientes VIH, los factores asociados y desarrollar un score de predicción de mortalidad en estos pacientes

2. MÉTODOS

Estudio de cohortes retrospectivo de todos los pacientes VIH con un primer episodio de NPJ en el Complexo Hospitalario Universitario de A Coruña en 1993-2013 (n=301).

Se recogieron variables sociodemográficas, uso de drogas, status VIH, tratamiento y variables durante la hospitalización. Se analizó la mortalidad precoz como la registrada en los 50 días siguientes al ingreso hospitalario.

Se realizó un análisis de regresión logística múltiple. Las posibles asociaciones no lineales se evaluaron mediante el uso de splines cúbicos restringidos. Se desarrolló un score de predicción basado en puntos utilizando la metodología de Sullivan et al. (2004).

Se analizó la capacidad predictiva mediante análisis de discriminación y calibración. El área bajo la curva ROC se estimó mediante técnicas bootstrap (1000 réplicas).

3. RESULTADOS

La media de edad fue 36.9±9.5 años, 76,1% varones. Ingresaron en UCI el 14,6%, con una estancia hospitalaria de 25.2±32.0 días. El porcentaje de reingresos fue del 9,3%. La mortalidad fue del 13,6% (n=41 pacientes).

En el análisis multivariante se asociaron a mayor mortalidad la edad (p=0,009), el uso reciente de drogas (p=0,001), el gradiente alveolo-arterial de oxígeno (p=0,001) y los valores de hemoglogina (p=0,020), creatinina (p=0,013) y albúmina (p<0,001) al ingreso. El análisis mediante splines reveló un

efecto no lineal de los valores de creatinina sobre la mortalidad, aumentando la misma tanto para valores bajos como elevados en este parámetro.

Se construyeron tres herramientas predictivas de mortalidad por NPJ:

a) Nomograma
b) Ecuaciones de predicción de mortalidad a partir de los coeficientes del modelo de regresión logística. La probabilidad de fallecer se estima como P=1/(1+exp(-LP)), donde

LP=-0,402+0,060xEdad+1,606xDrogas+0,030xGradA-a-0,287xHemoglobina+1,28xCreatinina-1,438xAlbúmina, si Creatinina≥0,9

LP=3,745+0,060xEdad+1,606xDrogas+0,030xGradA-a-0,287xHemoglobina-3,480xCreatinina-1,438xAlbúmina, si Creatinina<0,9

c) Score basado en puntos.

Todos los modelos mostraron buena capacidad discriminante (área bajo la curva ROC corregida por bootstrap=0,89) y similar calibración.

4. CONCLUSIONES

Se proponen tres herramientas de predicción de mortalidad por NPJ en pacientes VIH, basadas en 6 variables y fácilmente aplicables en la práctica, con adecuados valores de discriminación y calibración.

REFERENCIAS

Sullivan LM, Massaro JM, D'Agostino RB. (2004) Presentation of multivariate data for clinical use: The Framingham Study risk score functions. Stat Med, 23(10):1631-60

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

NOMOGRAMA Y VALIDEZ DE UN MODELO PARA PREDECIR DESNUTRICIÓN EN PACIENTES EN LISTA DE TRASPLANTE HEPÁTICO

López Calviño Beatriz[1] , García Rodríguez María Teresa[2], Pita Fernández Salvador[1], Pértega Díaz Sonia[1], Piñón Villar María del Carmen[2] , Seoane Pillado Teresa[1]

[1] Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario A Coruña (CHUAC)-Instituto de Investigación Biomédica A Coruña (INIBIC)
[2] Servicio de Digestivo. Complexo Hospitalario Universitario A Coruña (CHUAC)

RESUMEN

Se considera relevante determinar la prevalencia de desnutrición en pacientes con hepatopatía crónica por ser un factor de riesgo para su morbimortalidad. Debido a que no existe un gold stándar para evaluar el estado nutricional, se estudia la concordancia de los métodos de valoración. Finalmente se construye un nomograma para predicción del riesgo de desnutrición, tras estudiar los factores predictivos. Objetivándose que valores bajos en la fuerza de agarre (dinamometría) y la presencia de asicitis aumentan el riesgo de desnutrición.

Palabras clave: liver transplant, malnutrition, diagnostic validity, nomogram

1. INTRODUCIÓN

La desnutrición es un factor pronóstico de la evolución de la enfermedad y debido a la no existencia de un gold stándard se evalúa el estado nutricional mediante los diferentes métodos existentes. De este modo, se pretente identificar los factores asociados y construir un score de predicción.

El objetivo de este estudio ha sido determinar un modelo predictivo y construir un nomograma para identificar el riesgo de desnutricion en pacientes en lista de trasplante hepático.

2. PACIENTES Y MÉTODOS

Se realizó un estudio observacional de seguimiento prospectivo de n=110 pacientes (precisión=±9,4%;seguridad=95%) de 18 y más años en lista de trasplante hepático que den su consentimiento informado a participar.

Se estudiaron variables demográficas (edad y sexo), descompensaciones debidas a la enfermedad, disfunción hepática según la clasificación de la escala Child-Pugh, valoración del estado nutricional según diferentes métodos (Sistema de Control Nutricional(CONUT), Test de Valoración Global Subjetiva(TVGS), criterios SENPE , Índice de Riesgo Nutricional (IRN) y Índice Nutricional Pronóstico de Onodera (PNI-O)) y la fuerza de agarre a través de la dinamometría.

Se realizó un análisis de regresión logística múltiple, estudiando la validez diagnóstica del modelo predictivo mediante el área bajo la curva ROC.

Se construye un nomograma predictivo que permite calcular el riesgo de desnutrición de una manera sencilla para ser aplicado por personal sanitario.

El nomograma asigna a cada variable un rango de puntuación de 0 a 100 a través de una representación gráfica. Dicha puntuación se obtiene a través de los coeficientes de regresión mediante la predicción lineal (LP) de la siguiente manera:

$$\text{Puntos}=100*\frac{\left(\text{LP}_i-\min\left(\text{LP}\right)\right)}{\left(\max\left(\text{LP}\right)-\min\left(\text{LP}\right)\right)}$$

Para obtener la puntuación total y la probabilidad del evento se sigue la siguiente metodología:

1. Puntos por unidad de predictor lineal:

$$\frac{100}{\max\left(LP\right)-\min\left(LP\right)}$$

2. Unidades de predictor lineal por puntos:

$$\frac{1}{\text{Puntos por unidad de predictor lineal}}$$

3. LP para puntuación total de 0:

$$\beta_0+\text{X}\beta$$

con cada variable en el valor de referencia para una puntuación total de 0.

4. LP para puntuación total >0:

$$\ln\left(\frac{\text{Riesgo de Y=1}}{\text{1-Riesgo de Y=1}}\right)$$

5.Puntuación total:

$$\text{Puntuación total}=\left(\text{Puntos por unidad de predictor lineal}\right)*\left(LP_i-LP\text{ para puntuaciones totales de }0\right)$$

La interpretación del nomograma se basa en sumar los puntos de cada variable, obteniendo la puntuación total. A continuación, se traza la línea vertical desde la puntuación total hasta los ejes horizontales correspondiente con el pedictor lineal y con el riesgo del evento, obteniendo la probabilidad final.

El estudio está aprobado por el Comité de Ética de Investigación Clínica de Galicia (CEIC 2010/081) y ha sido financiado por FIS ISC.III (PI11/01255).

3. RESULTADOS

La edad media fue 56,9±8,2 años siendo un 72,7%hombres. El 48,6% tiene una disfunción hepática grado B (clasificación Child-Pugh). Las descompensaciones más frecuentes son asicitis (79,1%), encefalopatía (69,9%) e hipertensión portal (68,2%). El porcentaje de desnutrición varía según el método de valoración: 90,9% con CONUT; 89,1% con IRN; 87,3% con PNI-O; 64,6% con SENPE; 50,9% con TVGS. No existiendo una buena concordancia entre los métodos, Kappa Global K=0,364; p<0,001. Estableciéndose como criterio estar desnutrido por al menos 4 o más métodos, observándose una prevalencia de desnutrición del 70,0%.

Se estudia la fuerza de las manos a través de la dinamometría obteniendo los siguientes valores: 20,0±9,73kg para la mano derecha, 18,0±9,62kg para la mano izquierda y 19,5±9,44kg para ambas manos (obtenido como la media de las manos).

La dinamometría para ambas manos es menor en las mujeres que en los hombres (10,23±5,66 vs.22,29±8,42;p<0,001), en los >60 años que en los ≤60 años (17,12±9,80 vs.20,76±8,81;p=0,043) y en los pacientes desnutridos por 4 o más métodos que en los nutridos (17,36±9,30 vs. 22,8±8,75; p=0,005).
Se realiza un modelo de regresión logística para determinar qué variables se asocian a estar desnutrido por 4 o más criterios ajustando por edad, sexo, dinamometría para ambas manos y tener ascitis. Objetivándose que a menor valor en la dinamometría (OR=0,922,p=0,010) y la presencia de ascitis (OR=3,006,p=0,03) incrementan el riesgo de desnutrición (Tabla 1). El modelo predictivo tiene una buena capacidad predictiva para identificar la desnutrición (área bajo la curva ROC AUC=0,715). (Figura 1).

Se construye un nomograma para la predicción del riesgo de desnutrición a partir de las puntuaciones obtenidas con los coeficientes de regresión del modelo predictivo (Figura 2).

Tabla 1. Regresión logística para determinar factores de riesgo asociados a la desnutrición

Variables	**B**	**E.T**	**p**	**O.R**	**I.C. 95% (OR)**	
					Inf	**Sup**
Edad	0,893	0,671	0,183	2,441	0,655	9,094
Sexo (hombre)	-0,033	0,029	0,263	0,968	0,914	1,025
Dinamometría (ambas manos)	-0,081	0,031	**0,010**	0,922	0,867	0,980
Ascitis	1,101	0,531	**0,038**	3,006	1,062	8,509
Constante	2,846	1,863	0,126	17,227		

R^2=0,175

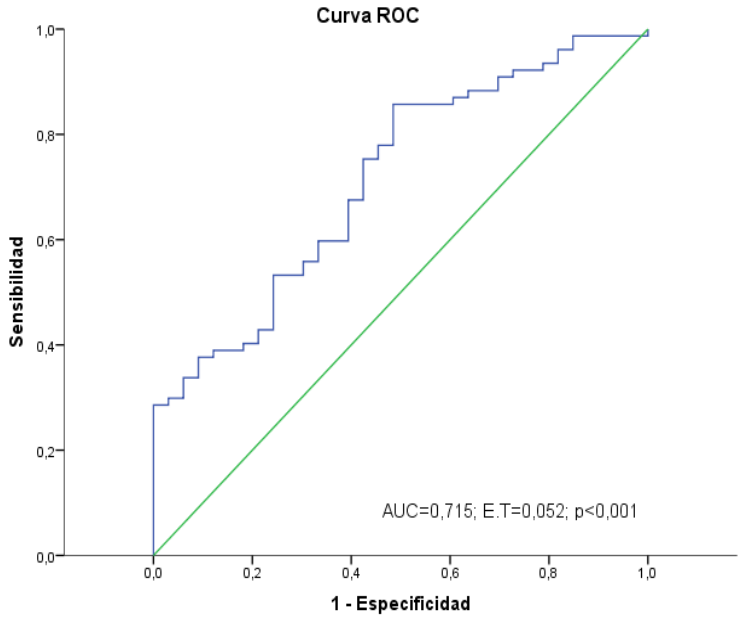


Figura 1. Curva ROC de la probabilidad del modelo de predicción para identificar la desnutrición

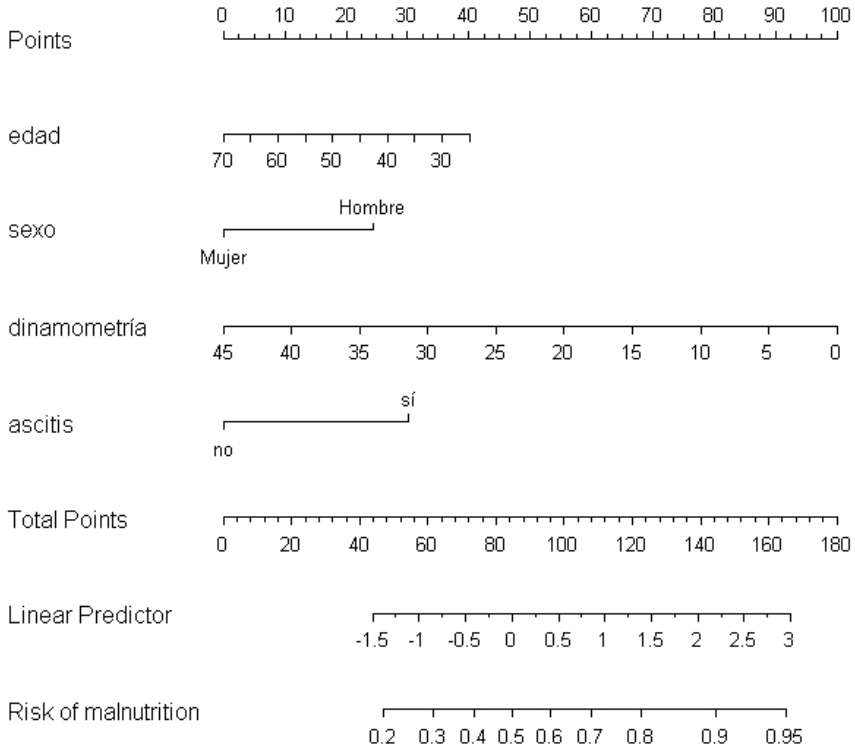


Figura 2. Nomograma para predicción del riesgo de desnutrición

4.CONCLUSIONES

La dinamometría y la presencia de ascitis son factores predictivos para determinar el riesgo de desnutrición.

REFERENCIAS

Frank E Harrell Jr (2013). rms: Regression Modeling Strategies. R package version 3.6-3. http://CRAN.R-project.org/package=rms

Lasonos A, Schrag D et al: How To Build and Interpret a Nomogram for Cancer Prognosis. Journal of Clinical Oncology, volume 26 , 2008.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

MODELOS DE RIESGOS COMPETITIVOS PARA DETERMINAR FACTORES PRONÓSTICOS ASOCIADOS A LA MORTALIDAD POR INFECCIÓN EN PACIENTES TRASPLANTADOS RENALES

López Calviño Beatriz[1] , Pérez Fontán Miguel[2], Pita Fernández Salvador[1], Pértega Díaz Sonia[1], Seoane Pillado Teresa[1]

[1] Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario A Coruña (CHUAC)-Instituto de Investigación Biomédica A Coruña (INIBIC)
[2] Servicio de Nefrología. Complexo Hospitalario Universitario A Coruña (CHUAC)

RESUMEN

Los pacientes trasplantados renales a lo largo de su seguimiento tienen riesgo de peritonitis y de peritonitis por infección gramnegativos provocando mortalidad debido a la infección. Se estudia a través de modelos de riesgos competitivos si el uso de protectores gástricos (Antagonistas H e Inhibidores) incrementa la probabilidad de mortalidad debido a peritonitis infecciosa por gramnegativos. Objetivándose que aumenta la probabilidad de mortalidad infecciosa, el uso de protectores gástricos en un momento inicial (incrementando el riesgo con un uso prolongado durante los primeros 6 meses), recibir tratamiento inmunosupresor y la edad.

Palabras clave: peritonitis, gastroprotectors, competitive risk, survival

1. INTRODUCIÓN

Debido a que la prevalencia de peritonitis por gramnegativos en pacientes trasplantados hepáticos es elevada y a su vez esta ocasiona en muchos casos mortalidad debida a la infección. Se estudia si el consumo de gastroprotectores (Antagonistas H) en un momento inicial y a lo largo de los primeros 6 meses aumenta la probabilidad de fallecer por peritonitis infecciosa de gramnegativos.

2. PACIENTES Y MÉTODOS

Se realizó un estudio observacional de seguimiento prospectivo de n=691 pacientes (precisión=±3,8%;seguridad=95%) con insuficiencia renal que den su consentimiento informado a participar.

Se estudiaron variables demográficas (edad y sexo), parámetros analíticos, recibir tratamiento inmunosupresor y tomar gastroprotectores (antagonistas H).

Se calcula la curva de incidencia acumulada de Kalbefleisch y Prentice, para el evento de interés (mortalidad por peritonitis infecciosa por gramnegativos), siendo el evento competitivo la mortalidad por otra causa y teniendo en cuenta los datos censurados (vivos).

Se realizó un análisis de modelos de riesgos competitivos para determinar los factores pronósticos asociados a la mortalidad por peritonitis infecciosa de gramnegativos, asumiendo como evento competitivo morir por otra causa.

3. RESULTADOS

La incidencia acumulada para mortalida infecciosa es del 3,53% al año y del 13,63% a los 5 años; siendo la incidencia acumulada de mortalidad por otra causa del 8,80% al año y del 51,27% por otra causa. (Figura 1).

Se realizan modelos de riesgos competitivos para determinar qué variables se asocian con la mortalidad debido a la peritonitis infecciosa por gramnegativos considerando como evento competitivo la mortalidad por otra causa. Se objetiva que la edad, niveles bajos del parámetro kr y el tratamiento inmunosupresor

incrementan dicha probabilidad. Observándose com el uso de antagonistas H en el momento inicial (Tabla 1) y durante los primeros 6 meses es un factor de riesgo para fallecer por infección (tabla 2).

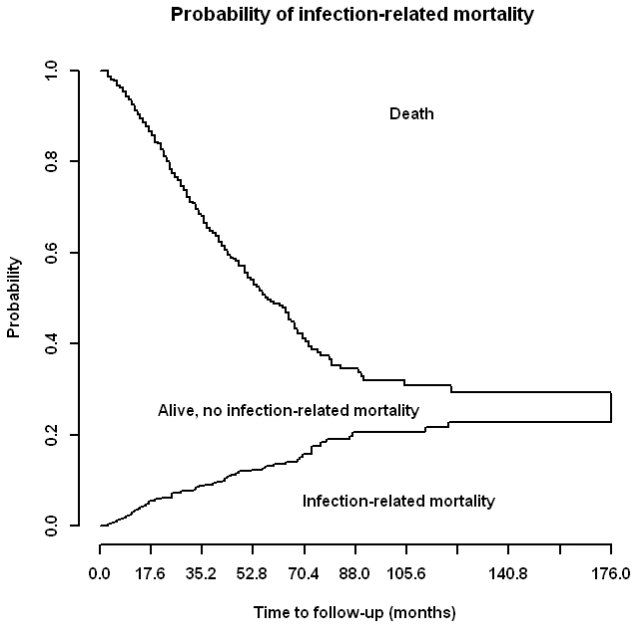


Figura 1. Incidencia acumulada de mortalidad infecciosa, ajustando por el evento competitvo mortalidad por otra causa, en posibles candidatos a la peritonitis infecciosa por gramnegativos

Tabla 1. Análisis de factores pronóstico asociados a la mortalidad infecciosa por gramnegativos ajustando por tomar antagonistas H en el momento inicial.

Variables	B	E.T	P	HR	I.C al 95% (H.R)
Edad	0.0397	0.0132	**0.0026**	**1.040**	(1.014; 1.070)
Hemoglobina	-0.0838	0.0832	0.3100	0.920	(0.781; 1.080)
Kr.med	-0.0654	0.0339	0.0540	0.937	(0.876; 1.000)
Portad	-0.4648	0.2682	0.0830	0.628	(0.371; 1.060)
Tratamiento inmunosupresor	1.0084	0.4322	**0.0200**	2.741	(1.175; 6.400)
K basal	-0.3058	0.2406	0.2000	0.737	(0.460; 1.180)
Antagonistas.H (basal)	0.7918	0.2608	**0.0024**	**2.207**	(1.324; 3.680)

Tabla 2. Análisis de factores pronóstico asociados a la mortalidad infecciosa por gramnegativos ajustando por tomar antagonistas H durante los primeros 6 meses.

Variables	B	E.T	P	HR	I.C al 95% (H.R)
Edad	0.0397	0.0132	**0.0026**	**1.040**	(1.014; 1.070)
Hemoglobina	-0.0838	0.0832	0.3100	0.920	(0.781; 1.080)
Kr.med	-0.0654	0.0339	0.0540	0.937	(0.876; 1.000)
Portad	-0.4648	0.2682	0.0830	0.628	(0.371; 1.060)
Tratamiento inmunosupresor	1.0084	0.4322	**0.0200**	2.741	(1.175; 6.400)
K basal	-0.3058	0.2406	0.2000	0.737	(0.460; 1.180)
Antagonistas.H (durante los primeros 6 meses)	0.7918	0.2608	**0.0024**	**2.207**	(1.324; 3.680)

4.CONCLUSIONES

A mayor edad, niveles bajos de Kr, tomar tratamiento inmunosupresor y el uso de gastroprotectores antagonistas H en el momento inicial; así como durante los primeros 6 meses incrementan la mortalidad por infección de gramnegativos.

REFERENCIAS

Gray RJ. A class of k-sample tests for comparing the cumulative incidence of a competing risk. Ann Stat. 1988; 16:1141-54.
Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999; 94:496-509.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

CVRCALC: A CARDIOVASCULAR RISK CALCULATOR

Seoane Pillado MT[1], Pita Fernández S[1], Hervada Vidal X[2], Pértega Díaz S[1], López Calviño B[1], Rodríguez Muíños MA[2]
[1]Unidad de Epidemiología Clínica y Bioestadística. Complejo Hospitalario Universitario A Coruña. Servizo Galego de Saúde.
[2]Dirección Xeral de Innovación e Xestión da Saúde Pública. Consellería de Sanidade. Xunta de Galicia.

ABSTRACT

Development of an application, CVRCALC, for the calculation of different cardiovascular risk scores (Framingham, Framingham-Wilson, REGICOR, DORICA and SCORE) by using the bulk load data. CVRCALC is a program developed in R (The Comprehensive R Archive Network)

Keywords: Cardiovascular risk equations. Risk factors. Software. Calculation.

1. BACKGROUND:

There is great interest in developing cardiovascular risk prediction models that will allow us to take action at the level of primary prevention. Software tools are available which allow us to automatically calculate this risk according to any established model. This type of program (commonly referred to as the "cardiovascular risk calculators") are usually based on risk prediction tables presenting estimates of the risk of cardiovascular disease.

But problems arise when we want to do an assessment of a set of individuals previously collected in a database. Current tools only allow us to perform these calculations manually by entering the requested information for each of the individual cases to assess.

2. METHODS:

CVRCALC is a program developed in R which allows us to estimate the cardiovascular risk of a database of individuals applying algorithms from different models. The project is hosted in R-Forge (http://r-forge.r-project.org/projects/cvrcalc/) and is available under the GNU/GPL license.

There are two ways to install the CVRCALC package in R: 1. -Download the file http://r-forge.r-project.org/R/?group_id=1731 and proceed to the installation from a ZIP archive. 2. -Run the following command from the R command console:

```
>install.packages("cvrcalc", repos="http://R-Forge.R-project.org")
```

3. RESULTS:

In R, load CVRCALC: **require (cvrcalc)** and call your graphical environment: cvrcalcgui().From the main window, we must supply the program, through an Excel file, with the variables: sex, age, cholesterol, hdl, tas, tad, smoking, diabetes and hypertrophy for each individual that have been previously collected. In addition, we must select the model that we want to apply: Framigham classic, Framingham-Wilson, Regicor, Dorica, High Risk Score or Low Risk Score. The result will appear as a new column in the Excel file supplied, with the estimation of the cardiovascular risk for each of the individuals on the database.

4. CONCLUSIONS:

CVRCALC is a cardiovascular risk calculator that allows to load a database, select the model you wish to apply and which allows us to calculate, in one step, the cardiovascular risk for each of the records in the database and store the results.

REFERENCIAS

Anderson KM, Wilson PWF, Odell PM, Kannel WB, An Updated coronary risk Profile. A Statement for healths profesionals. Circulation 1991; 83: 356-62. PubMed PMID: 1984895.

Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998 May 12;97(18):1837-47. PubMed PMID: 9603539.

Marrugat J, D'Agostino R, Sullivan L, Elosua R, Wilson P, Ordovas J, et al. An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas. J Epidemiol Community Health. 2003 Aug;57(8):634-8. PubMed PMID: 12883073; PubMed Central PMCID: PMC1732543.

Aranceta J, Pérez Rodrigo C, Foz Sala M, Mantilla T, Serra Majem L, Moreno B, et al; Grupo Colaborativo para el estudio DORICA fase 2. [Tables of coronary risk evaluation adapted to the Spanish population: the DORICA study]. Med Clin (Barc). 2004 Nov 20;123(18):686-91. Spanish. Erratum in: Med Clin (Barc). 2004 Dec 4;123(20):30. PubMed PMID: 15563815.

Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al.; SCORE project group. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003 Jun;24(11):987-1003. PubMed PMID: 12788299.

R Development Core Team. The R Manuals. [acceso 05.05.2015]. Disponible en: http://cran.r-project.org/manuals.html

R Core Team. 2012. Writing R extensions. [acceso 05.05.2015]. Disponible en: http://cran.rproject.org/doc/manuals/r-release/R-exts.pdf

Leisch, F. (2009). Creating R Packages: A Tutorial. [acceso 05.05.2015]. Disponible en: http://cran.r-project.org/doc/contrib/Leisch-CreatingPackages.pdf

Verzani J. (2014). gWidgets API for building toolkit-independent, interactive GUIs. [acceso 05.05.2015]. Disponible en:

http://cran.r-project.org/web/packages/gWidgets/index.html

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

SAÚDE MENTAL EM TEMPOS DE CRISE: ANÁLISE DOS CUIDADOS PRESTADOS NO AGRUPAMENTO DE CENTROS DE SAÚDE ARCO RIBEIRINHO, ÁREA METROPOLITANA DE LISBOA

Duarte Tavares[1], Helena Mouriño[2], Mário Durval[3] e João Carvalho das Neves[4]

[1] CMAF-CIO, Universidade de Lisboa
[2] Faculdade de Ciências, Universidade de Lisboa
[3] Unidade de Saúde Pública, ACES Arco Ribeirinho
[4] Instituto Superior de Economia e Gestão, Universidade de Lisboa

RESUMO

De acordo com os resultados mais recentes de um estudo epidemiológico nacional de Saúde Mental, em Portugal 22% dos habitantes apresentam, pelo menos, uma perturbação psiquiátrica. Esta percentagem é considerada a mais elevada prevalência de doenças psiquiátricas na Europa, em conjunto com a Irlanda do Norte (Caldas de Almeida e Xavier, 2013).

Desde o despoletar da crise financeira mundial nos Estados Unidos da América, em 2008, até aos dias de hoje e, em especial após a assinatura do memorando de entendimento entre a *Troika* e o Estado Português, em 2011, as condições de vida dos portugueses sofreram bastantes alterações, de onde se destaca o aumento da taxa de desemprego que afetou 16,3% da população ativa em 2013 (INE, 2014). A Península de Setúbal tem sido uma das regiões portuguesas mais afetadas pela crise, exibindo particular relevo o fenómeno do desemprego, salientando-se que no último trimestre de 2012 esta região registou a taxa de desemprego mais elevada em Portugal, atingindo os 21% da população activa (Pereira, 2013).

Durante os períodos de crise económica e financeira, a promoção de saúde mental e prevenção das perturbações mentais são uma prioridade (OMS, 2011). Há vários estudos que demonstram que os fatores sociais, como o desemprego, têm um impacto negativo na saúde mental e estão associados a aumentos e/ou alterações na prevalência de patologias de natureza psiquiátrica (Urbanos-Garrido e Lopez-Valcarcel, 2015; Economou *et al*., 2014; De Silva *et al*, 2005). No entanto, é consensual que os planos de de intervenção primária podem desempenhar um papel crucial na mitigação dos efeitos da crise, a nível das patologias psiquiátricas (Marqués-Calderón *et al*., 2014; Economou *et al*., 2013).

Para realizar um plano de intervenção primária adequado às necessidades da população, urge conhecer a realidade a nível de perturbações mentais. Para tal, estudou-se a população da Península de Setúbal residente na zona do Arco Ribeirinho (concelhos de Alcochete, Barreiro, Moita e Montijo), com pedido de internamento compulsivo entre janeiro de 2009 e dezembro de 2013, uma vez que este tipo de intervenção tem por objetivo tratar indivíduos com anomalia psiquiátrica grave (Portugal. Lei nº 36/98 de 24/07, Lei da Saúde Mental).

O presente estudo pretende verificar se a crise económica alterou a prevalência e as características dos internamentos compulsivos, nomeadamente: a duração do internamento, a terapêutica psicofarmacológica prescrita, existência de reincidências, patologias e diagnósticos. Para além disso, verificou-se se existem alterações significativas entre as zonas urbanas e rurais dos concelhos supra referidos, no que diz respeito às dinâmicas dos internamentos compulsivos.

Palavras chave: Saúde Mental. Internamentos Compulsivos. Crise. Análise de Tendência. Testes do Qui-Quadrado.

REFERÊNCIAS

Caldas de Almeida, J. M. e Xavier M. (2013) Estudo Epidemiológico Nacional de Saúde Mental – 1º Relatório. Faculdade de Ciências Médicas, Universidade Nova de Lisboa, Lisboa.

De Silva, M. J., McKenzie, K, Harpham, T. e Huttly, S. R. A. (2005) Social capital and mental illness: a systematic review. J Epidemiol Community Health, 59, 619-627.

Economou, M., Madianos M., Peppou, L. E., Patelakis, A. e Stefanis, C. N. (2013) Major depression in the Era of economic crisis: A replication of a cross-sectional study across Greece. Journal of Affective Disorders, 145, 308-314.

Economou, M., Madianos, M., Peppou, L. E., Souliotis, K., Patelakis, A. e Stefanis, C. (2014) Cognitive social capital and mental illness during economic crisis: A nationwide population-based study in Greece. Social Science & Medicine, 100, 141-47.

Instituto Nacional de Estatística (2014) Estatísticas do Emprego – 4.º Trimestre de 2013. Instituto Nacional de Estatística, Lisboa.

Márquez-Calderón, S., Villegas-Portero, R., Soler, V. G., Martínez-Pecino, F. (2013) Promocíón de la salud y prevención en tiempos de crisis: el rol del sector sanitario. Informe SESPAS 2014. Gac Sanit. 2014 Jun;28 Suppl 1:116-23.

Organização Mundial de Saúde (2011) Impact of economic crises on mental health. Geneva: WHO, Geneva.

Pereira, S., Campos, P. e Correia, L. (2013) Estimação em Pequenos Domínios do Desemprego ao Nível NUTS III. Instituto Nacional de Estatística, Lisboa.

Urbanos-Garrido, R. M. e Lopez-Valcarcel, B. G. (2015) The influence of the economic crisis on the association between unemployment and health: an empirical analysis for Spain. Eur J Health Econ, 16:2, 175-184.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

MODELIZACIÓN DE LA DEGRADACIÓN DE SMART-BIOMATERIALES PARA USO ODONTOLÓGICO

Yaroslava Robles-Bykbaev [1], Salvador Naya[2] y Javier Tarrío-Saavedra[2]

[1]Facultad de Informática. Departamento de Matemáticas. Campus de Elviña. Universidade da Coruña.

[2]Escuela Politécnica Superior. Departamento de Matemáticas. Grupo MODES. Ferrol. Universidade da Coruña.

RESUMEN

En este trabajo se propone el empleo de modelos estadísticos para el estudio de la fiabilidad de nuevos biomateriales, como el caso de los biopolímeros usados en diferentes aplicaciones médicas. Concretamente se analizarán modelos cinéticos que permiten el estudio del tiempo de vida de los andamios usados en odontología. Así, el principal objetivo es proponer la estimación de los procesos de degradación y el tiempo de vida de biopolímeros para aplicaciones dentales, como el caso de varias formulaciones de ácido poliláctico (PLA) propuestos para su futuro uso en andamios dentales.

Palabras clave: Degradation models; Nonlinear Mixed-Effects Models; Generalized Additive Models.

1. INTRODUCCIÓN

En los últimos años se está produciendo la irrupción de nuevos métodos estadísticos para el estudio de la fiabilidad de materiales. Por otra parte el campo de los biomateriales es también un campo de gran expansión con biopolímeros que permiten la regeneración ósea o muscular, y nuevos nanomateriales que posibilitan acceder a cualquier parte del cuerpo humano para la administración de medicamentes dirigidos. El estudio de la fiabilidad de estos biomateriales es un tema de gran interés donde la modelización estadística es un campo emergente.

Muchos de los biopolímeros usados para aplicaciones odontológicas, están basados en ácidos polilácticos (PLA) dadas sus propiedades de biodegradabilidad y biocompatibilidad (Anderson,1997).



Foto 1: Detalle de un andamio dental.

En el presente caso, los polímeros estudiados se obtuvieron para el desarrollo de bio-andamios. De hecho, después de la pérdida de tejido dental y periodontal, se suele buscar la regeneración ósea para lo que es preciso usar este tipo de bio-andamios con objeto de ayudar en el crecimiento del hueso, que se utiliza en algunos casos antes de la colocación de implantes de otros materiales (Hollister 2005).

El problema, desde el punto de vista estadístico, es la identificación de variables influyentes en la degradación, la modelización de la dependencia de la pérdida de masa con respecto a estas variables y la estimación de la vida útil de estos materiales en situaciones similares al cuerpo humano. Entendiendo que el tiempo de vida útil del material como aquel tiempo o duración después de que este material no puede cumplir con sus requisitos de diseño y por lo tanto no es fiable (Prime et al, 2009; Vyazovkin et al, 2009). El envejecimiento depende de diferentes variables como la masa y otras características mecánicas (Vyazovkin et al. 2009). Este envejecimiento puede ser producido por la acción de presión, el agua, el PH del cuerpo, la tensión mecánica, el calor, etc. Por tanto, para la estimación de modelos para la degradación y las estimaciones de la vida de los bio-polímeros utilizados serán cruciales analizar las relaciones dependiendo de estas variables.

En la literatura, hay algunos trabajos estadísticos dedicados a la degradación de andamios y a la modelización del tiempo de vida. Por ejemplo, en Chen et al (2011) se desarrolló un modelo estadístico teniendo en cuenta la hidrólisis y la masa para simular los procesos de degradación y erosión polimérica. En Hoque et al. (2012) también se modeló la pérdida de masa usando una expresión exponencial, suponiendo que la difusión de agua y la hidrólisis como las principales causas de los procesos de degradación.

El resto del resumen se presentará en la siguiente sección el problema de la modelización en función de una sola variable, para en otra sección posterior mostrar la diferencia con la inclusión de otras variables, como el pH o el tipo de polímero.

2. DISEÑO DE EXPERIMENTOS

El objetivo principal es proponer nuevos modelos para la estimación de la degradación de biomateriales usados en prótesis odontológicas, partiendo de diferentes factores como la temperatura, el pH o la humedad. En lo que se refiere al caso particular de la modelización aplicada a biomateriales degradables, existe un gran interés en el estudio de sus procesos de biodegradación, debido principalmente a que la incrustación de biomateriales degradables en el cuerpo humano da lugar a una serie de procesos de carácter biológico que pueden afectar a la salud del paciente, por lo que en la medida de lo posible los ensayos deben de ser llevados a cabo en laboratorio antes de su empleo en el ser humano, para lo que es necesario la realización de una serie de experimentos que reproduzcan las condiciones más parecidas posibles a las reales, en este caso concreto, se trataría de reproducir la temperatura, acidez, pH o humedad bucal para el posible uso odontológico.

Desde el punto de vista estadístico, como ya se comentó se trata de modelizar una variable, la degradación de los biomateriales (masa perdida en el experimento) usando diferentes factores y empleando diferentes técnicas como los modelos gam (Generalized Additive Models). Para ello se usará la librería de R, nlme (Wood, 2006) y la librería mgcv (Wood, 2013a, Wood 2013b).

Las muestras analizadas son 7 biopolímeros del tipo PLA, con una masa de 0.3g cada una. Las muestras se mezclaron en probetas con 1 ml de fosfato de solución salina (PBS). Una vez cerradas las muestras se mantienen en un baño a 37 °C, temperatura que imita la del cuerpo humano donde se sintetizan macromoléculas como el ADN y ARN similares a las de nuestro interés.

El diseño de esta investigación consistió en hacer 2 repeticiones por polímero PDLGA mediante el orden que responde a la aleatorización experimental. La variable respuesta era la masa en función del tiempo. Cada 3 días se analizó y se midieron en las muestras de cada polímero: la masa, el pH, el peso molecular y tipo de biopolímero.

3. MODELIZACIÓN DE LA DEGRADACIÓN EN FUNCION DEL TIEMPO

En un primer estudio, se realizó una modelización GAM con el objeto de estimar el tipo de relación entre la variable respuesta masa perdida y las variables independientes pH, tipo de polímero y tiempo. Se obtuvo que el efecto del pH sobre la pérdida de masa es de tipo asintótico, mientras que la relación entre la pérdida de masa y el tiempo es de tipo sigmoide.

Como puede observarse en la Tabla 1 el tipo de polímero usado es muy significativo siendo muy diferentes los valores estimados. La aplicación práctica de este estudio permitirá el empleo de estos diferentes polímeros en función de la aplicación concreta, ya que en algunos casos se persigue que el andamio se degrade rápidamente (caso de una regeneración del hueso mandíbula de un maxilar inferior para un implante dental), mientras que otros casos se requieren biopolímeros más perdurables.

Tipo de Polímero	Estimación	Error	Valor Z	p-valor
PolymerPDL02A	0.8976	0.3114	2.883	0.00394
PolymerPDLG5002	14.783	0.3169	4.665	3.08e-06
PolymerPDLG5004	13.217	0.3184	4.151	3.31e-05
PolymerPDLG5010	0.8167	0.3050	2.678	0.00741

PolymerPDLG7502	0.7643	0.3036	2.518	0.01181
PolymerPDLG7502A	25.022	0.3255	7.687	1.50e-14

Tabla 1: Estimación con un modelo *gam* para diferentes biopolímeros en función del tiempo.

Se analizó la importancia de la masa perdida solo en función del tiempo y del polímero mediante la aplicación de modelos de regresión no lineales de efectos mixtos, añadiendo un efecto aleatorio relacionado con el tipo de polímero (pues los polímeros estudiados son sólo una muestra de las posibles alternativas PLGA. Concretamente, se ha aplicado un modelo no lineal logístico para modelizar la pérdida de masa con respecto del tiempo.

$$Mass\ Loss_{ij} = \phi_{1i} + \frac{\phi_{2i} - \phi_{1i}}{\left(1 + \exp\left(\frac{Time_{ij} - \phi_{3i}}{\phi_{4i}}\right)\right)} + \epsilon_{ij}.$$

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{bmatrix} = \beta + b_i; \quad b_i \sim N(0, \Psi), \ \epsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

Donde los parámetros de la ecuación (1) representan: el índice i el tipo de polímero (hasta 7 estudiados), ϕ_{ji} representan los parámetros del modelo para diferentes momentos para cada uno de los i-polímeros estudiados (en el caso j=1 es la masa inicial, j=2 la masa final, j=3 es la masa en el tiempo de máxima degradación y j=4 será la estimación de la razón de masa degradada en el proceso). A su vez estos parámetros pueden descomponerse en dos nuevos parámetros β serán el térmico constante del modelo y b_i las desviaciones en cada polímero, siendo ε_{ji} los errores del modelo, $Time_{ij}$ el tiempo de degradación y $MassLoss_{ij}$ la masa perdida.

Uno de los ajustes resultante puede apreciarse en la Figura 1, mientras que la Figura 2 muestra los residuos del modelo empleado para este mismo ajuste.

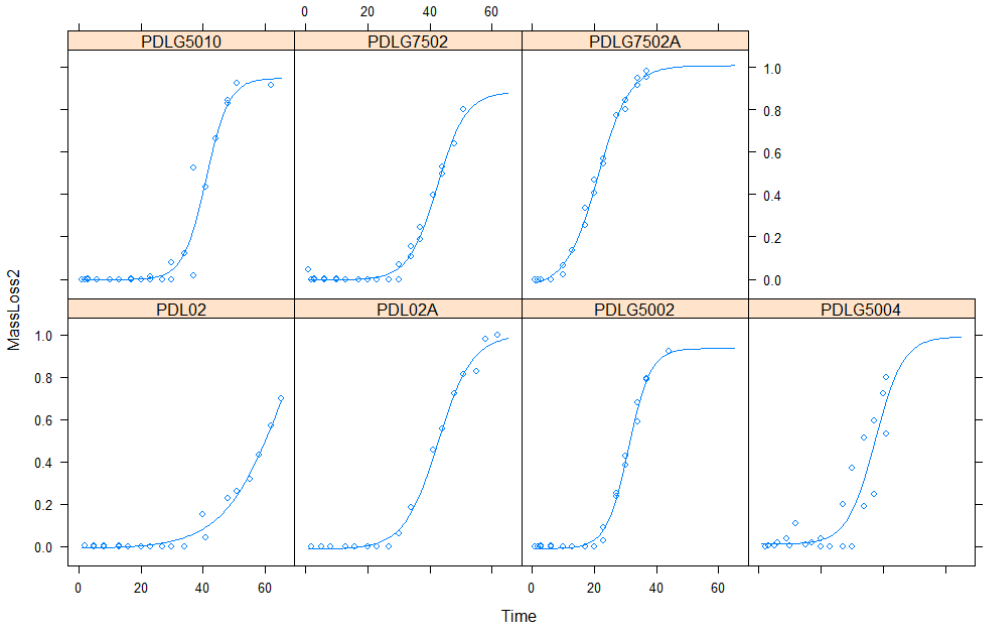


Figura 1: Ajuste de diferentes biopolímeros usando el modelo (1) en función del tiempo

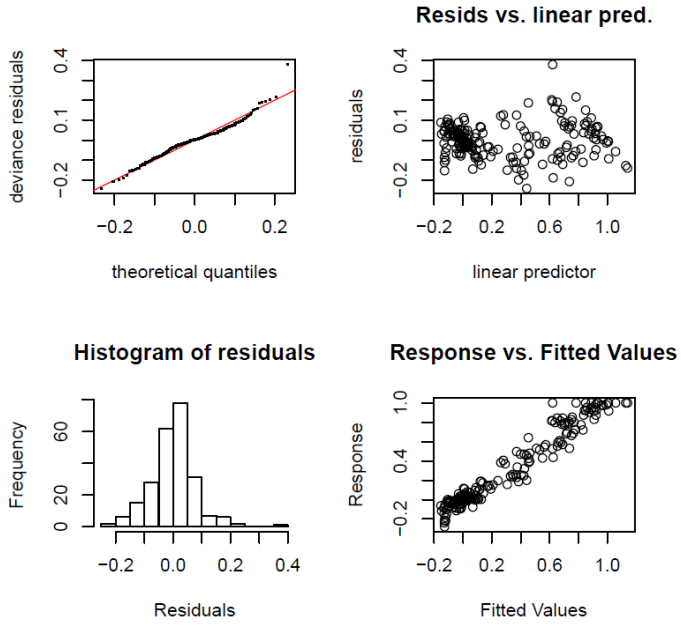


Figura 2: Gráfico de residuos del modelo ajustado.

4. CONCLUSIONES

En este trabajo, se presentan diferentes propuestas para modelizar la pérdida de masa de diferentes biopolímeros usados en aplicaciones odontológicas. Se ha aplicado un modelo GAM para estimar la pérdida de masa frente al tiempo pH y polímero. Además, se han aplicado modelos de regresión no lineal de efectos mixtos, usando una función logística, para modelizar la dependencia entre la masa perdida o avance de la degradación con respecto al tiempo y polímero. La relación entre la pérdida de masa y el tiempo permite comparar los caminos de degradación de este tipo de biopolímeros resultando que los denominados PDL02 y PDLG7502 comienzan a degradarse más tarde y el proceso de degradación es más lento, mientras que el PDLG7502A comienza su pérdida antes. Estos modelos pueden ayudar a comprender la degradación de estos biomateriales, lo que posibilitará su empleo en función de la aplicación odontológica que se requiera al ser materiales biocompatibles con el organismo humano.

Agradecimientos

Los autores agradecen la ayuda de los Proyecto del ministerio MTM2013-41383P (fondos FEDER incluidos), a la empresa DevolopBiosystems por la ayuda prestada, y a la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación del Ecuador (SENESCYT).

REFERENCIAS

Anderson, J.M. and Shive, M.S. 1997. Biodegradation and biocompatibility of PLA and PLGA microspheres. *Advanced Drug Delivery Reviews* 28 (1): 5–24.

ASTM (2012). American Society for Testing and Materials –ASTM D6400-12: Standard specification for labeling of plastics designed to be aerobically composted in municipal or industrial facilities.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*, Wiley.

Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Mixed Effects Models for Repeated Measurement Data,* Chapman and Hall.

Laird, N.M. and Ware, J.H. (1982) *Random-Effects Models for Longitudinal Data*, Biometrics, 38, 963-974.

Lindstrom, M.J. and Bates, D.M. (1990) Nonlinear Mixed Effects Models for Repeated Measures Data, *Biometrics*, 46, 673-687.

Hollister, .S.J. 2005. Porous scaffold design for tissue engineering. *Nature Materials* 4(7): 518–524.

Hoque M.E., Yong, L.C. and Ian, P. Mathematical modeling on degradation of 3d tissue engineering scaffold materials. *Regenerative Research* 1(1) 2012 58-61.

Pinheiro, J.C. and Bates., D.M. (1996) Unconstrained Parametrizations for Variance-Covariance Matrices, *Statistics and Computing*, 6, 289-296.

Pinheiro, J.C., and Bates, D.M. (2000) *Mixed-Effects Models in S and S-PLUS*, Springer.

Pitt, C.G. and Zhong-wei, G. 1987. Modification of the rates of chain cleavage of poly (ϵ-caprolactone) and related polyesters in the solid state. *Journal of controlled release* 4: 283-292.

Prime, R.B., Bair, H.E., Vyazovkin, S., Gallagher, P.K. and Riga, A. 2009. *Thermogravimetric analysis (TGA). In J.D. Menczel & R.B. Prime (eds), Thermal analysis of polymers. Fundamentals and applications*: 241-318. San José: John Wiley & Sons.

Tarrío-Saavedra, J., López-Beceiro, J. Naya. S. and Artiaga R. (2008). Effect of silica content on fumed silica/epoxy composites. *Polymer Degradation and Stability*. 93, 12, 2133-2137.

Vyazovkin, S., Burnham, A.K., Criado, J.M., Pérez-Maqueda, L.A., Popescu, C. & Sbirrazzuoli, N. 2011. *Thermochimica Acta* 520, 1–19.

Wood, S.N. (2013a) A simple test for random effects in regression models. *Biometrika* 100:1005-1010.

Wood, S.N. (2013b) On p-values for smooth components of an extended generalized additive model. *Biometrika* 100:221-228.

Wood, S.N. (2006) *Generalized Additive Models: an introduction with R*, CRC.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

On the Hill estimator: a comparison of methods

Marta Ferreira[1] and Márcio Rebelo[2]

[1]Center of Mathematics of Minho University and CEMAT
[2]Department of Mathematics and Applications of University of Minho

ABSTRACT

Extreme value theory (EVT) deals with the occurrence of extreme phenomena. The tail index is a very important parameter appearing in the estimation of the probability of rare events. Under a semiparametric framework, inference requires the choice of a number k of upper order statistics to be considered. This is the crux of the matter and there is no definite formula to do it, since a small k leads to high variance and large values of k tend to increase the bias. Several methodologies have emerged in literature, specially concerning the most popular Hill estimator (Hill, 1975). In this work we compare through simulation well-known procedures presented in Drees and Kaufmann (1998), Matthys and Beirlant (2000), Beirlant *et al.* (2002) and de Sousa and Michailidis (2004), with a heuristic scheme considered in Frahm *et al.* (2005) within the estimation of a different tail measure but with a similar context. We will see that the new method may be an interesting alternative.

Keywords: Extreme value theory, Tail index estimation Monte-Carlo simulations

1. INTRODUCTION

Extreme value statistics are being increasingly used in the environment, engineering, finance, among other sciences, given the need to account for the possible occurrence of extreme phenomena in the modeling.

Let $\{X_n\}_{n\geq 1}$ be a sequence of independent and identically distributed (iid) random variables (rv's), with common distribution function (df) F. If for some sequences of real constants $a_n > 0$ and b_n, $n \geq 1$, the limit

$$P(\max(X_1, \ldots, X_n) \leq a_n x + b_n) \underset{n\to\infty}{\longrightarrow} G_\xi(x)$$

exists for some non-degenerate function G_ξ with ξ real, then F is said to belong to the max-domain of attraction of G_ξ, denoted $F \in \mathcal{D}(G_\xi)$. This function is called generalized extreme value (GEV) and is given by

$$G_\xi(x) = \exp(-(1+\xi x)^{-1/\xi}), \quad 1 + \xi x > 0,$$

($G_0(x) = \exp(-e^{-x})$). Parameter ξ is the so-called tail index and determines the shape and weight of the tail. Thus, $\xi > 0$ implies a heavy tail (Fréchet max-domain of attraction) with polynomial decay and infinite right-end-point, whenever null means an exponential tail (Gumbel max-domain of attraction) and $\xi < 0$ corresponds to a light tail (Weibull max-domain of attraction) with finite right-end-point.

The tail index plays a determinant role concerning the inference within rare events like, for instance, the estimation of an unusual high quantile (e.g. the Value-at-Risk in finance) or the dual problem of estimating the probability of exceeding a high level x, i.e., $p = 1 - F(x)$. The Hill estimator (Hill 1975),

$$\widehat{\xi}_{k,n} := \frac{1}{k}\sum_{i=1}^{k} \log \frac{X_{n-i+1:n}}{X_{n-k:n}},$$

with $X_{n:n} \geq X_{n-1:n} \geq ... \geq X_{1:n}$ being the order statistics of $\{X_n\}_{n\geq 1}$, is perhaps the most applied in literature and assumes that F belongs to max-domain of attraction of a heavy tail ($\xi > 0$). Observe that it depends on the k upper order statistics, where $k \equiv k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$ (i.e., $k \equiv k_n$ is an intermediate sequence), in order to achieve consistency.

The choice of k is the estimation core, being somewhat ambiguous, since small values of k results in greater variance and large values lead to an increased bias. The plot $(k, \widehat{\xi}_{k,n})$, $1 \leq k < n$, usually shows a plateau region from which we may infer the values of k where $\widehat{\xi}_{k,n}$ approximates the true value (see Figure 1). This problem often accompanies extreme-values semi-parametric inference and has been much studied in the literature. See, e.g., Beirlant *et al.* (2004) for a survey.

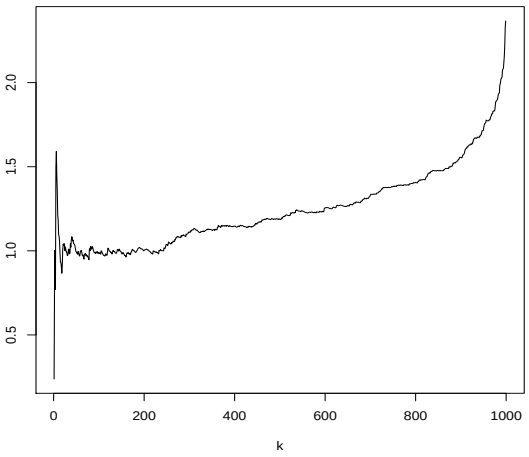


Figure 1: Hill plots of 1000 realizations of a Fréchet model with $\xi = 1$.

For non-null ξ, Beirlant *et al.* (2002) stated, under little restrictive conditions,

$$Y_i := (i+1)\log\frac{X_{n-i:n}\,\widehat{\xi}_{i,n}}{X_{n-(i+1):n}\,\widehat{\xi}_{i+1,n}} = \xi + b(n/k)\left(\frac{i}{k}\right)^{-\rho} + \epsilon_i,\ i = 1, ..., k, \tag{1}$$

where the error term ϵ_i is zero-centered and b is a positive function such that $b(x) \to 0$, as $x \to \infty$. The so called second order tail parameter ρ is usually replaced by $\rho = -1$. For more details, see Beirlant *et al.* 2004 and references therein. Least squares estimators of ξ and $b(n/k)$ derived from (1) are expressed as

$$\widetilde{\xi}^{LS}_{k,n} = \overline{Y}_k - \widetilde{b}^{LS}_{k,n}/(1-\rho) \quad \text{and} \quad \widetilde{b}^{LS}_{k,n} = \frac{(1-\rho)^2(1-2\rho)}{\rho^2}\frac{1}{k}\sum_{i=1}^{k}\left(\left(\tfrac{i}{k}\right)^{-\rho} - \tfrac{1}{1-\rho}\right)Y_i. \tag{2}$$

One method consists in finding the k value that corresponds to the smallest estimate of the Hill's asymptotic mean squared error (AMSE)

$$\text{AMSE}(\widehat{\xi}_{k,n}) = \tfrac{\xi^2}{k} + \left(\tfrac{b(n/k)}{1-\rho}\right)^2. \tag{3}$$

More precisely, for fixed $\rho = -1$, compute the least squares estimates of ξ and $b(n/k)$ by using (2). Obtain the Hill's AMSE in (3) by replacing ξ and $b(n/k)$ with the respective estimates. Take $\widehat{k}^1_{opt}$ as the value of k that minimizes the obtained estimates of the AMSE and estimate ξ as $\widehat{\xi}_{\widehat{k}^1_{opt},n}$.

Alternatively, the value of k for which the respective AMSE is minimal is

$$k_{opt} \sim b(n/k)^{-2/(1-2\rho)}k^{-2\rho/(1-2\rho)}\left(\tfrac{\xi^2(1-\rho)^2}{-2\rho}\right)^{1/(1-2\rho)} \tag{4}$$

Consider again the least squares estimates for ξ and $b(n/k)$ by using (2) and fixed $\rho = -1$, for $k = 3, ..., n$. Compute $\widehat{k}_{opt,k}$ according to the expression (4). Take $\widehat{k}^2_{opt} = median\{\widehat{k}_{opt,k},\ k = 3, ..., \lfloor\frac{n}{2}\rfloor\}$, where $\lfloor x\rfloor$ denotes the largest integer not exceeding x. Estimate ξ by $\widehat{\xi}_{\widehat{k}^2_{opt},n}$.

For more details about these two methods, see Beirlant *et al.* 2002 Matthys and Beirlant (2000) and references therein. We denote them, respectively, AMSE and Kopt.

The adaptive procedure of Drees and Kaufmann (1998), here denoted DK, is based on results for a class of Generalized Pareto models, known has Hall's class, where the maximum random fluctuation of $\sqrt{i}(\widehat{\xi}_{i,n} - \xi)$, $i = 1, ..., k-1$, with $k \equiv k_n$ an intermediate sequence, is of order $\sqrt{\log\log n}$. The derived stopping time criterion seeks the optimum k under which the bias starts to dominate the variance. More precisely, for $\rho = -1$:

1. State $r_n = 2.5 \times \widetilde{\xi} \times n^{0.25}$, where $\widetilde{\xi} = \widehat{\xi}_{2\sqrt{n},n}$.

2. Compute $\widetilde{k}(r_n) := \min\{k = 1, ..., n-1 : \max_{i<k} \sqrt{i}|\widehat{\xi}_{i,n} - \widehat{\xi}_{k,n}| > r_n\}$. If $\sqrt{i}|\widehat{\xi}_{i,n} - \widehat{\xi}_{k,n}| > r_n$ is not satisfied for any k, impute $0.9 \times r_n$ to r_n and repeat step 2. Else move to step 3.

3. For $\varepsilon \in (0, 1)$, usually $\varepsilon = 0.7$, obtain

$$\widehat{k}_{\mathtt{DK}} = \left\lfloor \frac{1}{3}(2\widetilde{\xi}^2)^{1/3} \left(\frac{\widetilde{k}(r_n^{\varepsilon})}{(\widetilde{k}(r_n))^{\varepsilon}} \right)^{1/(1-\varepsilon)} \right\rfloor$$

The procedure considered in Sousa and Michailidis (2004) looks for an estimate of the optimal value k based on the Hill sum plot, (k, S_k), $k = 1, ..., n-1$, where $S_k = k\widehat{\xi}_{k,n}$. Since $E(S_k) = k\xi$, then the sumplot is expected to be approximately linear for the values of k where $\widehat{\xi}_{k,n} \approx \xi$. The slope of this linear part is an estimator of ξ. Thus, an algorithm was developed to establish where there is a breakdown of linearity. Considering the regression model $y = X\xi + \delta$, with $y = (S_1, ..., S_k)'$, $X = [1\, i]_{i=1}^{k}$ and δ the error term, the method is based on a sequential testing procedure discussed in McGee and Carleton (1970). This process checks the null hypothesis that a new point y_0 is adjacent to the left or to the right of the set of points $y = (y_1, ..., y_k)$, through the statistics

$$F = s^{-2} \left((y_0 - \widehat{y}_0^*)^2 + \sum_{i=1}^{k} (\widehat{y}_i - \widehat{y}_i^*)^2 \right),$$

where $*$ denotes the predictions based on $k+1$ and $s^2 = (k-2)^{-1}(y'y - \widehat{\xi}X'y)$. We have F approximately distributed by $F_{1,k-2}$ and thus the null hypothesis is rejected whenever F is larger than the $(1-\alpha)$-quantile, $F_{1,k-2;1-\alpha}$. The algorithm runs along the following steps:

1. Fit a least-squares regression line to the initial $k = \nu n$ upper observations, $y = [y_i]_{i=1}^{k}$ (usually $\nu = 0.02$).

2. Using the test statistic F, determine if a new point $y_0 = y_j$ for $j > k$, belongs to the original set of points y. Go adding points until the null hypothesis is rejected.

3. Consider $k_{\mathtt{new}} = \max(0, \{j : F < F_{1,k-2;1-\alpha}\})$. If $k_{\mathtt{new}} = 0$, no new points are added to y and thus move forward to step 4. Return to step 1. if $k_{\mathtt{new}} > 0$ by considering $k = k_{\mathtt{new}}$.

4. Estimate ξ by considering the obtained k.

Although in a different context concerning the estimation of a bivariate tail dependence measure, but with a similar problematic, Frahm *et al.* (2005) presented a heuristic procedure aiming to find the plane area of the estimator sample path. Ferreira (2014, 2015) adapted this method to the Hill estimator, whose algorithm is described below:

1. Smooth the Hill plot $(k, \widehat{\xi}_{k,n})$ by taking the means of $2b+1$ successive points, $\overline{\widehat{\xi}}_{1,n}, ..., \overline{\widehat{\xi}}_{n-2b,n}$, where the bandwidth $b = \lfloor 0.005 \times n \rfloor$ (and thus each moving average consists in 1% of the data).

2. Define the regions $p_k = (\overline{\widehat{\xi}}_{k,n}, ..., \overline{\widehat{\xi}}_{k+m-1,n})$, $k = 1, ..., n-2b-m+1$, with length $m = \lfloor \sqrt{n-2b} \rfloor$. The algorithm stops at the first region satisfying

$$\sum_{i=k+1}^{k+m-1} \left| \overline{\widehat{\xi}}_{i,n} - \overline{\widehat{\xi}}_{k,n} \right| \leq 2s,$$

where s is the empirical standard-deviation of $\overline{\widehat{\xi}}_{1,n}, ..., \overline{\widehat{\xi}}_{n-2b,n}$.

3. Consider the chosen plateau region p_k* and estimate ξ as the mean of the values of p_k* (consider the estimate zero if no plane region fulfills the stopping condition).

In the sequel it will be referred as plateau method.

In this work we compare through simulation the performance of these methods within the estimation of the tail index and the exceedance probability. A simple estimator under $\xi > 0$ is given by (Dijk and de Haan, 1992)

$$\widehat{p}_{k,n} := \frac{k}{n} \left(\frac{x}{X_{n-k:n}} \right)^{-1/\widehat{\xi}}.$$

In this measure, the plateau algorithm will be applied to the respective sample path $(k, \widehat{p}_{k,n})$ in a similar manner of the variance estimation considered in Ferreira (2015). More precisely, we choose the plane region in step 2 at the same position of the one found for the ξ estimation.

2. SIMULATION STUDY

We consider 1000 independent samples of sizes $n = 100, 1000$, generated from the following models:

- Pareto(ξ) with d.f. $F(x) = 1 - x^{-1/\xi}$, $x > 1$, $\xi > 0$: ($\xi = 1$);
- Cauchy with d.f. $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$, x real: ($\xi = 1$);
- Burr(β,τ,λ), with d.f. $F(x) = 1 - (\beta/(\beta + x^\tau))^\lambda)$, $x > 0$, $\beta, \tau, \lambda > 0$, $\xi = 1/(\tau\lambda)$: we take Burr(1,1,1) ($\xi = 1$)
- Fréchet(ξ) with d.f. $F(x) = \exp\left(-x^{-1/\xi}\right)$, $x > 0$, $\xi > 0$: ($\xi = 1$);
- Log-Gamma(τ,λ), with d.f. $F(x) = \int_1^x \frac{\lambda^\tau}{\Gamma(\tau)} (\log t)^{\tau-1} t^{-\lambda-1} dt$, $x \geq 0$, $\tau, \lambda > 0$, $\xi = 1/\lambda$: we take Log-Gamma(2,1) ($\xi = 1$).

The bias and root mean squared error (rmse) values obtained for the tail index are reported in Table 1. The results concerning the exceedance probabilities, $p = 0.001, 0.0001$, respectively in the cases $n = 100, 1000$, are reported in Table 2. Regarding the tail index, the AMSE method presents the smaller rmse, followed by the DK except in the Pareto model. Indeed, this latter methodology was developed for the so-called Hall's class that leaves out the simple Pareto case considered here. The plateau procedure has also a good performance, in particular, for small sample sizes. The worst performance for $n = 100$ is associated with the Kopt method. The sumplot methodology was developed under the assumption of a Pareto tail behavior and therefore it is not surprising its best performance within the Pareto model. However, it revealed to be the worst method in the remaining models, in particular, for $n = 1000$. In what concerns the exceedance probability, the best results for $n = 100$ lie in the plateau and Kopt methods. In the case $n = 1000$, the methods have a broadly similar performance, except the sumplot with the highest rmse.

3. Acknowledgments

The first author research was supported by the Research Centre CEMAT through the Project UID/Multi/04621/2013.

bias (n=100)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.1136	0.1798	0.1338	0.0424	0.3085
Fréchet	0.0101	0.1489	0.0827	0.0362	0.2158
Cauchy	0.0070	0.1121	0.0282	-0.0306	0.2166
Log-Gamma	0.1512	0.2918	0.1803	0.2242	0.4240
Pareto	-0.0882	0.0349	-0.1396	-0.0139	-0.0311
rmse (n=100)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.2848	0.4546	0.2887	0.3058	0.4243
Fréchet	0.2261	0.4610	0.2371	0.2606	0.3327
Cauchy	0.2953	0.5366	0.2991	0.3737	0.4118
Log-Gamma	0.3003	0.5708	0.3595	0.3523	0.5045
Pareto	0.1975	0.3665	0.4256	0.2010	0.1262
bias (n=1000)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.0710	0.0475	0.0662	0.0142	0.3123
Fréchet	-0.0122	0.0620	0.0577	0.0107	0.2735
Cauchy	0.0199	0.0194	0.0064	-0.0147	0.1707
Log-Gamma	0.1496	0.1896	0.2646	0.2025	0.5249
Pareto	-0.0402	0.0187	-0.0901	-0.0002	-0.0002
rmse (n=1000)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.1319	0.1303	0.1279	0.1465	0.3372
Fréchet	0.1032	0.2446	0.1069	0.1166	0.2927
Cauchy	0.1115	0.1470	0.1152	0.1765	0.2124
Log-Gamma	0.1957	0.3568	0.2833	0.2380	0.5404
Pareto	0.0860	0.2116	0.3648	0.0760	0.0321

Table 1: Bias and rmse of the tail index estimation.

REFERENCES

Beirlant, J., Dierckx, G., Guillou, A. and Stărică, C. (2002) On Exponential Representation of Log-Spacings of Extreme Order Statistics. Extremes, 5, 157-180.

Beirlant,J., Goegebeur, Y., Segers, J. and Teugels, J.L. (2004). Statistics of Extremes: Theory and Applications. J. Wiley & Sons.

de Sousa, B. and Michailidis, G. (2004) A Diagnostic Plot for Estimating the Tail Index of a Distribution. Journal of Computational and Graphical Statistics, 13(4), 1-22.

Dijk, V. and de Haan, L. (1992). On the estimation of the exceedance probability of a high level. Order statistics and nonparametrics: Theory and applications. In Sen, P.K. and Salama I.A. (Editors), 79-92. Elsevier, Amsterdam.

Drees, H. and Kaufmann, E. (1998) Selecting the optimal sample fraction in univariate extreme value estimation. Stochastic Process Appl., 75, 149-172.

Ferreira, M. (2014) A Heuristic Procedure to Estimate the Tail Index, Proceedings of the 14th International Conference in Computational Science and Its Applications - ICCSA 2014, June 30 - July 3 (2014), Guimares, Portugal, IEEE-Computer Society 4264a241, 241-245.

Ferreira, M. (2015) Estimating the tail index: Another algorithmic method. ProbStat Forum, 08, 45-53.

Frahm, G., Junker, M. and Schmidt R. (2005) Estimating the tail-dependence coefficient: properties and pitfalls. Insurance: Mathematics & Economics, 37(1), 80-100.

Hill, B.M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. Ann. Stat., 3, 1163-1174.

Matthys, G. and Beirlant, J. (2000) Adaptive Threshold Selection in Tail Index Estimation. In: Extremes and Integrated Risk Management, ed. P. Embrechts, London: Risk Books.

bias (n=100)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.0002	-0.0008	-0.0004	-0.0009	0.0056
Fréchet	0.0002	-0.0008	0.0001	-0.0006	0.0118
Cauchy	0.0017	-0.0004	0.0009	-0.0006	0.0174
Log-Gamma	0.0003	-0.0008	0.0025	-0.0004	0.0338
Pareto	-0.0004	-0.0007	-0.0008	-0.0004	0.0381
rmse (n=100)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.0020	0.0009	0.0012	0.0009	0.0130
Fréchet	0.0011	0.0009	0.0015	0.0009	0.0241
Cauchy	0.0046	0.0012	0.0039	0.0011	0.0335
Log-Gamma	0.0011	0.0009	0.0053	0.0010	0.0878
Pareto	0.0008	0.0009	0.0009	0.0008	0.0603
bias (n=1000)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.0001	-0.0001	-0.0001	-0.0001	0.0004
Fréchet	0.0001	-0.0001	-0.0001	-0.0001	0.0012
Cauchy	0.0000	-0.0001	-0.0007	-0.0001	0.0008
Log-Gamma	0.0000	-0.0001	0.0003	-0.0001	0.0037
Pareto	-0.0001	-0.0001	-0.0001	-0.0001	0.0503
rmse (n=1000)	AMSE	Kopt	DK	Plateau	Sumplot
Burr	0.0001	0.0001	0.0001	0.0001	0.0007
Fréchet	0.0001	0.0001	0.0001	0.0001	0.0016
Cauchy	0.0001	0.0001	0.0008	0.0001	0.0012
Log-Gamma	0.0001	0.0001	0.0005	0.0001	0.0051
Pareto	0.0001	0.0001	0.0001	0.0001	0.0729

Table 2: Bias and rmse of the exceedance probability estimation, considering $p = 0.001$ for $n = 100$ and $p = 0.0001$ for $n = 1000$.

McGee, V.E. and Carleton, W.T. (1970) Piecewise Regression. Journal of the American Statistical Association, 65, 11091124.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

WEIGHTED ESTIMATION OF CONDITIONAL MEAN FUNCTION WITH TRUNCATED, CENSORED AND DEPENDENT DATA

Han-Ying Liang[1] and María del Carmen Iglesias-Pérez[2]

[1]Department of Mathematics, Tongji University, Shanghai, P. R. China
[2]Department of Statistics and OR, Universidad de Vigo, E. Ingeniería Forestal, Pontevedra

ABSTRACT

We have constructed a new weighted estimator of the conditional mean function for a left-truncated and right-censored model. The function includes regression function, conditional moment as well as conditional distribution function. Under the assumption that the observations form a stationary α-mixing sequence, we have derived the weak convergence with rate and asymptotic normality of the weighted estimator. The asymptotic normality shows that the weighted estimator preserves the bias and variance of the local linear estimator of the conditional mean function. In this work, we investigate the finite sample behavior of the estimator via a simulation study.

Palabras e frases chave: Asymptotic normality; conditional mean function; truncated and censored data; weighted estimator; α-mixing

1. INTRODUCTION

The study of the relationship between a variable of interest Y and a covariate X is one of the most important problems in Statistics. This can be done via estimating the function $m(x) = E(\phi(Y)|X = x)$, where the (known) transformation $\phi(\cdot)$ is introduced to include various targets of interest. For example, taking $\phi(y) = y^r$ we get the rth conditional moment (when $r = 1$, the 1th conditional moment is the regression function), and if we take $\phi(y) = I(y \le t)$ then $m(x)$ becomes the conditional distribution function of Y given $X = x$ at t.

In some fields as reliability, survival analysis or economics, right-censored or/and left-truncated data are often encountered. The random variable Y can be regarded as the lifetime of a device, the survival time of a patient or the duration time of some economic variable. Recently, Liang et al. (2015) constructed for the first time the Nadaraya-Watson (NW) and the local linear (LL) estimators of $m(x)$ and its derivatives for the left truncated and right censored (LTRC) model, and established the asymptotic normality for the NW and LL estimators of $m(x)$ under dependent observations. Motivated by the work of Liang (2012), we, in this paper, use empirical likelihood method to construct a new weighted estimator of $m(x)$ for the LTRC model, and investigate the weak consistency and asymptotic normality of the weighted estimator under α-mixing observations. These results are new, even in independent cases. Theoretical results and a simulation study show that the weighted estimator shares the advantages of both the NW method and LL fitting.

Let (X, Y, T, W) be a random vector, where Y is the lifetime with distribution function (df) F, T is random left truncation time with the df L and W denotes random right censoring time with the df G. Assume that X admits df $V(\cdot)$ and density $v(\cdot)$. In the LTRC model one observes (X, Z, T, δ) if $Z \ge T$, where $Z = \min(Y, W)$ and $\delta = I(Y \le W)$. When $Z < T$ nothing is observed. Clearly, if Y is independent of W, then Z has the df $H = 1 - (1 - F)(1 - G)$. Take $\theta = P(T \le Z)$, then necessarily, we assume $\theta > 0$. Let $(X_i, Z_i, T_i, \delta_i)$, for $i = 1, 2, \cdots, n$, be a stationary random sample from (X, Z, T, δ) which one observes then $(T_i \le Z_i, \forall i)$. Without loss of generality, we assume that Y, T and W are nonnegative random variables, as usual in survival analysis. Iglesias-Pérez and González-Manteiga (1999) defined a generalized product-limit estimator of the conditional distribution function $F(y|x)$ of Y given $X = x$ for the LTRC data by

$$\hat{F}_n(y|x) = 1 - \prod_{i=1}^{n} \Big(1 - \frac{I(Z_i \le y)\delta_i B_{ni}(x)}{\sum_{j=1}^{n} I(T_j \le Z_i \le Z_j) B_{nj}(x)}\Big), \tag{1}$$

where $B_{ni}(x) = K(\frac{x-X_i}{h_n})/\sum_{j=1}^n K(\frac{x-X_j}{h_n})$, K denotes a kernel function, and $0 < h_n \to 0$ is a bandwidth parameter. Under the i.i.d. assumptions, Iglesias-Pérez and González-Manteiga (1999) obtained for the first time an almost sure representation and asymptotic normality of $\hat{F}_n(\cdot)$. Liang et al. (2012) investigated asymptotic properties of $\hat{F}_n(\cdot)$ with α-mixing data.

In the sequel, $\{(X_i, Z_i, T_i, \delta_i) =: \zeta_i,\ 1 \le i \le n\}$ is assumed to be a stationary α-mixing sequence of random vectors. Recall that the sequence $\{\zeta_k, k \ge 1\}$ is said to be α-mixing if the α-mixing coefficient

$$\alpha(n) := \stackrel{\text{def}}{\sup_{k\ge 1}} \sup\{|P(AB) - P(A)P(B)| : A \in \mathcal{F}_{n+k}^{\infty}, B \in \mathcal{F}_1^k\}$$

converges to zero as $n \to \infty$, where $\mathcal{F}_l^m = \sigma\{\zeta_l, \zeta_{l+1}, \ldots, \zeta_m\}$ denotes the σ-algebra generated by $\zeta_l, \zeta_{l+1}, \ldots, \zeta_m$ with $l \le m$. Among various mixing conditions used in the literature, α-mixing is reasonably weak, and has many practical applications. Many stochastic processes and time series are known to be α-mixing. In fact, under very mild assumptions linear autoregressive and more generally bilinear time series models are strongly mixing with mixing coefficients decaying exponentially, i.e., $\alpha(n) = O(\rho^n)$ for some $0 < \rho < 1$. For motivation in the scope of survival analysis, see Cai and Kim (2003) and Liang et al. (2015).

2. ESTIMATORS

In the sequel, for any df $Q(y) = P(\eta \le y)$, its density function is denoted by $q(y)$, we denote the left and right support endpoints by $a_Q = \inf\{y : Q(y) > 0\}$ and $b_Q = \sup\{y : Q(y) < 1\}$, respectively.

For any given $x \in \mathbb{R}$, define $Q(y|x) = P(\eta \le y|X = x)$ and $Q^*(y) = P(\eta \le y|T \le Z)$, their density functions are denoted by $q(y|x)$ and $q^*(y)$, respectively. Set $\theta(x) = P(T \le Z|X = x)$,

$$C(y|x) = P(T \le y \le Z|X = x, T \le Z) \ \text{ and } \ H_1^*(y|x) = P(Z \le y, \delta = 1|X = x, T \le Z).$$

We observe that $V^*(x) = P(X \le x|T \le Z) = \theta^{-1}\int_{-\infty}^x \theta(t)v(t)dt$, which gives $v^*(x) = \theta^{-1}\theta(x)v(x)$. In addition, assume that Y, T and W are conditionally independent at $X = x$, and $F(y|x)$ and $G(y|x)$ are continuous with respect to y then it is easy to verify that

$$C(y|x) = \theta^{-1}(x)L(y|x)(1 - G(y|x))(1 - F(y|x)) = \theta^{-1}(x)L(y|x)(1 - H(y|x))$$

and $H_1^*(y|x) = \theta^{-1}(x)\int_0^y L(t|x)(1 - G(t|x))f(t|x)dt$, which gives $h_1^*(y|x) = \theta^{-1}(x)L(y|x)(1 - G(y|x))f(y|x)$. The estimator of $C(y|x)$ is defined by $\hat{C}_n(y|x) = \sum_{i=1}^n I(T_i \le y \le Z_i)B_{ni}(x)$.

From now on we only consider classes of functions ϕ that satisfy the following conditions:

$(\mathcal{H})$:$(\mathcal{H}1)$ ϕ vanishes outside the interval $[\tau_1, \tau_2]$ for some $a_{L(\cdot|x)} < \tau_1 \le \tau_2 < b_{H(\cdot|x)}$;

$(\mathcal{H}2)$ ϕ is a bounded function on $[\tau_1, \tau_2]$.

Remark. *If the lifetime Y has no left truncation (i.e., $T = 0$), then one can choose $\tau_1 = 0$ in the condition $(\mathcal{H})$, in this case, $(\mathcal{H})$ reduces to assumption $(\mathcal{H})$ of El Ghouch and Van Keilegom (2008).*

Under condition $(\mathcal{H})$, if $\Lambda(\cdot)$ is a bounded function with bounded support, and the functions $v^*(t)$ and $f(y|t)$ are continuous at $t = x$ we find

$$\begin{aligned}
& E\big\{\Lambda_{a_n}(X - x)[\delta\phi(Z)C^{-1}(Z|X)(1 - F(Z|X)) - m(x)]\big\} \\
= \ & \frac{1}{a_n}\int_{\mathbb{R}}\int_{\mathbb{R}} \Lambda\Big(\frac{s-x}{a_n}\Big)v^*(s)\phi(y)C^{-1}(y|s)(1 - F(y|s))h_1^*(y|s)dsdy - \frac{m(x)}{a_n}\int_{\mathbb{R}} \Lambda\Big(\frac{s-x}{a_n}\Big)v^*(s)ds \\
= \ & \int_{\mathbb{R}}\int_{\mathbb{R}} \Lambda(s)v^*(x + a_n s)\phi(y)f(y|x + a_n s)dsdy - \int_{\mathbb{R}} \Lambda(s)v^*(x + a_n s)ds\int_{\mathbb{R}} \phi(y)f(y|x)dy \\
= \ & \int_{\mathbb{R}} \Lambda(s)v^*(x + a_n s)\Big[\int_{\mathbb{R}} \phi(y)(f(y|x + a_n s) - f(y|x))dy\Big]ds \to 0, \qquad (2)
\end{aligned}$$

where $\Lambda_{a_n}(\cdot) = \Lambda(\cdot/a_n)/a_n$ and $0 < a_n \to 0$.

If $m(z)$ is assumed to have $(p+1)$th continuous derivative in a small neighborhood of x, it can be approximated by a polynomial function as

$$m(z) \approx m(x) + \cdots + m^{(p)}(x)(z-x)^p/p! := \beta_0 + \cdots + \beta_p(z-x)^p.$$

Note that (2) suggests that the estimation of $m(x)$ can be viewed as a nonparametric regression of $Y_i^* = \delta_i \phi(Z_i) C^{-1}(Z_i|X_i)(1-F(Z_i|X_i))$ on X_i. Therefore, based on the idea of the local polynomial smoother, the estimator of $(m(x), \cdots, m^{(p)}(x)/p!)^\tau$ is defined as $(\hat{\beta}_0, \cdots, \hat{\beta}_p)^\tau$, which minimizes

$$\sum_{i=1}^{n} \Big(\hat{Y}_i^* - \sum_{j=0}^{p} \beta_j (X_i - x)^j \Big)^2 \Lambda_{a_n}(X_i - x), \tag{3}$$

where $\hat{Y}_i^* = \delta_i \phi(Z_i) \hat{C}_n^{-1}(Z_i|X_i)(1 - \hat{F}_n(Z_i|X_i))$.

From (3), when $p = 0$, we obtain the following NW type estimator of $m(x)$

$$\hat{m}_{NW}(x) = \sum_{i=1}^{n} \hat{Y}_i^* w_i^{NW}(x) \quad \text{with} \quad w_i^{NW}(x) = \frac{\Lambda_{a_n}(X_i - x)}{\sum_{j=1}^{n} \Lambda_{a_n}(X_j - x)},$$

when $p = 1$, the LL estimator of $m(x)$ is

$$\hat{m}_{LL}(x) = \sum_{i=1}^{n} \hat{Y}_i^* w_i^{LL}(x), \quad \text{where} \quad w_i^{LL}(x) = \frac{\Lambda_{a_n}(X_i - x)\{s_{n2} - (X_i - x)s_{n1}\}}{s_{n0}s_{n2} - s_{n1}^2}$$

with $s_{nj} = \sum_{i=1}^{n} (X_i - x)^j \Lambda_{a_n}(X_i - x)$.

Remark. *The estimators $\hat{m}_{NW}(x)$ and $\hat{m}_{LL}(x)$ are constructed for the first time in Liang et al. (2015), who investigated the asymptotic normality of $\hat{m}_{NW}(x)$ and $\hat{m}_{LL}(x)$ under dependent assumptions.*

It is easy to see that the LL weights $w_i^{LL}(x)$ satisfy:

$$\sum_{i=1}^{n} w_i^{LL}(x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} (X_i - x) w_i^{LL}(x) = 0. \tag{4}$$

But for the NW type weights $w_i^{NW}(x)$ this moment condition is not fulfilled.

In view of the information (4), we use the empirical likelihood method to define the new weighted estimator of $m(x)$ for the LTRC model as follows. We first introduce the empirical likelihood function $\mathcal{L} = \prod_{i=1}^{n} p_i(x)$, where $p_1(x), \cdots, p_n(x)$ are subject to the restrictions:

$$p_i(x) \geq 0, \ \sum_{i=1}^{n} p_i(x) = 1 \ \text{ and } \ \sum_{i=1}^{n} p_i(x)(X_i - x)\Lambda_{a_n}(X_i - x) = 0. \tag{5}$$

The maximum of $\mathcal{L}$ can be found via Lagrange multipliers. It may be shown that $\mathcal{L}_{\max} = \prod_{i=1}^{n} \hat{p}_i(x)$, where $\hat{p}_i(x) = \frac{1}{n} \cdot \frac{1}{1+\eta(X_i - x)\Lambda_{a_n}(X_i - x)}$, $i = 1, \cdots, n$, and η is the solution of the following equation:

$$\sum_{i=1}^{n} \frac{(X_i - x)\Lambda_{a_n}(X_i - x)}{1 + \eta(X_i - x)\Lambda_{a_n}(X_i - x)} = 0. \tag{6}$$

The proposed weighted estimator of $m(x)$ is

$$\hat{m}_n(x) = \sum_{i=1}^{n} \hat{Y}_i^* w_{ni}(x) \quad \text{with} \quad w_{ni}(x) = \frac{\hat{p}_i(x)\Lambda_{a_n}(X_i - x)}{\sum_{j=1}^{n} \hat{p}_j(x)\Lambda_{a_n}(X_j - x)}.$$

3. ASYMPTOTIC PROPERTIES

Throughout the paper, let $C, C_1, \cdots$ and $c_0, c, c_1, \cdots$ denote generic finite positive constants, whose values may change from line to line, and let $\Phi(u)$ stand for the standard normal distribution function and $[t]$ be the integer part of t. $A_n = O(B_n)$ means $|A_n| \le C|B_n|$ and $f^{(i,j)}(x,y) := \partial^{i+j} f(x,y)/\partial x^i \partial y^j$. Let $I = [x_1, x_2]$ be an interval contained in the support of $m^*(\cdot)$ such that $\inf_{x \in I_\delta} m^*(x) \ge \delta_0 > 0$ for $I_\delta = [x_1 - \delta, x_2 + \delta]$ with small $\delta > 0$.

Throughout the paper we assume that $\alpha(n) = O(n^{-\lambda})$ for some positive constant λ. In order to formulate the asymptotic results, we need the following assumptions.

(A1) $K(\cdot)$ is a Lipschitz-continuous density function with bounded support and $\int_{\mathbb{R}} tK(t)dt = 0$.

(A2) (i) For $s \in I_\delta$, Y, T and W are conditionally independent at $X = s$;
(ii) τ_1 and τ_2 are two real numbers such that $a_{L(\cdot|x)} < \tau_1 \le \tau_2 < b_{H(\cdot|x)}$ and $a_{L(\cdot|x)} < a_{H(\cdot|x)}$ for $x \in I_\delta$.

(A3) The first 2 derivatives of functions $\theta(s)$ and $v(s)$ are bounded for $s \in I_\delta$, and the first 2 partial derivatives with respect to s of functions $L(y|s)$, $G(y|s)$, $F(y|s)$, $l(y|s)$, $g(y|s)$ and $f(y|s)$ are bounded for $(s,y) \in I_\delta \times \mathbb{R}$.

(A4) For all integers $j \ge 1$, the joint conditional density $v_j^*(\cdot,\cdot)$ of X_1 and X_{j+1} exists on $\mathbb{R} \times \mathbb{R}$ and satisfies $v_j^*(s_1, s_2) \le C_1$ for $(s_1, s_2) \in I_\delta \times I_\delta$.

(A5) (i) $h_n^{-2}(nh_n/\ln(n))^{-(\lambda-3)/2} = O(1)$; (ii) $\sum_{n=1}^{\infty} h_n^{-2}(nh_n/\ln(n))^{-(\lambda-3)/2} < \infty$.

(A6) The kernel $\Lambda(\cdot)$ is a symmetric and bounded density function with a bounded support $[-1, 1]$.

(A7) Function $m(\cdot)$ has continuous second order derivatives at $x \in I$.

(A8) Assume that $na_n \to \infty$, and that the sequence $\alpha(n)$ satisfies for positive integers $q := q_n$ that $q = o((na_n)^{1/2})$ and $\lim_{n\to\infty}(na_n^{-1})^{1/2}\alpha(q) = 0$.

Remark. *Conditions (A1)-(A3) and (A6)-(A7) are standard regularity conditions and used commonly in the literature, see, e.g., Iglesias-Pérez and González-Manteiga (1999) for conditions (A1)-(A3), El Ghouch and Van Keilegom (2008) for conditions (A6)-(A7); Condition (A4) is mainly technical, which is employed to simplify the calculations of covariances in the proof, these assumptions are redundant for the independent setting; the role of condition (A5) is to obtain the rate of convergence of the GPLE $\hat{F}_n(\cdot)$. Condition (A8) is used to prove asymptotic normality for an α-mixing sequence.*

Set $\Delta_{ij} = \int_{\mathbb{R}} s^i \Lambda^j(s)ds$, $\sigma^2(x) = \theta(x)\int_{\mathbb{R}} \frac{\phi^2(y)f(y|x)dy}{L(y|x)(1-G(y|x))} - m^2(x)$ and $\Gamma^2(x) = \frac{\theta\Delta_{02}\sigma^2(x)}{\theta(x)v(x)}$.

Theorem 1 *Let $x \in I$ and $\alpha(n) = O(n^{-\lambda})$ for some $\lambda \ge 2.1$. Suppose that conditions ($\mathcal{H}$), (A1)-(A4), (A5)(i) and (A6)-(A7) are satisfied. If $na_n^{1+r_0} = O(1)$ for some constant $r_0 \ge 2$, then*

$$\hat{m}_n(x) - m(x) = \frac{a_n^2}{2}m''(x)\Delta_{21} + O_p((na_n)^{-1/2}) + o_p(a_n^2) + O_p\Big(\Big(\frac{\ln(n)}{nh_n}\Big)^{1/2} + h_n^2\Big).$$

Theorem 2 *Under the assumptions of Theorem 1 and $\Gamma^2(x) > 0$, if (A8) holds, then*

$$\sqrt{na_n}\Big\{\hat{m}_n(x) - m(x) - \frac{a_n^2}{2}m''(x)\Delta_{21} + o_p(a_n^2) + O_p\Big(\Big(\frac{\ln(n)}{nh_n}\Big)^{1/2} + h_n^2\Big)\Big\} \xrightarrow{\mathcal{D}} N\big(0, \Gamma^2(x)\big).$$

Remark. (a) *It may be seen from Theorem 1 that the weighted estimator $\hat{m}_n(x) \to m(x)$ in probability with a rate, which implies that $\hat{m}_n(x)$ is consistent.*

(b) *For the LL estimator $\hat{m}_{LL}(x)$ of $m(x)$, Liang et al. (2015) proved that*

$$\sqrt{na_n}\Big\{\hat{m}_{LL}(x) - m(x) - \frac{a_n^2}{2}m''(x)\Delta_{21} + o_p(a_n^2) + O_p\Big(\Big(\frac{\ln(n)}{nh_n}\Big)^{1/2} + h_n^2\Big)\Big\} \xrightarrow{\mathcal{D}} N\big(0, \Gamma^2(x)\big).$$

Obviously, the estimators $\hat{m}_{LL}(x)$ and $\hat{m}_n(x)$ of $m(x)$ have same asymptotic normality, i.e., $\hat{m}_n(x)$ matches both the asymptotic bias and the asymptotic variance of $\hat{m}_{LL}(x)$.

4. SIMULATION STUDY

In this section, we investigate with simulated data the finite sample performance of the proposed estimator $\hat{m}_n(x)$in the case $\phi(y) = y$. In particular, (i) we compare the global mean squared errors of the estimators $\hat{m}_n(x)$, $\hat{m}_{NW}(x)$ and $\hat{m}_{LL}(x)$ for different sample sizes, truncation rates and percentage of censoring, and we explore their graphical fit to the true underlying curve; (ii) we examine how good the asymptotic normality is by quantile-quantile-plots of the estimator $\hat{m}_n(x)$ at a specific point x.

In order to obtain an α-mixing observed sequence $\{X_i, Z_i, T_i, \delta_i\}$, we generate the observed data as in the simulated study by Liang el al. (2015), which is as follows.

(1) Drawing of the first observation $(X_1, Z_1, T_1, \delta_1)$ in the final sample.

Step 1. Draw $e_1 \sim N(0,1)$, take $X_1 = 0.5e_1$;

Step 2. Compute Y_1 and W_1, respectively, from the model $Y_1 = \sin(\pi X_1)+\phi_1(1+0.3\cos(\pi X_1))\epsilon_1$, $W_1 = \sin(\pi X_1) + 0.5\phi_2(1 + 0.3\cos(\pi X_1)) + \phi_3(1 + 0.3\cos(\pi X_1))\tilde{\epsilon}_1$, where both ϵ_1 and $\tilde{\epsilon}_1$ are $N(0,1)$ random variables, ϵ_1, $\tilde{\epsilon}_1$ and X_1 are mutually independent, and ϕ_i $(i = 1, 2, 3)$ are chosen (see below) to control the percentage of censoring. Take $Z_1 = \min(Y_1, W_1)$, $\delta_1 = I(Y_1 \leq W_1)$;

Step 3. Draw independently $T_1 \sim N(\mu, 1)$, where μ is adapted in order to get different values of θ. If $Z_1 < T_1$, reject the datum $(X_1, Z_1, T_1, \delta_1)$ and go back to *Step 2*; do this until $Z_1 \geq T_1$.

(2) Drawing of the second observation $(X_2, Z_2, T_2, \delta_2)$ in the final sample.

Step 4. Draw X_2 according to the AR(1) model $X_2 = \rho X_1 + 0.5e_2$, where $e_2 \sim N(0,1)$ is independent of X_1, and $|\rho| < 1$ is some constant, which is chosen to control the dependence of the observations;

Step 5. Compute Y_2 and W_2, respectively, from the model $Y_2 = \sin(\pi X_2)+\phi_1(1+0.3\cos(\pi X_2))\epsilon_2$, $W_2 = \sin(\pi X_2) + 0.5\phi_2(1 + 0.3\cos(\pi X_2)) + \phi_3(1 + 0.3\cos(\pi X_2))\tilde{\epsilon}_2$, where both ϵ_2 and $\tilde{\epsilon}_2$ are $N(0,1)$ random variables, and ϵ_2, $\tilde{\epsilon}_2$ and X_2 are mutually independent. Take $Z_2 = \min(Y_2, W_2)$, $\delta_2 = I(Y_2 \leq W_2)$;

Step 6. Draw independently $T_2 \sim N(\mu, 1)$. If $Z_2 < T_2$, reject the datum $(X_2, Z_2, T_2, \delta_2)$ and go back to *Step 5*; do this until $Z_2 \geq T_2$.

By replicating the process (2) above, we generate the observed data $(X_i, Z_i, T_i, \delta_i)$, $i = 1, \cdots, n$. The generating process shows that $X_i = \rho X_{i-1} + 0.5e_i$, $Y_i = \sin(\pi X_i) + \phi_1(1 + 0.3\cos(\pi X_i))\epsilon_i$, $W_i = \sin(\pi X_i) + 0.5\phi_2(1 + 0.3\cos(\pi X_i)) + \phi_3(1 + 0.3\cos(\pi X_i))\tilde{\epsilon}_i$, $Z_i = \min(Y_i, W_i)$, $\delta_i = I(Y_i \leq W_i)$, where $e_i \sim N(0,1)$, $\epsilon_i \sim N(0,1)$, $\tilde{\epsilon}_i \sim N(0,1)$ and $T_i \sim N(\mu, 1)$, and everything is distributed conditionally on $Z_i \geq T_i$. Note that the α-mixing property of the observable X_i is immediately transferred to the $(X_i, Z_i, T_i, \delta_i)$. Hence, the true underlying regression function is $m(x) = \mathbb{E}(Y|X = x) = \sin(\pi x)$.

For the proposed estimators, we employ the kernel $K(x) = \Lambda(x) = \frac{15}{16}(1 - x^2)^2 I(|x| \leq 1)$. In addition, the parameters ϕ_i $(i = 1, 2, 3)$ allow to control the percentage of censoring (PC) which is given by

$$\begin{aligned}\text{PC} =& P(Y_i > W_i | X_i = x) = 1 - \Phi\Big(\frac{0.5\phi_2}{\sqrt{\phi_1^2 + \phi_3^2}}\Big) \\ &\overset{\phi_1=\phi_3=0.7}{=} 1 - \Phi\Big(\frac{5\sqrt{2}}{14}\phi_2\Big) = \begin{cases} 10\%, & \text{when } \phi_2 = 2.537, \\ 30\%, & \text{when } \phi_2 = 1.038, \\ 50\%, & \text{when } \phi_2 = 0. \end{cases}\end{aligned}$$

In simulation below, we take $\phi_1 = \phi_3 = 0.7$.

(i) Consistency

For comparing the global performance of the estimators $\hat{m}_n(x)$, $\hat{m}_{NW}(x)$ and $\hat{m}_{LL}(x)$, we compute for each estimator $\hat{m}_{h,a}$ of m the global mean square error (GMSE) along $M = 500$ Monte Carlo trials and a grid of bandwidths $h := h_n$ and $a := a_n$; the GMSE is defined as

$$GMSE(h,a) = \frac{1}{Mn}\sum_{l=1}^{M}\sum_{k=1}^{n}[\hat{m}_{h,a}(X_{k,l}) - m(X_{k,l})]^2$$

where $X_{k,l}$ is the k-th datum in the l-th trial. The minimal values of $GMSE(h,a)$ along the grid, and the corresponding bandwidths minimizing the error, are reported in Table 1. Specifically, for $n = 100$, h_n and a_n range from 0.05 to 0.80 with a step 0.01; for $n = 300$, h_n ranges from 0.05 to 0.80 with step 0.01 and a_n ranges from 0.05 to 0.60 with step 0.01. When $h_n = 0.05$ was obtained, then the grid was expanded.

ρ	θ	PC	n	$\hat{m}_n$	a_n	h_n	$\hat{m}_{NW}$	a_n	h_n	$\hat{m}_{LL}$	a_n	h_n
0.1	60%	10%	100	5.9923×10^{-2}	0.36	0.04	6.2453×10^{-2}	0.32	0.03	6.0215×10^{-2}	0.41	0.04
			300	3.0086×10^{-2}	0.33	0.07	3.1718×10^{-2}	0.29	0.06	2.9861×10^{-2}	0.37	0.10
		50%	100	9.7991×10^{-2}	0.45	0.48	9.3781×10^{-2}	0.43	0.49	9.4899×10^{-2}	0.52	0.48
			300	5.4613×10^{-2}	0.37	0.41	5.2431×10^{-2}	0.41	0.68	5.2705×10^{-2}	0.44	0.45
	90%	10%	100	4.3330×10^{-2}	0.40	0.22	4.4696×10^{-2}	0.36	0.34	4.1301×10^{-2}	0.46	0.26
			300	2.0164×10^{-2}	0.32	0.24	2.0731×10^{-2}	0.31	0.45	1.9873×10^{-2}	0.36	0.23
		50%	100	8.5158×10^{-2}	0.46	0.72	8.6258×10^{-2}	0.42	0.67	8.3940×10^{-2}	0.51	0.71
			300	4.5092×10^{-2}	0.38	0.53	4.3951×10^{-2}	0.39	0.68	4.3736×10^{-2}	0.47	0.66
0.5	60%	10%	100	6.0644×10^{-2}	0.39	0.05	6.2637×10^{-2}	0.36	0.05	5.9977×10^{-2}	0.44	0.07
			300	3.1768×10^{-2}	0.34	0.09	3.2710×10^{-2}	0.31	0.09	3.1498×10^{-2}	0.35	0.09

Table 1: Minimum GMSE's of $\hat{m}_n$, $\hat{m}_{NW}$ and $\hat{m}_{LL}$ along $M = 500$ Monte Carlo trials, and corresponding optimal bandwidths, for several truncation rates and percentage of censoring (PC).

In Table 1 we see that the minimum GMSE decreases as the sample size n or the no truncation proportion θ increase, and that it increases as the dependence of the observations increases (ρ increases) or as the percentage of censoring PC increases. All these features were expected. More interestingly, we can appreciate how the new estimator $\hat{m}_n$ outperforms the NW estimator in most of the cases and how the local linear smoother outperforms both of them. It is also seen in Table 1 that the bandwidth a_n decreases as the sample size increases or the percentage of censoring decreases, and that a_n for the new estimator is between the a_n for the NW estimator and the a_n for the LL estimator. The bandwidth h_n is very similar for the three estimators and it increases as the percentage of censoring increases. For the sample sizes of this study we can not see clearly that h_n is a decreasing function of n.

In Figure 1, we plot the target curve $m(x) = \sin(\pi x)$ together with the estimators $\hat{m}_n(x)$ (denoted by m_{NMP}), $\hat{m}_{NW}(x)$ and $\hat{m}_{LL}(x)$, respectively, averaged along the 500 Monte Carlo replications for the two different sample sizes, $\rho = 0.1$, $PC = 10\%$, $\theta = 60\%$, a_n and h_n from Table 1. From this Figure it is seen that all estimators perform better with a larger n, and that the new NWP estimator and the LL estimator fit better the target particularly at the boundaries.

(ii) Asymptotic normality

We examine how good is the asymptotic normal approximation of the new estimator $\hat{m}_n(x)$ at $x = 1$. Note that the regression function at $x = 1$ is $sin(\pi) = 0$. In Figure 2, we plot the histograms and the Q-Q plots of $\hat{m}_n(x)$ for $\rho = 0.1$, $PC = 10\%$, $\theta = 60\%$, a_n and h_n from Table 1, based on $M = 250$ replications with sample size $n = 100$ and 300, respectively. From Figure 2, it is seen that the normality in the distribution of the estimators is improved when n increases. Also, the consistency of the estimator can be observed.

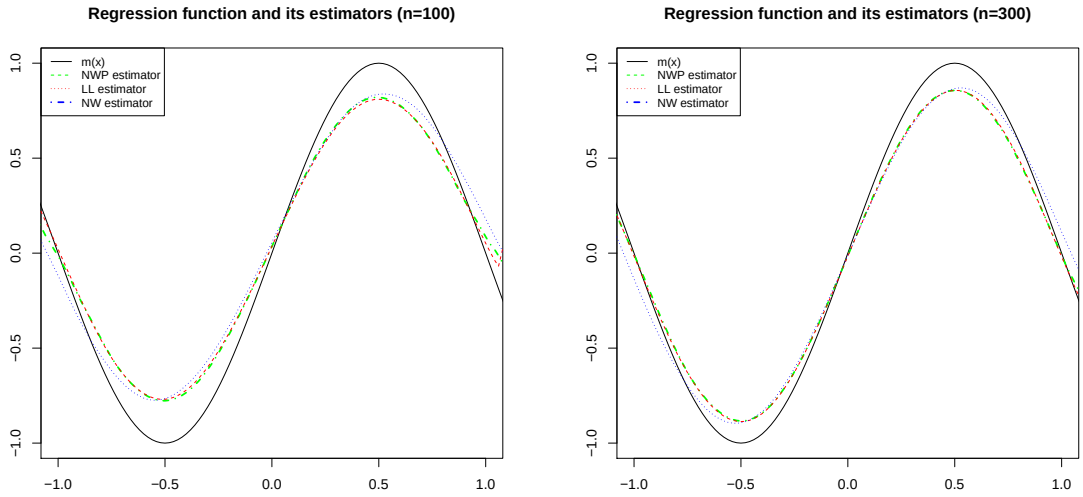


Figure 1. Regression function $m(x) = \sin(\pi x)$ and its estimators averaged along 500 Monte Carlo trials with $\rho = 0.1$, $PC = 10\%$, $\theta = 60\%$, a_n and h_n from Table 1, n =100 (left) and 300 (right).

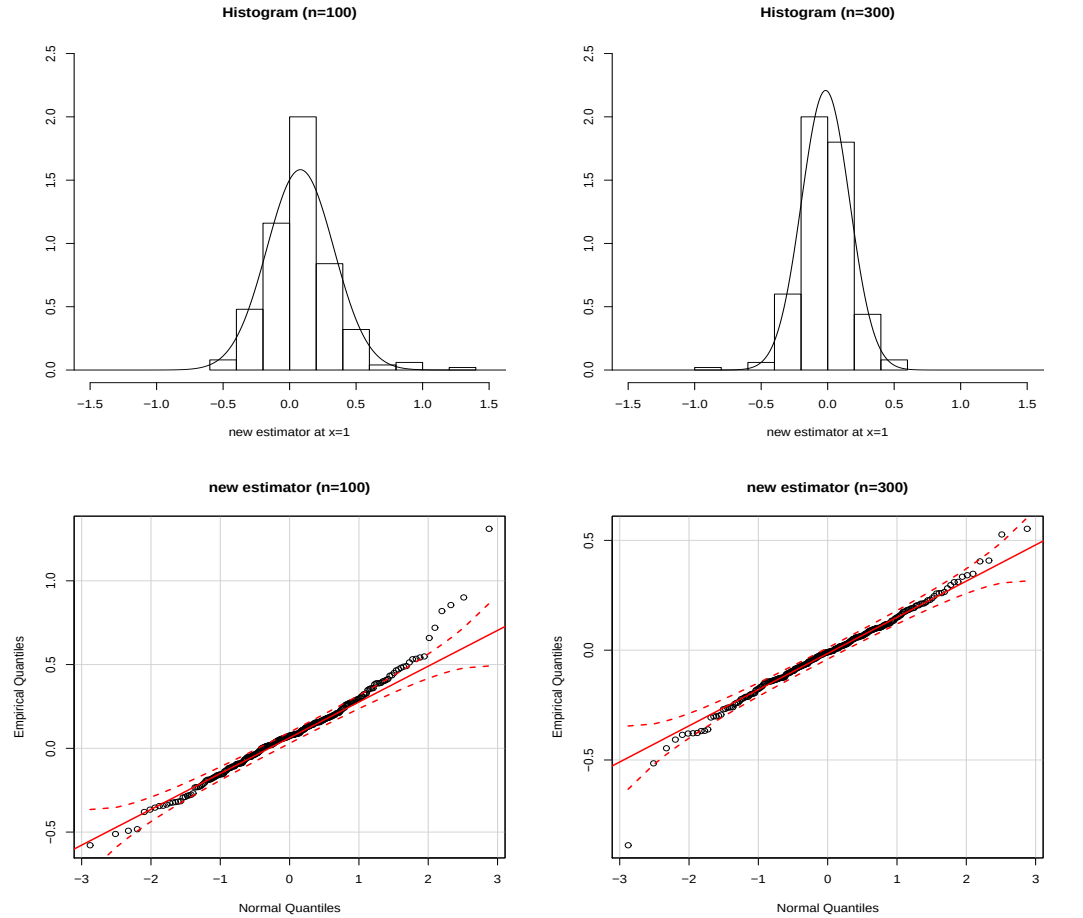


Figure 2. The histograms and de qq-plots of $\hat{m}_n(x)$ at $x = 1$, based on 250 Monte Carlo trials, with $\rho = 0.1$, $PC = 10\%$, $\theta = 60\%$, a_n and h_n from Table 1, n =100 and 300, respectively.

ACKNOWLEDGEMENTS

The first author was supported by the National Natural Science Foundation of China (11271286) and the Specialized Research Fund for the Doctor Program of Higher Education (20120072110007) of China. The second author was supported by the Grant MTM2011-23204 (FEDER support included) of the Spanish Ministry of Science.

REFERENCES

Cai, J. and Kim, J. (2003) Nonparametric quantile estimation with correlated failure time data. Lifetime Data Analysis, 9, 357-371.

El Ghouch, A. and Van Keilegom, I. (2008) Non-parametric regression with dependent censored data. Scandinavian Journal of Statistics, 35(2), 228-247.

Iglesias-Pérez, C. and González-Manteiga, W. (1999) Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. Journal of Nonparametric Statistics, 10, 213-244.

Liang, H. Y. (2012) Weighted nonparametric regression estimation with truncated and dependent data. Journal of Nonparametric Statistics, 24(4), 1051-1073.

Liang, H. Y. and de Uña-Álvarez, J. and Iglesias-Pérez, M. C. (2012) Asymptotic properties of conditional distribution estimator with truncated, censored and dependent data. Test, 21(4), 790-810

Liang, H. Y., de Uña-Álvarez, J. and Iglesias-Pérez, M. C. (2015) A CLT in nonparametric regression with truncated, censored and dependent data. Scandinavian Journal of Statistics, 42, 256 269.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

A comparative simulation study about the model checks for nonparametric regression with missing data

Cotos-Yáñez, T. R.[1], Pérez-González, A.[1] e González-Manteiga, W.[2]

[1] Departamento de Estadística e Investigación Operativa, Universidad de Vigo
[2] Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela

ABSTRACT

Our objective is to analyze the behavior of the goodness of fit test for regression models using test statistics based on an estimation of the integrated regression function, when there are missing observations in the response variable or in some of the covariates. We propose several empirical process using only the complete observations or replacing the missing observations with imputed values. In the case of missing covariates we use a link model to fill the missing observations from the other complete covariates. In all the situations, the calibration test is obtained by Bootstrap methodology. A simulation study compares the proposed tets with smoothed tests previously studied. The study reflects the effect of the correlation between the covariates in the tests based on the imputed sample for the case of missing covariates. Finally two real data examples are shown.

Keywords Empirical process, goodness of fit test, missing observations, regression Model.

1. IntroduciNón

Our interest lies in analyzing, through a simulation study, the behavior of several test statistics based on empirical processes in both settings: missing in the response and missing in the covariates. The test statistics are calculated over the complete data, imputed data and weighting the observations by the missing data model. The study compares our tests with smoothed tests previously studied by González-Manteiga *et al.* (2006) for the case of missing response. Moreover, the behavior of the tests assuming missing data model known and estimated are also compared. The calibration of the tests is obtained by Bootstrap methodology. we propose a strategy to reduce the computational cost based on binning techniques. The implementation of this strategy can be seen in the application to real data.

2. Missing data in the response variable

Let us consider the regression model:

$$Y = m(X) + \eta,$$

where η is the error, assumed to have a zero mean. In missing response data, it may be possible that Y_i is not observed for any index i, which implies that we are faced with: (X_i, Y_i) if Y_i is observed, and $(X_i, ?)$ if Y_i is missing.

To control whether or not an observation is complete, a new variable δ is introduced into the model, constituting an indicator of the missing observations. Thus for each index i, $\delta_i = 1$ if Y_i is observed, and zero if Y_i is missing.

$$P(\delta = 1|Y, X) = P(\delta = 1|X) = p(X), \quad X \in \mathbb{R}^d. \tag{1}$$

The empirical processes considered in this setting are given by:

$$R_{n,S}(x) = n_1^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \delta_i \left(Y_i - \widehat{m}_{\theta,S}(X_i)\right),$$

$$R_{n,S,w}(x) = n_1^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \frac{\delta_i}{\widehat{p}(X_i)} \left(Y_i - \widehat{m}_{\theta,S}(X_i)\right),$$

where $\widehat{m}_{\theta,S}(\cdot)$ is a parametric estimator of the regression function under the null hypothesis with the complete data, $n_1 = \sum_{i=1}^{n} \delta_i$, and $\widehat{p}$ is a estimator of the missing data model (1).

Also, we propose three processes based on the imputed samples. The first process, called imputed, is defined as:

$$R_{n,I}(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left[\left(\delta_i Y_i + (1-\delta_i)\widehat{m}_{h,S}(X_i)\right) - \widehat{m}_{\theta,S}(X_i)\right],$$

where $\widehat{m}_{h,S}(\cdot)$ is a nonparametric estimator of the regression function with the complete data and with bandwidth parameter h. The second one considers the nonparametric estimations of all the responses. We denote it as the averaged process

$$R_{n,A}(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left(\widehat{m}_{h,S}(X_i) - \widehat{m}_{\theta,S}(X_i)\right).$$

And finally a mixture of the two previous processes:

$$R_{n,I,w}(x) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\{X_i \leq x\}} \left[\frac{\delta_i}{\widehat{p}(X_i)} \left(Y_i - \widehat{m}_{\theta,S}(X_i)\right) + \left(1 - \frac{\delta_i}{\widehat{p}(X_i)}\right) \left(\widehat{m}_{h,S}(X_i) - \widehat{m}_{\theta,S}(X_i)\right)\right].$$

As test statistics, we consider the Kolmogorov-Smirnov and the Cramér-von Mises functionals applied to the previously defined processes.

3. Missing data in the covariate

We split the covariate X as $X = (X^c, X^m)$, where X^c contains those variables that are completely observed in all individuals whereas X^m represents the variables with missing observations. Let us reconsider the regression model:

$$Y = m(X^c, X^m) + \eta, \tag{2}$$

where η is the error term, assumed to have a mean of zero.

To control if an observation is complete, a new variable ξ is introduced into the model, constituting an indicator of the missing observations. We assume that the whole vector X^m is not observed. Without loss of generality, we suppose that X^m is one-dimensional.

In order to obtain information about the missing covariate we consider the following regression model:

$$X^m = g(X^c) + \zeta, \tag{3}$$

where X^m is now considered as the response variable, X^c are the covariates totally observed and ζ has a mean of zero.

We use the previous model (3) to impute the missing observations of the covariate X^m. In order to obtain a consistent estimation of g in (3), it is necessary to assume a missing model MAR:

$$P(\xi = 1|Y, X^c, X^m) = P(\xi = 1|X^c) = \pi(X^c).$$

First, we consider the the simplified process using only the complete observations:

$$R_{n,S}(x^c, x^m) = n_1^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\left\{(X_i^c, X_i^m) \leq (x^c, x^m)\right\}} \xi_i \left(Y_i - \widehat{m}_{\theta,S}(X_i^c, X_i^m)\right), \tag{4}$$

where $n_1 = \sum_{i=1}^{n} \xi_i$ and $\widehat{m}_{\theta,S}(\cdot)$ is a parametric estimator of the regression function under the null hypothesis using only the complete observations.

The other proposal imputes the missing observations.

$$R_{n,I}(x^c, x^m) = n^{-\frac{1}{2}} \sum_{i=1}^{n} I_{\left\{\left(X_i^c, \hat{X}_i^m\right) \leq (x^c, x^m)\right\}} \left[\left(Y_i - \widehat{m}_{\theta,I}\left(X_i^c, \hat{X}_i^m\right)\right)\right], \tag{5}$$

where $\hat{X}_i^m = \xi_i X_i^m + (1 - \xi_i)\, \hat{g}_S(X_i^c)$, $\hat{g}_S$ an estimation of the (3) with complete data (simplified version). Finally, $\widehat{m}_{\theta,I}(\cdot)$ is an estimator of the regression function under the null hypothesis with the imputed covariate. As test statistics, we consider once more the statistics of Kolmogorov-Smirnov and the Cramér-von Mises, applied to the previous processes.

4. Simulation Study

4.1. Missing Response

We consider the following regression model:

$$Y_i = 5X_i + aX_i^2 + \sigma(X_i)\,\varepsilon_i, \tag{6}$$

where X_i were generated from the uniform distribution in the interval $[0, 1]$, $\varepsilon_i \sim Normal(0, 1)$ and $a = 0, 1, 3, 5, 7, 9$, the parameter $a \neq 0$ represents the deviation of a linear model. The sample sizes considered were $n = 50$ and $n = 100$. The missing data model considered was MAR given by $p(x) = 0.4 + 0.5\,(\cos(2x + 0.4))^2$.

Our objective is to test the null hypothesis:

$$H_0 : m \in \{m_\theta(x) = \theta_0 + \theta_1 x\}$$

Wild Bootstrap resampling was performed 1000 times for each sample in order to approximate the quantile of order $1 - \alpha$ with $\alpha = 0.01, 0.05$ and 0.1.

The bandwidth parameter of $\widehat{m}_{h,S}$ was selected by Cross-validation method over the complete subsample, i.e. $\{(X_i, Y_i, \delta_i = 1)\}$.

The experiment was iterated 1000 times and the percentage of rejections was calculated. Table 1 shows a scheme of the test statistics studied for this context.

	Kolmogorov-Smirnov	Cramér-von Mises
Complete sample	$D_{n,C} = \sup_x \lvert R_n(x)\rvert$	$W_{n,C}^2 = \int_{\mathbb{R}^d} [R_n(x)]^2 F_n(dx)$
Simplified	$D_{n,S} = \sup_x \lvert R_{n,s}(x)\rvert$	$W_{n,S}^2 = \int_{\mathbb{R}^d} [R_{n,S}(x)]^2 F_n(dx)$
Simplified Weighted	$D_{n,S,w} = \sup_x \lvert R_{n,s,w}(x)\rvert$	$W_{n,S,w}^2 = \int_{\mathbb{R}^d} [R_{n,S,w}(x)]^2 F_n(dx)$
Imputed	$D_{n,I} = \sup_x \lvert R_{n,I}(x)\rvert$	$W_{n,I}^2 = \int_{\mathbb{R}^d} [R_{n,I}(x)]^2 F_n(dx)$
Averaged	$D_{n,A} = \sup_x \lvert R_{n,A}(x)\rvert$	$W_{n,A}^2 = \int_{\mathbb{R}^d} [R_{n,A}(x)]^2 F_n(dx)$
Imputed Weighted	$D_{n,I,w} = \sup_x \lvert R_{n,I,w}(x)\rvert$	$W_{n,I,w}^2 = \int_{\mathbb{R}^d} [R_{n,I,w}(x)]^2 F_n(dx)$

Table 1: Kolmogorov-Smirnov and the Cramér-von Mises statistics over empirical processes when the response variable has missing observations.

In a first step, we assume that $p(x)$ is known. Table 2 shows the behavior of the test statistics for $n = 50$ and $n = 100$ with different levels $\alpha = 0.01, 0.05$ and 0.1. As expected the results are better with $n = 100$ and the test approximate better the values of α.

	n=50			n=100		
$\mathbf{a=0}$	1%	5%	10%	1%	5%	10%
$D_{n,C}$	0.3	4.2	11.7	1.3	6.0	11.0
$D_{n,S}$	0.3	4.8	9.8	0.3	4.9	9.8
$D_{n,I}$	1.2	6.4	11.4	0.8	4.5	10.1
$D_{n,S,w}$	0.8	6.5	12.5	1.2	5.1	10.4
$D_{n,A}$	0.9	6.4	12.1	0.9	5.3	10.7
$D_{n,I,w}$	1.1	6.7	13.4	1.3	5.8	11.6
$W^2_{n,C}$	0.9	4.7	10.5	0.7	5.1	10.4
$W^2_{n,S}$	0.5	5.2	10.9	0.9	5.3	9.2
$W^2_{n,I}$	1.5	6.8	11.6	1.5	4.4	9.3
$W^2_{n,S,w}$	0.9	6.3	13.3	1.5	5.2	10.3
$W^2_{n,A}$	0.8	6.6	13.2	1.2	5.6	11.1
$W^2_{n,I,w}$	1.0	6.2	12.5	1.1	5.4	10.6

Table 2: Percentage of times that the null hypothesis is rejected with $\sigma(x) = x + 1$, with $n = 50$, $n = 100$ and $\alpha = 1\%, 5\%, 10\%$

The previous simulations were made using the true model of missing data, the following table shows the results obtained when the missing probability is nonparametically estimated (rows with $\widehat{p}$). We can see that the percentage of times that the null hypothesis is rejected is similar with the true p than with the estimated p. This shows that, in general, the effect of to estimate the missing data model (1), does not introduce major changes in the behavior of the tests.

$\mathbf{b=1}$		**a=0**	**a=1**	**a=3**	**a=5**	**a=7**	**a=9**
$D_{n,S,w}$	p	5.3	7.8	15.0	32.9	55.2	78.3
	$\widehat{p}$	4.8	8.0	12.5	30.4	49.8	74.0
$D_{n,I,w}$	p	6.9	8.1	15.7	34.7	56.5	80.7
	$\widehat{p}$	5.5	8.6	14.7	32.0	52.4	75.7
$W^2_{n,S,w}$	p	5.8	7.8	16.8	36.1	62.3	84.5
	$\widehat{p}$	5.7	7.1	14.9	35.2	61.4	83.0
$W^2_{n,I,w}$	p	6.0	7.6	16.3	36.2	60.6	83.8
	$\widehat{p}$	5.9	7.0	15.8	35.3	60.4	81.9

Table 3: Percentage of times that the null hypothesis is rejected with $\sigma(x) = x + 1$ when $p(x)$ is real p or estimated nonparametrically $\widehat{p}$ with $n = 100$.

4.2. Missing Covariate

We consider the following bidimensional regression model:

$$Y_i = X_i^c + X_i^m + a\,(X_i^c)^2 + \eta_i,$$

where η_i and X_i^c are independent variables $Normal\,(0, 0.5)$. We consider

$$X_i^m = \gamma X_i^c + \zeta, \tag{7}$$

with $\zeta \sim N(0, 0.5)$. The missing data model considered was $p(x) = (1 + \exp(-0.5 - 2x))^{-1}$. To simulate the model and draw conclusions is very important how is the relationship between the variables (X^c, X^m). For this, we have considered scenarios varying the correlation: $\rho_{(X^c, X^m)} = 0.64, 0.80, 0.98$.

Various imputation techniques were chosen to compared with the complete (subscript C) and simplified case (subscript S):

- Subscript *true*: corresponds to imputation under the true model, i.e., if $\delta_i = 0$, $\hat{X}_i^m = \gamma X_i^c$.
- Subscript *ls*: corresponds to imputation under the null hypothesis, i.e., in this case $\delta_i = 0$, $\hat{X}_i^m = \hat{\gamma} X_i^c$ with $\hat{\gamma}$ is estimated by least squares.

- Subscript *np*: corresponds to a nonparametric imputation, i.e, if $\delta_i = 0$, $\hat{X}_i^m = \hat{g}\left(X_i^c\right)$ with $\hat{g}\left(\cdot\right)$ is a nonparametric regression estimator.

Table 4 shows a scheme of the test statistics considered for this scenario.

	Kolmogorov-Smirnov	Cramér-von Mises
Complete sample	$D_{n,C} = \sup_x \lvert R_n\left(x\right)\rvert$	$W_{n,C}^2 = \int_{\mathbb{R}^d} \left[R_n\left(x\right)\right]^2 F_n\left(dx\right)$
Simplified	$D_{n,S} = \sup_x \lvert R_{n,s}\left(x\right)\rvert$	$W_{n,S}^2 = \int_{\mathbb{R}^d} \left[R_{n,S}\left(x\right)\right]^2 F_n\left(dx\right)$
True Model Imputed	$D_{n,I_{true}} = \sup_x \lvert R_{n,I}\left(x\right)\rvert$	$W_{n,I_{true}}^2 = \int_{\mathbb{R}^d} \left[R_{n,I}\left(x\right)\right]^2 F_n\left(dx\right)$
Least Squares Imputed	$D_{n,I_{ls}} = \sup_x \lvert R_{n,I}\left(x\right)\rvert$	$W_{n,I_{ls}}^2 = \int_{\mathbb{R}^d} \left[R_{n,I}\left(x\right)\right]^2 F_n\left(dx\right)$
Nonparametric Imputed	$D_{n,I_{np}} = \sup_x \lvert R_{n,I}\left(x\right)\rvert$	$W_{n,I_{np}}^2 = \int_{\mathbb{R}^d} \left[R_{n,I}\left(x\right)\right]^2 F_n\left(dx\right)$

Table 4: Kolmogorov-Smirnov and the Cramér-von Mises statistics over empirical processes when the a covariate has missing observations.

In all cases, we have applied the nonparametric bootstrap , adapted to this context, in order to approximate the percentage of rejections under the null hypothesis.

Table 5 shows the percentage of rejections under the null hypothesis and alternative hypothesis, i.e. $a = 0$ and with $a = 0.05, 0.1$. We can observe than the imputed tests based on Cramer-von Misses statistic $\left(W_{n,I_{true}}, W_{n,I_{ls}}, W_{n,I_{np}}\right)$ perform better as long as the correlation coefficients grows. The imputed tests under the null hypothesis perform better than the simplified tests $(D_{n,S}, W_{n,S})$. Under the alternative happens a interesting situation; when the correlation coefficient is not very large the simplified tests perform better than the imputed tests, the percentage of rejections are higher. However when $\rho_{(X^c,X^m)}$ is high the imputed versions perform better because the covariate X^c provides information sufficient to impute suitably. The Table 6 shows, for $\rho_{(X^c,X^m)} = 0.8$, the behavior of the tests for different sample sizes and $\alpha = 1\%, 5\%, 10\%$. It should be noted that the model used for the simulation of covariates X^m (7), implies that the correlation coefficient influence not only in the case of imputation, but also in the case of complete data. $(D_{n,C}, W_{n,C})$.

	a=0			a=0.05			a=0.1		
$\rho_{(X,Z)}$	0.64	0.80	0.98	0.64	0.80	0.98	0.64	0.80	0.98
$D_{n,C}$	4.2	5.2	4.5	4.6	6.3	96.0	6.0	10.9	99.9
$D_{n,S}$	6.1	4.1	4.5	6.1	6.5	88.4	7.2	10.7	97.8
$D_{n,I_{true}}$	4.6	4.1	5.0	5.0	5.3	93.7	5.6	9.20	99.9
$D_{n,I_{ls}}$	5.0	4.2	5.3	4.9	4.9	93.7	5.6	8.8	99.9
$D_{n,I_{np}}$	5.7	3.3	5.3	5.3	5.2	91.6	5.5	8.8	99.8
$W_{n,C}$	5.8	5.6	5.9	6.4	7.8	99.6	8.5	16.9	100
$W_{n,S}$	6.5	4.8	6.1	7.2	8.7	97.9	10.4	14.9	100
$W_{n,I_{true}}$	6.4	5.9	5.5	6.7	7.4	99.1	8.5	14.1	100
$W_{n,I_{ls}}$	6.3	6.8	5.5	7.0	7.5	98.9	8.2	14.6	100
$W_{n,I_{np}}$	7.7	6.0	6.4	7.5	7.2	97.4	8.7	11.9	100

Table 5: Percentage of times that H_0 is rejected with $n = 50$, $a = 0, 0.05, 0.1$ and $\alpha = 5\%$.

	n=50			n=100		
$\rho_{(X,Z)} = 0.80$	1%	5%	10%	1%	5%	10%
$D_{n,C}$	0.5	5.2	10.9	0.4	4.8	8.9
$D_{n,S}$	0.9	4.1	11.7	0.7	3.9	9.8
$D_{n,I_{true}}$	0.8	4.1	9.9	0.3	4.6	9.3
$D_{n,I_{ls}}$	0.4	4.2	10.8	0.3	4.4	9.9
$D_{n,I_{np}}$	0.4	3.3	10.7	0.5	5.5	10.4
$W_{n,C}$	0.3	5.6	11.7	1.0	4.2	9.8
$W_{n,S}$	0.7	4.8	11.4	0.8	4.9	11.3
$W_{n,I_{true}}$	0.7	5.9	12.2	0.9	3.8	9.7
$W_{n,I_{ls}}$	0.6	6.8	11.2	0.9	4.6	9.9
$W_{n,I_{np}}$	0.9	6.0	11.9	1.0	5.0	11.9

Table 6: Percentage of times that H_0 is rejected with $a = 0$, $\rho_{(X,Z)} = 0.80$, $n = 50, 100$ and $\alpha = 1\%, 5\%, 10\%$.

5. Application to Real data

5.1 Missing Response

To illustrate a situation with missing data in the response variable, we consider the Wind Energy data. The data set consists of 219 daily observations of daily Electric Power (MW in Megawatts) and daily average of Wind Velocity (Wv in m/s). We are interested to check the linear model of the Electric Power versus Wind Velocity. There are 8 observations with missing data in the Electric Power variable (about 3.8%). The Figure 1 is a plot of daily electric power against daily average wind velocity with a estimates from a linear regression on the complete cases. A simple regression model is of interest with missing response data, thus the null hypothesis considered in this case is:

$$H_0 : E\,(MW|Wv) = \beta_0 + \beta_1 Wv$$

for some β_0 and β_1. That is, we want to check whether the model is linear or not.

We applied our test statistics as in the simulation study to these data. All p-values are less than 0.0001 except for simplified process ($R_{n,S,R}$) with 4% and 1.44% for KS and CvM, respectively. In Figure 2 shows the approximate bootstrap density of Kolmogorov-Smirnov statistics test.

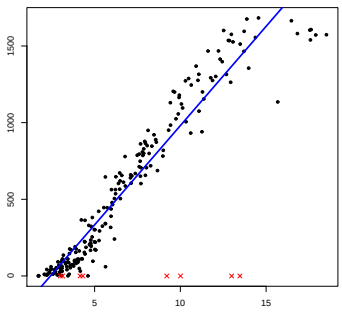


Figure 1: Daily electric power (MW) againts daily average wind velocity (m/s). The solid straight line is the linear-regression fit to the complete case analysis. Solid circles are the 211 complete data cases and points with $\times$ correspond to missing data in response variable.

5.2 Missing Covariate

We apply the statistical test to check whether relationship between annual Salary (S in €) and the covariables Duration of employment (D in days) and Age (A in years) is linear (Data are provided by the Galician Institute of Statistics (IGE) and obtained from the "Muestra continua

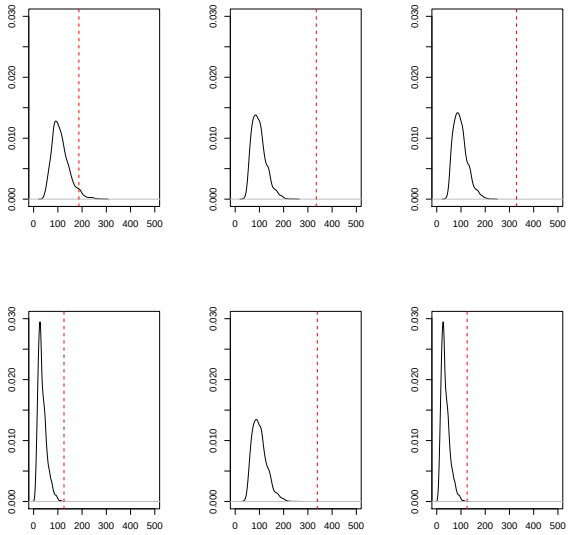


Figure 2: Bootstrap distribution of statistics test based on Kolmogorov-Smirnov. From top to bottom and left to right $D_{n,S,R}, D_{n,S}, D_{n,I,r}, D_{n,I}, D_{n,A}, D_{n,Sm}$.

de vidas laborales (MCVL) 2011" elaborated by the Ministry of Employment and Social Security in Spain):

$$H_0 : E\left[S/A, D\right] = \beta_0 + \beta_1 A + \beta_2 D$$

In this case, the missing observations are in covariate D and they are imputed under a linear model with missing observations in the response. A total of 37405 observations with 3409 (9.11%) missing in the covariate D. The linear correlation between D and A is 0.377. The tests statistics based on the sub-complete sample and in the least squares imputed sample were computed. The results have been obtained using the binning techniques. The linear binning has been used with $(40, 25)$ bins on each dimension.

All bootstrap p-values obtained with $B = 1000$, are less than 0.0001. Therefore, in all cases the null hypothesis of linearity can be rejected.

References

González-Manteiga W, Pérez-González A (2006). Goodness-of-fit tests for linear regression models with missing response data. Can J Stat; 34(1):149-170.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

ANÁLISIS ESTADÍSTICO DE LA EVALUACIÓN SENSORIAL DE UN MEDICAMENTO MEDIANTE UN PANEL DE CATADORES

Maruxa García Quiroga[1], M. Nieves Muñoz Ferreiro[12], M. Ángeles Romero Rodríguez[13] y M. Lourdes Vázquez Ordériz[13]

[1] TASTELAB (Spin-off de la USC, Consultoría de análisis sensorial para la industria alimentaria)
[2] Departamento de Estadística e Investigación Operativa (Universidad de Santiago de Compostela)
[3] Áreas de Tecnología de Alimentos y de Nutrición y Bromatología del Departamento de Química Analítica, Nutrición y Bromatología (Universidad de Santiago de Compostela)

RESUMEN

En este trabajo se realiza la comparación de un total de 9 formulaciones de un medicamento a partir de su evaluación sensorial. Una de las formulaciones es un producto farmacéutico de uso comercial y el resto son modificaciones en los compuestos responsables del aroma-sabor (bien en cantidad o tipo). El objetivo es comparar las 9 formulaciones entre sí mediante pruebas sensoriales con catadores entrenados.

La variable de interés es, inicialmente, el nivel de intensidad del sabor amargo. Partiendo de un grupo de 20 catadores semi-entrenados se realiza una selección mediante pruebas de umbrales de percepción del sabor amargo (cafeína). Tras la selección, resulta un panel de 12 catadores que realizan un re-entrenamiento específico en la detección del sabor amargo. Los catadores deben cumplimentar una ficha de cata para caracterizar el nivel de sabor amargo en el momento de probar el producto (tiempo 0), después de 10 segundos tras introducir la muestra en boca (persistencia) y el nivel de sabor dulce de cada una de las formulaciones. Por el número de formulaciones y por tratarse de productos con un elevado y persistente sabor amargo, se emplea un diseño en bloques incompletos balanceado (BIB) evitando así que cada catador pruebe todas las formulaciones.

El análisis de la varianza muestra diferencias significativas entre las 9 formulaciones en las tres características estudiadas. Las comparaciones dos a dos mediante la prueba de Tukey reflejan que en el sabor amargo en tiempo 0 y en el sabor dulce, sólo el producto de uso comercial tiene diferencias significativas con el resto de formulaciones. En la persistencia del sabor amargo a los 10 segundos, sólo una de las formulaciones con cambios en el sistema aromático presenta diferencias estadísticamente significativas con la formulación de partida y tampoco en esta variable hay diferencias significativas entre las 8 formulaciones restantes. Sin embargo, desde el punto de vista multivariante, el test T^2 de Hotteling ayuda a seleccionar las formulaciones que se diferencian en mayor medida del resto. Se comprueba la estabilidad de esta conclusión haciendo un análisis cluster a partir de las valoraciones medias del grupo de catadores en los descriptores evaluados.

Palabras y frases clave: Diseño en bloques incompleto. Prueba T^2 de Hotteling. Análisis cluster.

REFERENCIAS

Cochran. W. G. and Cox, G.M. (1957): Experimental Designs, John Wiley & Sons.

Francois Husson, Sebastien Le and Marine Cadoret (2014). SensoMineR: Sensory data analysis with R. R package version 1.20. http://CRAN.R-project.org/package=SensoMineR

Meilgaard, M. C. et all (2007): Sensory Evaluation Techniques. CRC Press.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

ESTADÍSTICAS SOBRE PRODUCTOS GALLEGOS CON DOP E IGP: CONOCIMIENTO, CONSUMO Y PERCEPCIÓN

Marta Vázquez Pita[1], Jorge García Luque[1], M. Nieves Muñoz Ferreiro[24], M. Ángeles Romero Rodríguez[34] y M. Lourdes Vázquez Ordériz[34]

[1] Estudiante del Máster en Innovación en Seguridad y Tecnologías Alimentarias (Universidad de Santiago de Compostela)
[2] Departamento de Estadística e Investigación Operativa (Universidad de Santiago de Compostela)
[3] Áreas de Tecnología de Alimentos y de Nutrición y Bromatología del Departamento de Química Analítica, Nutrición y Bromatología (Universidad de Santiago de Compostela)
[4] TASTELAB (Spin-off de la USC-Consultoría de análisis sensorial para la industria alimentaria)

RESUMEN

Las Denominaciones de Calidad, entre las que se incluyen Denominaciones Geográficas como las Denominaciones de Origen Protegidas (DOP) y las Indicaciones Geográficas Protegidas (IGP), se pueden considerar esquemas de calidad de tipo público que las administraciones otorgan a productos tradicionales que, bien por la naturaleza u origen de las materias utilizadas, bien por los métodos de producción empleados, alcanzan una especial notoriedad o prestigio en el mercado. Desde el punto de vista de la política agraria comunitaria, con las Denominaciones Geográficas se pretende fomentar la diversificación de la producción agrícola y promocionar productos que presenten características diferenciadas para así beneficiar al mundo rural. Son productos con ventajas organolépticas, de retorno social, económico e incluso medioambiental respecto a otros productos sin vinculación al origen.

En estos productos el nombre de la denominación, a modo de marca colectiva, tiene como objetivo facilitar su identificación por los consumidores. Por ello, en este estudio, mediante encuestas personales, se recopiló información para valorar la percepción y el grado de conocimiento y consumo de los productos gallegos con DOP/IGP.

Los encuestados que participan perciben los productos gallegos con DOP/IGP como productos de buena calidad y de procedencia garantizada aunque caros, dando gran importancia a su origen gallego. El grado de conocimiento y el consumo de los 32 productos alimenticios gallegos con Denominación Geográfica es muy variable entre productos y en algunos casos está relacionado con factores sociodemográficos. También se observa que el etiquetado con símbolos diferenciadores juega un papel fundamental a la hora de identificar productos con Denominación Geográfica.

Palabras y frases clave: Denominaciones de Origen Protegidas (DOP). Indicaciones Geográficas Protegidas (IGP). Muestreo por cuotas. Análisis de correspondencias.

REFERENCIAS

Richard M. Heiberger, Naomi B. Robbins (2014). Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications. Journal of Statistical Software, 57(5), 1-32. http://www.jstatsoft.org/v57/i05/.

Francois Husson, Julie Josse, Sebastien Le and Jeremy Mazet (2015). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.31.3. http://CRAN.R-project.org/package=FactoMineR

MAGRAMA (2015). Ministerio De Agricultura, Alimentación y Medioambiente. Denominaciones de Origen e Indicaciones Geográficas Protegidas. http://goo.gl/5hWtCi.

ENCUESTA SOBRE COMPRA Y CONSUMO DE PRODUCTOS PESQUEROS FRESCOS Y CONGELADOS EN LUGO: ANÁLISIS DE RESULTADOS MEDIANTE MÉTODOS ESTADÍSTICOS EXACTOS

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 y Bromatología y de Tecnología de Alimentos, Grupo GASFQ, Dpto. de Química Analítica, Nutrición e Bromatoloxía, Facultade de Ciencias, Universidade de Santiago de Compostela.
[2] Grupo MODESTYA, Dpto. de Estatística e Investigación Operativa, Facultade de Ciencias, Universidade de Santiago de Compostela.

RESUMEN

Se presenta un estudio de consumidores basado en la entrevista personal realizado en la ciudad de Lugo. En primer lugar, se muestra un análisis exploratorio de los hábitos de compra y consumo de productos pesqueros frescos y congelados por parte del consumidor. En segundo lugar, se examina la posible influencia en los hábitos de compra y consumo de productos pesqueros frescos y congelados de ciertos factores de interés, demográficos - sexo, edad, nivel de estudios y situación laboral- y sobre la situación socioeconómica del hogar - número de personas, personas con ingresos y nivel de ingresos. Para examinar la posible asociación de un factor demográfico o socioeconómico de interés con un hábito de compra o consumo, se presentan los resultados del correspondiente contraste de independencia basados en los tests sugeridos en Lydersen y col. (2007): el test exacto de Fisher- Freeman-Halton, el test exacto basado en el estadístico chi cuadrado de Pearson y el test exacto basado en el estadístico de razón de verosimilitudes. En aquellos casos en los que los tests de independencia previos permiten concluir que hay evidencias de una relación entre el hábito y el factor, para determinados pares (i,k) de niveles del factor, resulta de interés comparar la probabilidad de tener el hábito en el grupo i con la probabilidad de tener el hábito en el grupo k. La comparación se lleva a cabo mediante los intervalos de confianza para las diferencias de proporciones basados en el método exacto de Wang (2010). Para los análisis estadísticos que se presentan se utilizan el paquete estadístico IBM SPSS Statistics (Mehta, C. R. y Patel, N. R. (1996)) y el entorno estadístico R (Shan, G. y Wang, W. (2013)).

Palabras y frases clave: test exacto de asociación; diferencia de proporciones; intervalos de confianza exactos; estudios de consumidores; hábitos de compra de pescado.

AGRADECIMIENTOS

Este trabajo fue parcialmente financiado mediante el proyecto de investigación de referencia 10TAL010CT de la Xunta de Galicia.

REFERENCIAS

Lydersen, S., Pradhan, V., Senchaudhuri, P. y Laake, P. (2007). Choice of test for association in small sample unordered r x c tables. *Statistics in Medicine*, 26, 4328-4343.

Mehta, C. R. y Patel, N. R. (1996). *SPSS Exact Tests™ 7.0 for Windows*. SPSS Inc.

Shan, G. y Wang, W. (2013). ExactCIdiff: An R Package for Computing Exact Confidence Intervals for the Difference of Two Proportions. *The R Journal*, 5/2 December, 62-70.

Wang, W (2010). On construction of the smallest one-sided confidence interval for the difference of two proportions. *Annals of Statistics*, 38, No. 2, 1227-1243.

Aplicación del análisis factorial múltiple para la estimación del riesgo de incendios en Galicia

M. Gema Maestro Requena[1], Manuel Francisco Marey Pérez[1]

[1] Grupo de investigación GI-1716 PROePLA, Universidad de Santiago de Compostela, España

RESUMEN

Este trabajo se enfoca ante la necesidad de poseer herramientas sólidas y robustas diseñadas para cada unidad de gestión (nivel regional) y a una escala de trabajo que permita detectar las variables de mayor riesgo como causa de estos incendios y, en consecuencia, poder tomar decisiones enfocadas a orientar cualquier decisión en la prevención de los incendios forestales. Los resultados muestran que un análisis factorial múltiple ayudaría, para el caso de estudio (comunidad de Galicia), a clasificar una serie de variables de distinta naturaleza (fisiográficas, socio-económicas, de infraestructuras y ambientales) consideradas importantes en el riesgo de incendio, pudiendo así optimizar recursos y focalizar soluciones a problemas concretos.

Palabras e frases chave: Incendio forestal, análisis factorial múltiple, "clusterización", gestión, Galicia

1. INTRODUCIÓN

En los últimos años el problema generado por los incendios forestales ha aumentado (Plucinski et al. 2013), del mismo modo recientes estudios arrojan contundentes datos de superficie quemada, estimándose que cada año arden 3,5 millones de km2 (Giglio et al., 2010), el cambio climático parece incidir directamente sobre la ocurrencia de incendios forestales como ya han documentado varios autores (Kasischke y Turetsky 2006), al tratarse de un fenómeno complejo, los incendios forestales dependen directamente de multitud de factores (Moreira et al., 2011) pero a diferencia de otros peligros naturales como terremotos o huracanes son uno de los más predecibles, dejando a las sociedades modernas algún margen de maniobra para implantar estrategias y contrarrestarlos (Birot et al., 2009). Todas estas premisas nos llevan a destacar la gran importancia que tiene el establecimiento de la elaboración de escenarios futuros de las condiciones de riesgo de incendio.

Una de las regiones más afectadas del mundo es la Europa Mediterránea (Rodrigues et al., 2013; Fuentes-Santos et al., 2013, Rodrigues et al., 2014a y 2014b) y España está incluida dentro de los cinco países del sur de Europa con más incendios forestales con medias anuales de 19.705 incendios y 130.714 hectáreas quemadas (Barreal el al., 2012). Dentro del total Español, es destacable el caso de Galicia, que con el 6 % de la superficie nacional representa alrededor del 45 % del total de incendios ocurridos en España, suponiendo estos el 23 % de la superficie afectada (Barreal el al., 2012). La actual evolución es dispar debido a la situación social de las zonas agroforestales del territorio (Marey et al., 2007). De cualquier forma existen causas análogas en el entorno europeo: el aumento de la superficie forestal (Catry et al. 2009); superposición entre las casas y los bosques (Lampin-Maillet et al., 2010); nuevos patrones de paisaje agrario (Catry et al. 2009; Martínez et al., 2009; Moreira et al., 2009, 2011; Verde and Zêzere, 2010); cambios en los patrones de paisajes agrarios tradicionales (Ruiz-Mirazo et al., 2012); abandono de la actividad agraria (Aranzabal et al., 2008); la expansión de la superficie de matorral (Koutsias et al., 2009; Montané et al. 2009; Moreira et al., 2009); cambios socioeconómicos (Romero-Calcerrada et al., 2010; Vilar et al., 2010); reacciones post fuego (Conedera et al., 2011) y cambios en la distribución de la renta y niveles de desempleo más altos (González-Olabarria et al., 2011).

Al ritmo que crecía la problemática de los incendios se han incrementado las herramientas y los instrumentos para estudiarlo. Se han considerado diferentes enfoques para resolver y explicar las dificultades de modelar la causa humana de ocurrencia los incendios forestales en Europa con un enfoque espacial. Algunos estudios utilizan modelos de regresión logística para para predecir la probabilidad de ocurrencia de incendio forestal (Catry et al., 2009; Chuvieco et al., 2010); otros se decantan por la utilización de redes neuronales artificiales (Gomez et al., 2011, Ruiz-Mirazo et al., 2012, Bisquert et al., 2012); pesos de la evidencia (Romero-Calcerrada et al., 2008 y 2010); estadisticos multivariantes (Herrero-Corral et al., 2012. Verdu et al., 2012); regresiones lineales generalizadas (Oliveira et al., 2013); distribuciones generalizadas de Pareto (Zea-Bermudez et al., 2009); lógica difusa (Iliadis et al., 2010); modelos bayesianos (Verde and Zezere, 2010), modelos de Poisson (Martinez et al, 2009); modelos weibull (Curt et al., 2013); modelos aditivos generalizados (Vilar et al., 2010); metodos kernel y no-parametricos (Kuter et al., 2012, Gonzalez-Olabarria et al., 2012) y simulaciones de Montecarlo (Conedera et al., 2011, Moreira et al., 2011).

El presente trabajo se centra en agrupar las variables espaciales de diferente naturaleza, utilizadas en trabajos anteriores, con el fin de concretar "causas" fisiográficas, socio-económicas, de infraestructuras y ambientales de incendios forestales en la comunidad de Galicia.

2. MATERIAL Y MÉTODOS

Área de estudio

Los incendios en Galicia son un problema recurrente que en las últimas décadas ha incrementado su seberidad (Chas-Amil et al. 2010). Desde 1961 hasta diciembre de 2013 se han producido un total de 251.106 incendios (Ministry of Agriculture, 2014). Se ha quemado un área total de 1829330 ha, lo que equivale al 65,4% del área geográfica de Galicia y aproximadamente el 128,5% del área forestal arbolada de la comunidad. De acuerdo con Catry et al. (2009) y al igual que en otras regiones, el aumento del número de encendidos de fuego se produce a pesar de la asignación de más recursos para la vigilancia, lucha y prevención de incendios, incluidos los planes de gestión, campañas de educación pública y la aplicación de la legislación más restrictiva en relación con las actividades humanas susceptibles de provocar incendios forestales. Los incendios forestales afectan principalmente a los municipios rurales situados en el sur de la región que tienen densidades de población bajas y las dinámicas demográficas regresivas (Balsa-Barreiro y Hermosilla, 2013; Fuentes-Santos et al. 2013). Estos municipios también se han visto afectadas por los patrones de inmigración extranjera recientes, que, combinado con el éxodo rural del siglo pasado, ha dado lugar a fuertes caídas en la población. Además, las estructuras económicas se basan en los sectores primarios (Gonzalez et al., 2007).

V**ariables**

Las variables utilizadas para este análisis pueden enmarcarse en cuatro grupos, según su naturaleza: (i) Fisiográficas, superficie del territorio visible desde puntos de vigilancia, (ii) Socioeconómicas, población, evolución de la población en los últimos 10 años, producción agraria, superficie de matorral, superficie forestal arbolada, (iii) Ambientales, superficie protegida por criterios ambientales e (iv) Infraestructuras, densidad de carreteras. Del mismo modo que la variable dependiente, se han agregado en pixeles de 1 km^2. La descripción de las variables y sus unidades se recogen en la tabla 1.

Tipo	**Variable**	**Descripción**	**Unidades**	**Fuente**
Fisiográficas	VISISON_TOR	Superficie visible del terreno desde cada punto de vigilancia fijo (torres o casetas)	celdas visibles/km^2	Elaboración propia
Socio-económicas	EV_POP_99_	Evolución de la población en el periodo 1999-2010	habitantes/municipio	INE
	POP_99_10	Número de habitantes en 2010	habitantes/municipio	INE
	DISTANCIA_	Distancia al núcleo de población más próximo	m	Elaboración propia
	PET_2009	Producción estándar total en el año 2009	euros totales/municipio	INE
Infraestructuras	CARRETERAS	Densidad de carreteras	m/km^2	Elaboración propia-NAFTEC
Ambientales	SUP_MATORR	Superficie de matorral	m^2/km^2	IV IFN
	SUP_FOREST	Superficie forestal arbolada	m^2/km^2	IV IFN
	SUP_PROTEG	Superficie protegida	m^2/km^2	Ministerio de Medio Ambiente

Tabla 1.- Variables consideradas para la elaboración del modelo de regresión

Se dispone del valor de cada una de las variables mencionadas por cada celda de 1 km^2 de la superficie de Galicia que haya sufrido uno o más incendios forestales en el periodo 2003-2008.

Análisis multivariante de ordenación

Se realizó un Análisis Factorial Múltiple (AFM), (Abdi et al., 2013) para sintetizar en un número reducido de dimensiones la variabilidad elegidas en trabajos anteriores, descritas en el apartado anterior, para modelizar el riesgo de incendio para toda la superficie de la comunidad de Galicia. El AFM es una extensión del Análisis de Componentes Principales (ACP) que permite analizar variables estructuradas en grupos diferentes, considerando explícitamente los grupos de pertenencia a la hora de producir los resultados finales, además de poder utilizar conjuntamente variables categóricas y continuas. Se utilizó un total de 57347 registros (uno por cada celda de 1 km2 de un mapa raster que cubre la región de Galicia) para realizar el análisis, que correspondían al total de los datos disponibles. No fue necesaria la elección de una muestra, ya que la capacidad computacional del equipo utilizado permitió el cálculo. Los valores de cada variable se normalizaron inicialmente (Busqué et al., 2014). Se escogieron las primeras dimensiones ortogonales del AFM cuya suma de valores propios representara el 63,98% de la suma total de todos los valores propios.

Con las puntuaciones de las observaciones de la muestra en las primeras dimensiones ortogonales elegidas, se realizó una clasificación no supervisada por particiones de las celdas mediante el algoritmo de k-medias (Legendre y Legendre, 2012). Este algoritmo busca, mediante un método iterativo, los centroides de los k grupos que se pretende definir, y asigna cada observación al grupo con el centroide más próximo. De cara a probar particiones con distinto número de grupos, se utilizó la función *cascadeKM* del paquete de R *vegan*, probando particiones para formar desde dos hasta 15 grupos. Se utilizó el índice de Calinski-Harabasz (Legendre y Legendre, 2012) como uno de los criterios para elegir la partición más adecuada. Todos estos análisis se realizaron con el programa estadístico R especialmente con funciones de sus librerías *FactoMineR* (Husson et al., 2015) y *vegan* (Oksanen et al., 2015).

3. RESULTADOS

Las tres primeras dimensiones del AFM recogieron el 63,98% de la varianza total. La primera dimensión supuso el 35,30%, la segunda el 15,65%, y la tercera el 13,03% de la varianza total. En la primera dimensión se alcanzaron puntuaciones más elevadas para las variables de superficie forestal y de matorral (SUP_FOREST y SUP_MATORR) y para la densidad de carreteras (CARRETERAS). Por el contrario, la menor puntuación se obtuvo para la distancia al núcleo de población más cercano (DISTANCIA). En la segunda dimensión la mayor puntuación fue para densidad de carreteras (CARRETERAS) y la menor para la visibilidad desde puntos de vigilancia fijos (VISIÓN_TOR). Por el contrario, en la tercera dimensión la mayor puntuación la obtiene la visibilidad (VISIÓN_TOR). La menor puntuación es para la superficie protegida (SUP_PROTEG) (Tabla 2).

Variables	Dimensión 1	Dimensión 2	Dimensión 3
VISION_TOR	0.5095585	-0.57148849	0.6329930
EV_POB_99_	-0.5754582	0.29773051	0.3687022
POB_99_10	0.1137860	0.43109762	0.2470304
DISTANCIA_	-0.7645169	-0.16079700	0.1048261
PET_2009	0.5333973	0.16215295	0.1292675
CARRETERAS	0.7396480	0.55067578	0.1350637
SUP_PROTEG	0.3040116	-0.35470451	-0.5295749
SUP_FOREST	0.7037674	0.05012485	-0.2265534
SUP_MATORR	0.7104703	-0.31601085	-0.3064266

Tabla 2: Puntuaciones de cada variable para cada una de las tres dimensiones de análisis

Respecto a la estratificación, el índice de Calinski-Harabasz señaló como más adecuadas las particiones generando de 7 a 15 grupos. El valor de los centroides de cada uno de los siete estratos en las tres primeras dimensiones del AFM mostró una diferenciación clara en este espacio tridimensional (Figura 2).

De la agrupación final de las celdas de 1 km^2 en los 7 grupos mencionados se interpreta que el primer grupo (grupo 1) contiene los incendios caracterizados por una evolución baja de la población y una distancia a carreteras elevada.

El grupo 2 presenta una evolución de población muy baja, pero si que se centra en los mayores núcleos urbanos (se observan las menores superficies forestales y de matorral y superficie protegida) (Figura 1). Por el contrario, el tercer grupo, presenta bajos valores de población pero una elevada densidad de carreteras. El grupo 5 está formado por las superficies mejor vigiladas desde puntos de vigilancia fijos y presenta elevados valores en cuanto a superficie forestal y alta densidad de carreteras. El grupo 6 engloba grandes superficies de matorral y elevadas superficies visibles (Figura 2), pero valores medios para el resto de variables. Por último, en el grupo 7 es en el destacan los elevados valores de superficies protegidas (altos también para superficies forestales y de matorral), y baja productividad y evolución de la población.

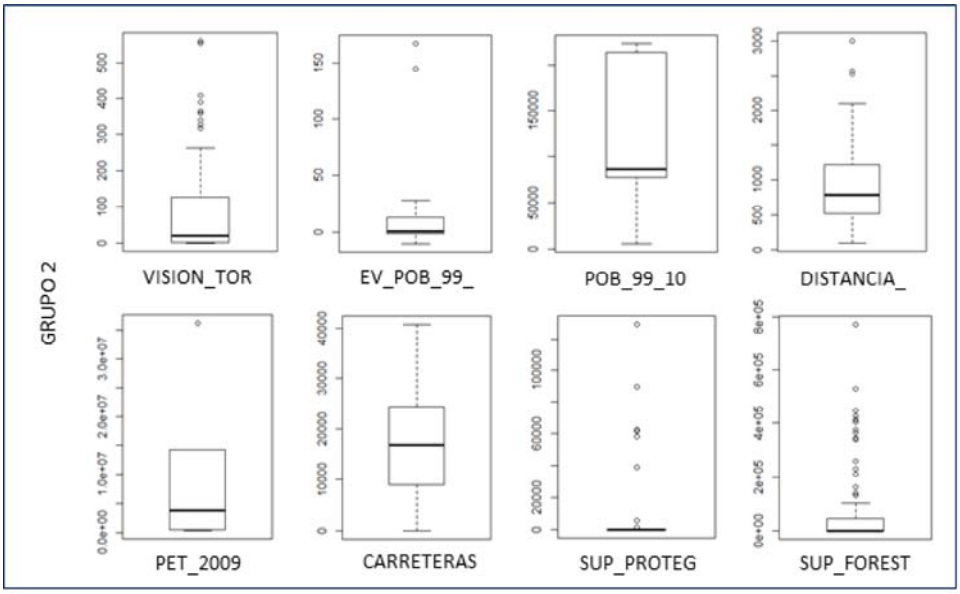


Figura 1.- Ejemplo: distribución de las variables explicativas utilizadas en el trabajo en el grupo 2

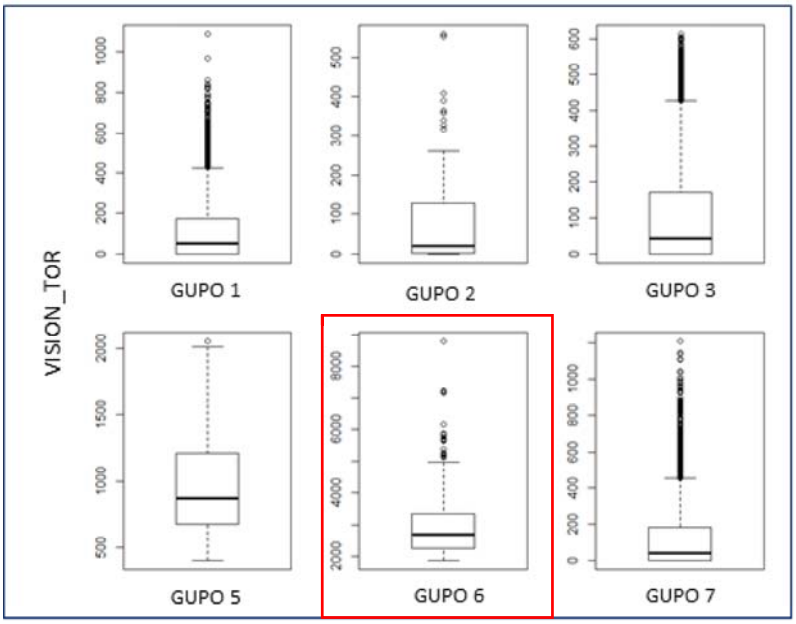


Figura 2.- Ejemplo: Las mayores superficies visibles se engloban en el grupo 6

Cabe mencionar que el grupo 4 engloba todas las “esquinas” del raster que cubre el territorio regional, por lo tanto contenían valores 0 para todas y cada una de las variables utilizadas en el análisis.

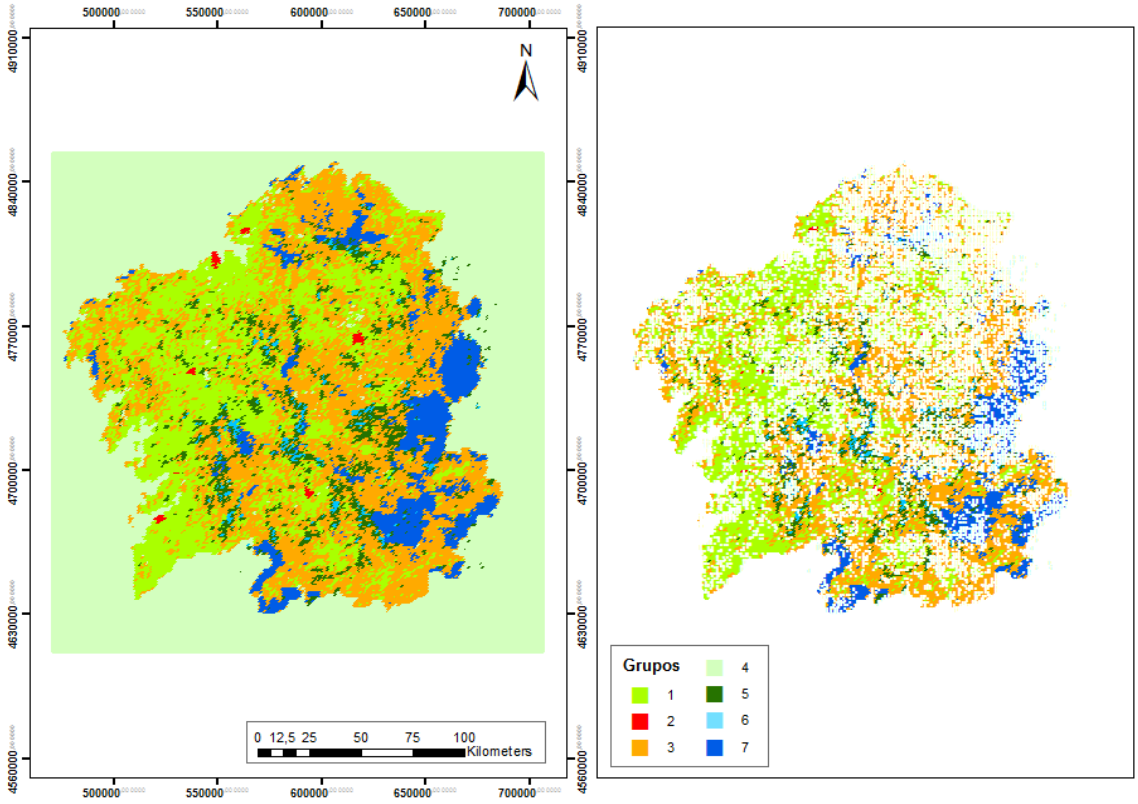


Figura 2.- A la izquierda, la superficie regional clasificada en los grupos resultado del análisis. A la derecha, clasificación de la superficie de Galicia afectada por algún incendio en el periodo 2003-2008

4. CONCLUSIÓNS

En cada grupo se establecerán unas medidas de gestión específicas; por ejemplo el grupo 6, que presenta una elevada carga de combustible forestal, requeriría de una serie de medidas enfocadas a tratamientos en la masa para reducir carga o eliminar su continuidad, el grupo 2, medidas enfocadas a la interfaz, bien a la distribución y carga del combustible alrededor de la zona urbanizada o bien a la disposición de la vivienda respecto al entorno y al fuego, o el grupo 3, con una densidad elevada de carreteras, sería objeto del diseño de medidas especiales de vigilancia y control en épocas de elevado riesgo de incendio. Para los grupos que presentan valores extremos en las variables socioeconómicas podrían requerir medidas de tipo educativo, informativo u otro tipo de campañas de sensibilización y concienciación.

Por lo tanto, la "clusterización" de variables fisiográficas, socio-económicas, ambientales y de infraestructuras para la identificación de características de incendios en la comunidad de Galicia define con claridad 6 grupos, permitiendo enfocar recursos administrativos y económicos en el diseño y planificación de soluciones concretas a cada casuística.

Sería interesante implementar la metodología llevada a cabo en este trabajo, mediante técnicas de análisis estadístico y sistemas de información geográfica que originen mapas temáticos, y elegir nuevas variables causantes de incendios forestales que agrupadas por temáticas, podrían ser estudiadas en detalle para cada medida de gestión.

REFERENCIAS

Abdi, H., Williams, L.J. and Valentin, D. (2013) Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Computational Statistics, 5, 149-179.

Aranzabal, I., Schimitz, M.F., Aauilera, P. and Pineda, F.D. (2008) Modelling of landscape changes derived from the dynamics of socio-ecological systems: a case of study in a semiarid Mediterranean landscape. Ecological Indicators, 8, 672-685.

Balsa-Barreiro, J. B., and Hermosilla, T. (2013). Socio-geographic analysis of the causes of the 2006's wildfires in Galicia (Spain). Forest Systems, 22(3), 497-509.

Barreal, J., Loureiro, M., Picos, J. (2012) Estudio de la causalidad de los incendios forestales en Galicia. Economía Agraria y Recursos Naturales. 12, 1: 99-114.

Barreal, J., Loureiro, M.L. and Picos, J. 2014. On insurance as a tool for securing forest restoration after wildfires. Forest Policy and Economics, 42, 15–23

Birot, Y., Borgniet, L., Camina, A., Dupuy, J.L., Fernandes, P., Goldammer, J.G., Gonzalez-Olabarria, J.R., Jappiot, M., Lampin-Maillet, C., Mavsar, R., Montiel-Molina, C., Moreira, F., Moreno, J.M., Rego, F., Rigolot, E., San-Miguel, J., Vallejo, R., Velez, R. (2012) Convivir con los incendios forestales: Lo que nos revela la ciencia. European Forest Institute. Discussion Paper 15.

Bisquert, M., Caselles, E.,Sanchez, J.M. and Caselles, V. 2012. Application of artificial neural networks and logistic regression to the prediction of forest fire danger in Galicia using MODIS data. International Journal of Wildland Fire, 21, 1025–1029.

Busqué, J., Maestro, M.G. y Jiménez, S. (2014) Estratificación ambiental de Cantabria: metodología, resultados y aplicaciones de interés pascícola. Actas de la 53ª reunión de la SEEP. Cantabaria. 53,113-120.

Catry, F.X., Rego, F.C. and Baçao, F. (2009) Modeling and mapping wildfire ignition risk in Portugal. International Journal of Wildland Fire 2009, 18, 921–931

Chas-Amil, M., Touza, J. and Prestemon, J.P. (2010) Spatial distribution of human-caused forest fires in Galicia (NW Spain). Ecology and the Environment, Vol 137, 247 - 258

Chuvieco, E., Aguado, I., Yebra, M., Nieto, H., Salas, J., Martín, M. P. and Zamora, R. (2010). Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecological Modelling, 221(1), 46-58.

Conedera, M., Torriani, D., Neff, C., Ricotta, C., Bajocco, S., and Pezzatti, G. B. (2011) Using Monte Carlo simulations to estimate relative fire ignition danger in a low-to-medium fire-prone region. Forest Ecology and Management, 261(12), 2179-2187.

Curt, T., Borgniet, L. and Bouillon, C. (2013) Wildfire frequency varies with the size and shape of fuel types in southeastern France: Implications for environmental management. ScienceDirect. Journal of Environmental Management 117, 150-161.

Jari Oksanen, J., Blanchet, F.G. Kindt R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson G. L., Solymos, P., Stevens, M. H. H. and Wagner, H. (2015) vegan: Community Ecology Package. R package version 2.3-0. http://CRAN.R-project.org/package=vegan

Husson, F., Josse, J., Le, S. and Mazet, J. (2015) FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.31.3. http://CRAN.R-project.org/package=FactoMineR

Fuentes-Santos, I., Marey-Pérez, M. F., and González-Manteiga, W. (2013) Forest fire spatial pattern analysis in Galicia (NW Spain). Journal of environmental management, 128, 30-42.

Giglio, L., Randerson, J. T., Van der Werf, G. R., Kasibhatla, P. S., Collatz, G. J., Morton, D. C., and DeFries, R. S. (2010) Assessing variability and long-term trends in burned area by merging multiple satellite fire products.Biogeosciences, 7(3).

Gomez, I., Pilar-Martin, M. (2011) Prototyping an artificial neural network for burned area mapping on a regional scale in Mediterranean areas using MODIS images. International Journal of Applied Earth, 13, 741-752.

Gómez-Vázquez, I., Álvarez-Álvarez, P., and Marey-Pérez, M. F. (2009) Conflicts as enhancers or barriers to the management of privately owned common land: A method to analyze the role of conflicts on a regional basis.Forest Policy and Economics, 11(8), 617-627.

González-Olabarria, J. R., and Pukkala, T. (2011) Integrating fire risk considerations in landscape-level forest planning. Forest Ecology and Management, 261(2), 278-287.

González, J. R., and Pukkala, T. (2007) Characterization of forest fires in Catalonia (north-east Spain). European Journal of Forest Research, 126(3), 421-429.

Herrero-Corral, G., Jappiot, M., Bouillon, C., Long-Fournel, M. (2012) Application of a geographical assessment method for the characterization of wildland–urban interfaces in the context of wildfire prevention: A case study in western Madrid. Applied Geography 35, 60-70.

Iliadis, L.S., Vangeloudh, M., Spartalis, S. (2010) An intelligent system employing an enhanced fuzzy c-means clustering model: Application in the case of forest fires. Computers and Electronics in Agriculture 70, 276-284.

Kasischke, E. S., and Turetsky, M. R. (2006) Recent changes in the fire regime across the North American boreal region—spatial and temporal patterns of burning across Canada and Alaska. Geophysical Research Letters, 33(9).

Koutsias, N., Mallinis, G., and Karteris, M. (2009) A forward/backward principal component analysis of Landsat-7 ETM+ data to enhance the spectral signal of burnt surfaces. ISPRS Journal of Photogrammetry and Remote Sensing, 64(1), 37-46.

Kuter, S., Usul, N., Kuter, N. (2012) Bandwidth determination for kernel density analysis of wildfire events at forest sub-district scale. Ecological Modelling 222, 3033-3040.

Lampin-Maillet, C., Mantzavelas, A., Galiana, L., Jappiot, M., Long, M., Herrero, G. and Thanassis, P. (2010) Wildland urban interfaces, fire behaviour and vulnerability: characterization, mapping and assessment.Towards Integrated Fire Management-Outcomes of the European Project Fire Paradox, Lampin-Maillet, C., 21-p.

Legendre, P. Y Legendre, L. (2012) Numerical Ecology. Third English Edition. Elsevier. Amsterdam. 990p.

Martínez, J., Vega-Garcia, C., and Chuvieco, E. (2009) Human-caused wildfire risk rating for prevention planning in Spain. Journal of Environmental Management, 90(2), 1241-1252.

Ministerio de Agricultura, Alimentacion y Medio Ambiente. (2014) [en linea]. Disponible en http://www.magrama.gob.es/es/

Molinero, F., Cascos, C., Garcia, A., and Baraja, E. (2008) Dynamics of forest fires in Castilla y León (Spain). Boletín de la AGE, 48, 397-401.

Montané, F., Casals, P., Taull, M., Lambert, B., and Dale, M. R. (2009) Spatial patterns of shrub cover after different fire disturbances in the Pyrenees. Annals of Forest Science, 66(6), 1-8.

Moreira, F., Vaz, P., Catry, F., and Silva, J. S. (2009) Regional variations in wildfire susceptibility of land cover types in Portugal: implications for landscape management to minimize fire hazard. International Journal of Wildland Fire, 18(5), 563-574

Moreira, F., Viedma, O., Arianoutsou, M., Curt, T., Koutsias, N., Rigolot, E., Barbati, A., Corona, P., Vaz, P., Xanthopoulus, G., Mouillot, F. and Bilgili, E. (2011) Landscape-wildfire interactions in southern Europe: Implications for landscape management. Journal of environmental Management 92 (2011), 2389-2402.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., Pereira, J. (2012) Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest.Forest Ecology and Management Volume 275, 1 July 2012, 117–129.

Plucinski, M. P. (2013) Modelling the probability of Australian grassfires escaping initial attack to aid deployment decisions. International journal of wildland fire, 22(4), 459-468.

Rodrigues, M., San Miguel, J., Oliveira, S., Moreira, F. and Camia A. (2013) An insight into spatial-temporal trends of fire ignitions an burned areas in the European Mediterranean countries. Journal of Earth Science an Engineering. 3 (2013), 497-505

Rodrigues, M., and de la Riva, J. (2014) An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environmental Modelling & Software, 57, 192-201.

Rodrigues, M., de la Riva, J., and Fotheringham, S. (2014) Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. Applied Geography, 48, 52-63.

Romero-Calcerrada, R., Novillo, C.J., Millington, J.D.A., and Gomez-Jimenez, I. (2008) GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain). Landscape Ecology 23(3), 341-354.

Romero-Calcerrada, R., Barrio-Parra, F., Millington, J. D. A., and Novillo, C. J. (2010) Spatial modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain).Ecological Modelling, 221(1), 34-45.

Ruiz-Mirazo, J., and Robles, A. B. (2012) Impact of targeted sheep grazing on herbage and holm oak saplings in a silvopastoral wildfire prevention system in south-eastern Spain. Agroforestry systems, 86(3), 477-491.

Verde, J. C., and Zêzere, J. L. (2010) Assessment and validation of wildfire susceptibility and hazard in Portugal. Natural Hazards and Earth System Science, 10(3), 485-497.

Verdu, F., Salas, J., Vega-Garcia, C. (2012) Amultivariate analysis of biophysical factors and forest fires in Spain, 1991–2005. International Journal of Wildland Fire 21, 498–509.

Vilar, L., Woolford, D., Martell, D.L. and Martín, M.P. (2010) A model for predicting human-caused wilfire occurrence in region of Madrid, Spain. International Journal of Wildland Fire 19 (2010): 325-337

Zea-Bermudez, P., Mendes, J., Pereira, J.M.C., Turkman, K.F. and Vasconcelos, M.J.P. (2009) Spatial and temporal extremes of wildfire sizes in Portugal (1984–2004). International Journal of Wildland Fire 18, 83–991.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Spatial pattern analysis of wildfires in Galicia.

F. Manuel Marey-Pérez[1], Wenceslao González-Manteiga[1], Paula Saavedra-Nieves[1], Isabel Fuentes-Santos[2]

[1]University of Santiago de Compostela, Spain.

[2]Instituto de Investigaciones Marinas, CSIC, Spain.

ABSTRACT

Wildfires are the major cause of disturbance and change in Galician forests. Therefore, studying their behaviour is indispensable in forest management in order to establish enhanced fire prevention and fire fighting plans. Our aim is to analyze the evolution of the spatial distributions of arson and non-arson fires separately for a sequence of ten years from 1999 to 2008. From the corresponding ignition points coordinates, the first order intensity was estimated and the contour maps were represented for extracting some interesting conclusions.

Keywords: Forest management, intensity function, kernel estimator, bandwidth matrix.

INTRODUCTION

Forest fire is considered one of the most important disturbance factors in natural ecosystems which directly depends on social behaviour, whether voluntary or involuntary. In fact, in many populated regions such as Galicia in the North West of Spain, most of fires are caused by arsonists, see Figure 1. Therefore, studying the behaviour of arson fires is key in forest management in order to prevent or reduce their incidence and to mount significant prevention efforts and to design specific fire prevention campaigns.

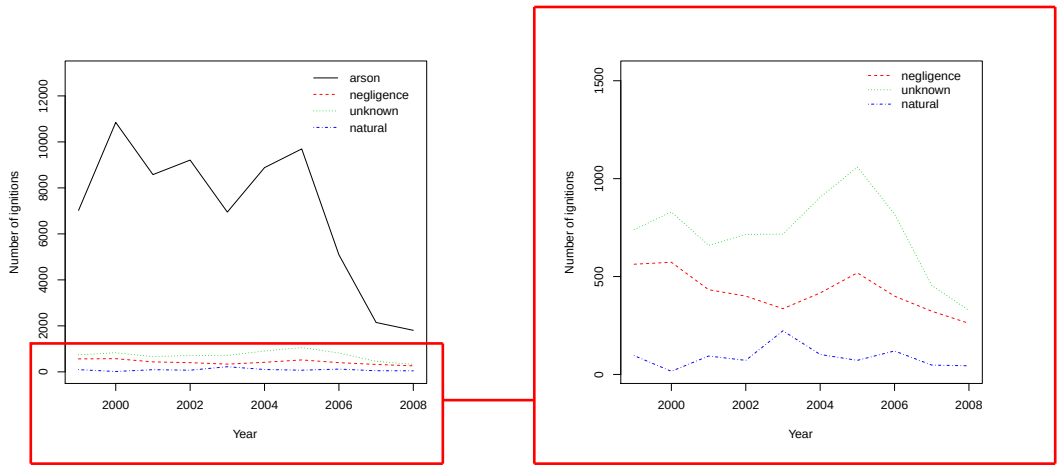


Figure 1: Evolution of the numbers of ignitions depending on their cause (arson, negligence, unknown or natural) during the period 1999-2008 in Galicia.

In this work, the forest fires of Galicia will be analyzed using a real data set that contains the longitudinal and latitudinal coordinates of ignition from 1999 to 2008. Therefore, this wildfire data set can be seen as a realization of a spatial point process. The main goal here is to study the spatial distribution of arson and non-arson forest fires separately estimating the first order intensity of the ignition points for each one of the years considered. The notation established of non-arson fires includes the set of fires whose cause is a negligence, unknown or natural. The non-parametric kernel estimator for the intensity function proposed by Diggle (1985) is considered selecting the bandwidth matrix according to the plug-in alternative in Fuentes-Santos et al. (2015). The corresponding intensity and contour maps have been derived for comparing the evolution of arson and non-arson fires for the period of time established. The results obtained show, besides the different behaviour of arson and non-arson fires and the risk areas, that 2006 could be considered as a change-point in the trend of arson fires.

REFERENCES

Diggle P.J. (1985) A kernel method for smoothing point process data. Applied Statistics 34, 138-147.

Fuentes-Santos I., González-Manteiga W., Mateu J. (2015). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial Poisson point patterns. (submitted)

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

PESQUISA E VALIDAÇÃO DO MPP GLOBAL NA ANÁLISE DA FIABILIDADE ESTRUTURAL DOS COMPÓSITOS

Luísa N. Hoffbauer[1] e Carlos C. António[2]

[1]Instituto Superior de Engenharia do Porto (ISEP), Instituto Politécnico do Porto
[2]Faculdade de Engenharia (FEUP), Universidade do Porto

RESUMO

Um dos principais problemas na análise da fiabilidade estrutural dos compósitos laminados é a possível existência de múltiplos MPP (*Most Probable failure Point*) ou pontos de falha mais provável das funções de estado limite. O problema é similar ao da existência de vários mínimos locais em otimização de estruturas. Embora o projeto ótimo de estruturas baseado num ótimo local seja um procedimento aceitável, as implicações da existência de mínimos locais em análise de fiabilidade podem ser catastróficas. Com efeito um MPP local pode não corresponder à situação mais desfavorável em termos de probabilidade de falha da estrutura. Por esta razão a pesquisa de múltiplos MPP é necessária e importante na análise da fiabilidade estrutural, uma vez que apenas o MPP global representa a fiabilidade da estrutura. Neste trabalho propõe-se uma técnica numérica inversa baseada no cálculo de sensibilidades e no método das bissecções sucessivas para a pesquisa do MPP global em função da anisotropia dos compósitos laminados. É utilizado um método Bayesiano na estimação da probabilidade de falha usando simulação de Monte Carlo para validar os resultados obtidos através da metodologia inversa proposta. O processo de validação demonstra que a metodologia proposta é adequada para estimar a probabilidade de falha das estruturas compósitas laminadas.

Palavras e frases chave: Estruturas compósitas, Probabilidade de falha, Aproximação inversa, Método de Lind Hasofer, Simulação de Monte Carlo.

1. INTRODUÇÃO

A análise mais realista de falhas de estruturas sob incertezas está associada ao uso de métodos probabilísticos. A análise probabilística da integridade estrutural, ou fiabilidade estrutural, toma em consideração as incertezas existentes nas variáveis e parâmetros de projeto de estruturas em engenharia e estuda o impacto que estas incertezas têm na caracterização da falha destes sistemas físicos. Os métodos de fiabilidade aproximados tais como os métodos de primeira ordem (FORM) ou de segunda ordem (SORM) visam obter o chamado ponto de falha mais provável (MPP) para estimar a probabilidade de falha. A pesquisa do MPP é equivalente à resolução de um problema de otimização que tem sido efectuada, até recentemente, usando técnicas baseadas em gradientes. Quando os métodos utilizados se baseiam em gradientes, é-se confrontado com a possibilidade de existência de múltiplos pontos de falha, ou seja, múltiplos óptimos locais. Neste caso é necessária a utilização de técnicas de simulação, como por exemplo, o Método de Monte Carlo.

2. PROBLEMA FUNDAMENTAL DA FIABILIDADE ESTRUTURAL

O problema fundamental da fiabilidade estrutural consiste no cálculo do seguinte integral

$$p_f = P(G(\mathbf{X}) \leq 0) = \int_{G(\mathbf{x})\leq 0} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{1}$$

em que $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ é um vetor aleatório definido no espaço de probabilidade $(\Omega, \mathcal{F}, \mathcal{P})$ onde Ω é o espaço amostral, podendo ser identificado como o conjunto de todos os cenários de projecto estrutural, $\mathcal{F}$ é a σ-álgebra dos subconjuntos de Ω e P é a medida de probabilidade. O vector $\mathbf{X}$ representa as variáveis aleatórias que contêm informação sobre as incertezas que afetam a segurança da estrutura em consideração e $f_{\mathbf{X}}(\mathbf{x})$ é a função densidade de probabilidade conjunta de $\mathbf{X}$. As variáveis básicas geralmente descrevem a aleatoriedade na geometria, nas propriedades dos materiais ou nas cargas aplicadas.

Para cada realização de $\mathbf{X}$, determina-se o estado da estrutura através de quantidades, tais como deslocamentos, deformações, tensões, medidas de dano, etc. O estado de uma estrutura pode ser de falha ou de segurança. O estado de falha acontece quando uma realização de $\mathbf{X}$ se encontra no domínio de falha definido por $D_f = \{\mathbf{x} \in \Re^n : G(\mathbf{x}) \leq 0\}$. A função $G(\mathbf{X}, \Psi(\mathbf{X}))$ é designada por função de estado limite e depende das variáveis de entrada, $\mathbf{X}$, e da resposta, $\Psi(\mathbf{X})$. Por uma questão de simplificação, só vai ser usada a notação $\mathbf{X}$ para referir todas as variáveis aleatórias envolvidas na análise. O conjunto $\{\mathbf{x} : G(\mathbf{x}) = 0\}$ define a superficie estado limite, representando o limite entre os dois estados.

Consideram-se então dois conjuntos:

- $D_f = \{\mathbf{x} \in \Re^n : G(\mathbf{x}) \leq 0\}$ que se designa por domínio de falha;
- $D_s = \{\mathbf{x} \in \Re^n : G(\mathbf{x}) > 0\}$ que se designa por domínio de segurança.

O estado de segurança é o estado para o qual uma estrutura, ou parte dela, é capaz de cumprir todas as funções para as quais foi concebida. Por outro lado, o estado limite é visto como aquele estado para além do qual uma estrutura, ou parte dela, não está mais no estado de segurança.

Na análise da fiabilidade de estruturas compósitas, as variáveis aleatórias que se assumem ser não correlacionadas, definem o vector $\mathbf{X}$.

Nesta aplicação, as variáveis aleatórias são as propriedades mecânicas dos compósitos laminados da estrutura e a função de estado limite pode ser escrita como

$$G(\mathbf{x}) = \bar{R} - 1 \tag{2}$$

onde $\bar{R}$ é o número de Tsai crítico definido por

$$\bar{R} = \min(R_1, \cdots, R_k, \cdots, R_{Ns}) \tag{3}$$

sendo N_s o número total de pontos onde o vector tensão é calculado. O número de Tsai, $\bar{R}$, definido como sendo a razão entre a tensão máxima admissível e a tensão aplicada, é obtido do critério de falha interactivo de Tsai-Wu calculado no k-ésimo ponto da estrutura, resolvendo a equação

$$1 - (F_{ij}\sigma_i\sigma_j) R_k^2 + (F_i\sigma_i) R_k = 0 \quad i, j = 1, 2, 6 \tag{4}$$

onde σ_i são as componentes do vector tensão e F_{ij} e F_i são os parâmetros de resistência associados com os laminados reforçados unidireccionais definidos de um ponto de vista macro-mecânico.

Sendo conhecida a distribuição das variáveis básicas, X_i, e a superfície estado limite, $G(\mathbf{x})$, a probabilidade de falha pode ser usada como uma medida de fiabilidade. Utilizando o método de Lind Hasofer vai ser obtido um valor aproximado para esta probabilidade.

3. MÉTODO DE LIND HASOFER

Em 1974 Hasofer e Lind apresentaram um método assumindo implicitamente a normalidade e a independência das variáveis de entrada associadas à aleatoriedade ou incerteza do problema.

O problema originalmente definido no espaço normal $\mathbf{x}$ é transformado num problema definido no espaço normal padronizado $\mathbf{u}$ de coordenadas $u_1, u_2, \cdots, u_n$ através da transformação

$$U_i = \frac{X_i - \mu_i}{\sigma_i} \tag{5}$$

Sejam $U_1, U_2, \ldots, U_n$ variáveis aleatórias normais padronizadas independentes.

Consideremos a função de estado limite projectada neste espaço. Seja agora
$M = g(U_1, U_2, \ldots, U_n)$.

A equação (1) pode ser reescrita na forma

$$p_f = \int_{g(\mathbf{u})\leq 0} \varphi(\mathbf{u})\, d\mathbf{u} \tag{6}$$

onde $\varphi(\mathbf{u})$ é a função densidade de probabilidade da multinormal padronizada. Esta função tem um máximo na origem e decresce exponencialmente com $\|u\|_2^2$. Então, os pontos que mais contribuem para o integral anterior são aqueles do domínio de falha que estão mais próximos da origem.

Assim, o primeiro passo consiste na determinação do ponto do domínio de falha mais próximo da origem no espaço normal padronizado.

No caso mais simples em que a função de estado limite é linear, esta função pode ser escrita na forma $g(\mathbf{u}) = \beta - \alpha^T\mathbf{u}$, onde β é a distância do hiperplano $g(\mathbf{u}) = 0$ à origem e α é um vector unitário normal ao hiperplano e direccionado para fora. Prova-se que $\varphi(\mathbf{u})$ tem um máximo em $\mathbf{u}^* = \beta\alpha$. Por isso, $\mathbf{u}^*$ é designado por ponto de falha mais provável (*Most Probable failure Point*, MPP). Este ponto da superfície de estado limite, que fica à menor distância da origem, representa a pior combinação das variáveis aleatórias.

A variável aleatória $M = \beta_{LH} - \alpha^T\mathbf{U}$ segue distribuição normal de valor médio $E(M) = \beta$ e variância $\mathrm{V(M)} = 1$. A probabilidade de falha é dada por

$$p_f = P\left(\beta - \alpha^T\mathbf{U} \leq 0\right) = \Phi(-\beta) \tag{7}$$

No caso em que $g(\mathbf{u})$ não é linear, o cálculo do ponto de falha mais provável (MPP) é um problema de otimização com restrições

$$\mathbf{u}^* = \min_{g(\mathbf{u})=0} \|\mathbf{u}\|_2 \tag{8}$$

Este problema pode ser resolvido utilizando métodos apropriados de otimização. O método clássico do cálculo variacional consiste em resolver o problema equivalente

$$\min \mathrm{L}(\mathbf{u}, \lambda) = \|\mathbf{u}\|_2 + \lambda g(\mathbf{u}) \tag{9}$$

em que λ é o multiplicador de Lagrange associado à equação de estado limite $g(u) = 0$.

Se o número de variáveis básicas é elevado ou se as equações estado limite são complicadas, é preferivel optar por um esquema iterativo para resolver este problema.

Após a determinação do MPP, o correspondente índice de fiabilidade é definido como

$$|\beta| = \|\mathbf{u}^*\|_2 \tag{10}$$

O segundo passo consiste em aproximar a superficie estado limite.

No FORM considera-se a função de estado limite linearizada no MPP, $\mathbf{u}^*$. Tal linearização corresponde a aproximar a superficie de falha $g(\mathbf{u}) = 0$ pelo hiperplano normal ao vector $\mathbf{u}^*$ que define o MPP. A equação deste hiperplano pode ser escrita como $\beta - \alpha^T\mathbf{u}$ onde o vector unitário α é dado por $\alpha = \frac{\nabla g|_{\mathbf{u}^*}}{\|\nabla g|_{\mathbf{u}^*}\|_2}$. O estudo é então análogo ao caso em que a função de estado limite é linear. A probabilidade de falha é dada pela equação (7):

$$p_f = \int_{g(\mathbf{u})\leq 0} \varphi(\mathbf{u}) d\mathbf{u} \approx p_{fFORM} = \int_{\beta-\alpha^T\mathbf{u}\leq 0} \varphi(\mathbf{u}) d\mathbf{u} = \Phi(-\beta) \tag{11}$$

A aproximação de primeira ordem dá bons resultados se as curvaturas da superfície de estado limite não forem elevadas. Para o caso da superfície de estado limite ser fortemente não linear, foram desenvolvidas aproximações de segunda ordem para melhorar os resultados. A aproximação de segunda ordem da superfície de estado limite obtém-se desenvolvendo $g(\mathbf{u})$ em polinómio de Taylor de segunda ordem em torno de u^*, ou seja

$$g(\mathbf{u}) \approx \nabla g|_{\mathbf{u}^*}(\mathbf{u} - \mathbf{u}^*) + \frac{1}{2}(\mathbf{u} - \mathbf{u}^*)^T \nabla^2 g\big|_{\mathbf{u}^*}(\mathbf{u} - \mathbf{u}^*) \tag{12}$$

onde $\nabla^2 g = \left[\frac{\partial^2 g}{\partial u_i \partial u_j}\right]$ é o Hessiano da função de estado limite.

Se as variáveis de entrada não forem normais ou não forem independentes, Rackwitz e Fiessler sugerem que o problema originalmente definido no espaço-x seja transformado num problema definido no espaço-u onde as variáveis aleatórias de entrada são normais padronizadas e independentes. A transformação do espaço-x para o espaço-u depende da informação estatística disponível, se as variáveis de entrada são normais ou não e se são independentes ou não.

4. SIMULAÇÃO DE MONTE CARLO (FORMULAÇÃO BAYESIANA) NA FIABILIDADE

Para estimar a probabilidade de falha, os métodos aproximados anteriores usam o chamado ponto de falha mais provável (MPP). A pesquisa deste ponto é um problema de otimização que envolve técnicas baseadas em gradientes, pesquisa evolucionária ou outras. Quando os métodos utilizados se baseiam em gradientes, é-se confrontado com a possibilidade de existência de múltiplos pontos de falha, ou seja, múltiplos óptimos locais. Este problema é exacerbado quando o número de variáveis aleatórias envolvidas é grande ou quando o grau de não linearidade das funções de estado limite associadas à resposta do sistema estrutural é elevado. Neste caso, não há garantia de que a fiabilidade estrutural tenha sido correctamente avaliada. De forma a ultrapassar estas possiveis dificuldades, são propostas aproximações complementares como, por exemplo, o método de Monte Carlo. Estas aproximações complementares são usadas essencialmente para fornecer valores de referência com vista à validação dos resultados obtidos pelos métodos anteriores.

Segundo a formulação Bayesiana, na estimação de um parâmetro θ, a distribuição *a posteriori* $h_{\theta|\mathbf{x}}(\theta|\mathbf{x})$ é o produto da verosimilhança da amostra $f_{\mathbf{x}|\theta}(\mathbf{x}|\theta)$ pela distribuição *a priori* $g_\theta(\theta)$ a menos de uma constante. Para o caso particular em que o parâmetro a estimar é a probabilidade de falha p_f, pelo Teorema de Bayes, a distribuição *a posteriori* de p_f, $h_{p_f|k,N}(p_f|k,N)$ é dada por

$$h_{p_f|k,N}(p_f|k,N) = \frac{f_{k,N|p_f}(k,N|p_f)\, g_{p_f}(p_f)}{\int_0^1 f_{k,N|p_f}(k,N|p_f)\, g_{p_f}(p_f)\, dp_f} \tag{13}$$

onde $g_{p_f}(p_f)$ é a distribuição *a priori* e deve pertencer a uma família tal que a distribuição *a posteriori* $h_{\theta|\mathbf{x}}(\theta|\mathbf{x})$ seja da mesma família. Para a verosimilhança da amostra, tem-se o modelo binomial

$$f_{k,N|p_f}(k,N|p_f) = \binom{N}{k} (p_f)^k (1-p_f)^{N-k} \quad k = 0,1,2,\ldots N \tag{14}$$

donde se deduz que a distribuição *a priori* deve ser uma distribuição Beta.

A distribuição *a posteriori* para uma verosimilhança binomial e uma *a priori* Beta é

$$h_{p_f|k,N}(p_f|k,N) = \begin{cases} \frac{1}{B(a+k,N-k+b)} (p_f)^{a+k+1} (1-p_f)^{N-k+b-1} & 0 \le p_f \le 1 \\ 0 & \text{outros valores} \end{cases} \tag{15}$$

que é uma distribuição Beta $(\mathrm{k}+\mathrm{a}, \mathrm{N}-\mathrm{k}+\mathrm{b})$ com valor médio

$$E(p_f|k,N) = \frac{k+a}{N+a-b} \tag{16}$$

e variância

$$V(p_f|k,N) = \frac{(a+k)(b+N-k)}{(a+b+N)^2 (a+b+N+1)} \tag{17}$$

Um intervalo de confiança, com coeficiente de confiança γ é dado por (p_{fmin}, p_{fmax}) com p_{fmin} tal que

$$\int_0^{p_{fmin}} \mathrm{Beta}(\mathrm{k}+\mathrm{a}, \mathrm{N}-\mathrm{k}+\mathrm{b})\, \mathrm{dp_f} = \frac{\gamma}{2} \tag{18}$$

e p_{fmax} é tal que

$$\int_0^{p_{fmax}} \mathrm{Beta}(\mathrm{k}+\mathrm{a}, \mathrm{N}-\mathrm{k}+\mathrm{b})\, \mathrm{dp_f} = 1 - \frac{\gamma}{2} \tag{19}$$

Vão ser aqui consideradas duas situações:

- $g_{p_f}(p_f)$ é não informativa

  Neste caso, o modelo adequado para $g_{pf}(p_f)$ segue uma distribuição $U(0,1)$, ou seja, $p_f \sim \text{Beta}(1,1)$. A distribuição *a posteriori* é uma distribuição $\text{Beta}(\text{k}+1, \text{N}-\text{k}+1)$. Um intervalo de confiança, com coeficiente de confiança γ é dado por (p_{fmin}, p_{fmax}) com p_{fmin} tal que

$$\int_0^{p_{fmin}} \text{Beta}(\text{k}+1, \text{N}-\text{k}+1)\,\text{dp}_\text{f} = \frac{\gamma}{2} \tag{20}$$

  e p_{fmax} é tal que

$$\int_0^{p_{fmax}} \text{Beta}(\text{k}+1, \text{N}-\text{k}+1)\,\text{dp}_\text{f} = 1 - \frac{\gamma}{2} \tag{21}$$

- $g_{p_f}(p_f)$ é informativa

  Se a fonte de informação for o conhecimento de uma simulação de Monte Carlo anterior (N' simulações e k' falhas), a distribuição *a priori* é uma distribuição $\text{Beta}(\text{k}'+1, \text{N}'-\text{k}'+1)$. A distribuição *a posteriori* é uma distribuição
  $\text{Beta}(\text{k}+\text{k}'+1, \text{N}-\text{k}+\text{N}'-\text{k}'+1)$. Um intervalo de confiança, com coeficiente de confiança γ é dado por (p_{fmin}, p_{fmax}) com p_{fmin} tal que

$$\int_0^{p_{fmin}} \text{Beta}(\text{k}+\text{k}'+1, \text{N}-\text{k}+\text{N}'-\text{k}'+1)\,\text{dp}_\text{f} = \frac{\gamma}{2} \tag{22}$$

  e p_{fmax} é tal que

$$\int_0^{p_{fmax}} \text{Beta}(\text{k}+\text{k}'+1, \text{N}-\text{k}+\text{N}'-\text{k}'+1)\,\text{dp}_\text{f} = 1 - \frac{\gamma}{2} \tag{23}$$

5. PROBLEMA DE FIABILIDADE INVERSO

É proposta uma abordagem baseada no projecto ótimo de estruturas compósitas com o objetivo de atingir um nível de fiabilidade especificado e é obtida a correspondente carga máxima. A função objetivo descrevendo o desempenho é definida como o quadrado da diferença entre o índice de fiabilidade estrutural, β_S, e o índice de fiabilidade prescrito, β_a. As variáveis de projeto são o ângulo da camada, a, e o fator carga, λ, estando ambos agregados no vetor $\mathbf{x}_d$. O vetor das variáveis aleatórias, $\mathbf{X}$, é constituído pelas propriedades elásticas e de resistência do material.

Assim, a função que descreve o desempenho do sistema estrutural é

$$F(\lambda, a, \mathbf{x}) = [\beta_S(\lambda, a, \mathbf{x}) - \beta_a]^2 \tag{24}$$

onde $\mathbf{x}$ representa uma realização das variáveis aleatórias. O vetor das cargas aplicadas é definido do seguinte modo: $\text{L} = \lambda \text{L}^{\text{ref}}$, onde L^{ref} é o vetor carga de referência. A minimização do funcional definido na equação (24) corresponde a um problema de fiabilidade estrutural inverso. Com efeito, para cada ângulo da camada, haverá um fator de carga ótimo λ^* associado a um índice de de fiabilidade estrutural prescrito β_a. Este fator de carga é estimado usando o método das bissecções sucessivas e o cálculo do β_S utilizando o Método de Lind Hasofer.

O índice de fiabilidade prescrito, β_a, é assim o objetivo a ser atingido pelo índice de fiabilidade estrutural.

6. APLICAÇÕES A ESTRUTURAS COMPÓSITAS

Para testar a abordagem proposta, a estrutura considerada é uma casca cilíndrica encastrada em compósito laminado representada na Figura 1. Ao longo do lado livre da estrutura (AB), é aplicado um conjunto de nove cargas verticais. A estrutura está dividida em quatro macro-elementos agrupando todos os elementos e cada macro-elemento é constituído por um único laminado. A distribuição dos laminados na estrutura está representada na Figura 1. A análise estrutural é baseada no modelo de Elementos Finitos de casca desenvolvido por Ahmad, com desenvolvimentos posteriores por Figueiras. Este elemento casca é obtido de um elemento tridimensional utilizando um processo degenerativo. É um elemento isoparamétrico com 8 nós e 5 graus de liberdade por nó

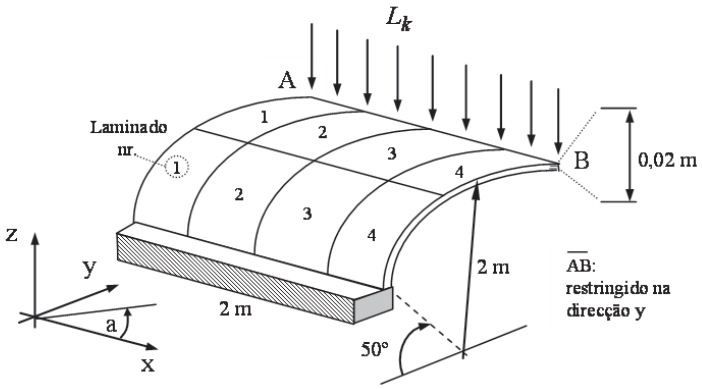


Figura 1: Casca cilíndrica e distribuição dos compósitos laminados

- três translações e duas rotações, baseado na teoria de cascas de Mindlin. São consideradas como variáveis aleatórias as propriedades mecânicas representadas por $\mathbf{X}_p$. Todas as variáveis aleatórias são não correlacionadas e seguem distribuições normais definidas pelo respectivos valor médio e desvio padrão. O grupo das propriedades mecânicas $\mathbf{X}_p$ inclui as seguintes dezasseis variáveis aleatórias: o módulo elástico longitudinal $\mathbf{E}_{1j}$, o módulo elástico transversal $\mathbf{E}_{2j}$, a resistência transversal em tracção $\mathbf{Y}_j$ e a resistência ao corte, $\mathbf{S}_j$, onde o índice $j = 1, \cdots, 4$ denota o número do laminado.

$a\,[^\circ]$	$\bar{L}\,(\beta_a)\,[N]$
0	3629
5	3645
10	3734
15	3918
20	4236
25	4739
30	5490
35	6601
40	8041
45	9815
50	12072
55	14756
60	17852
65	21320
70	25104
75	32725
80	33839
85	34704
90	35241

Tabela 1: Desempenho do sistema estrutural medido pela carga máxima para $\beta_a = 3$.

As camadas unidimensionais da casca cilíndrica representada na Figura 1 são constituídas pelo sistema compósito Vidro E/Resina de Epóxido (Scotchply 1002®). Os laminados são constituídos por oito camadas de igual espessura com sequência de empilhamento $[+a/+a/-a/-a]_S$ impondo condições de simetria. O ângulo da camada a refere-se ao eixo dos xx. Todos os laminados têm a

mesma espessura média igual a 0.02m.

O problema de fiabilidade inverso na equação (24) é resolvido e é obtida a carga máxima admissível $\bar{L}(\beta_a)$, pela estrutura compósita, considerando um índice de fiabilidade imposto de $\beta_a = 3$. Esta carga máxima, em função da variável de projecto ângulo da camada a, está apresentada na Tabela 1 para o coeficiente de variação do grupo das variáveis de entrada que representam as propriedades do material $\mathbf{X}_p$ igual a 6%.

Os resultados representados na Tabela 1 podem ser validados usando métodos da Estatística Bayesiana aplicados ao problema inverso, isto é, calculando a probabilidade de falha, p_f, associada com a carga máxima obtida, $\bar{L}(\beta_a)$, para cada variável ângulo da camada, a.

O objectivo dos métodos Bayesianos é o uso de toda a informação disponível (informação *a priori*) baseada na experiência anterior, intuição ou resultados obtidos em problemas anteriores similares. Neste caso, a fonte de informação é o conhecimento de uma simulação prévia de Monte Carlo ($N' = 2000$ simulações e k' falhas). Um intervalo de confiança para p_f com coeficiente de confiança γ é dado pelas equações (22) e (23) sendo $N = 5000$ e k o número de falhas na segunda simulação de Monte Carlo.

$a\,[^\circ]$	k'	k	p_{fmin}	p_{fmax}
0	0	8	0.00059	0.00225
5	1	8	0.00069	0.00244
10	2	8	0.00078	0.00263
15	2	8	0.00078	0.00263
20	2	7	0.00069	0.00244
25	2	8	0.00078	0.00263
30	3	7	0.00078	0.00263
35	3	8	0.00089	0.00281
40	4	7	0.00089	0.00281
45	5	8	0.00109	0.00317
50	5	7	0.00099	0.00299
55	5	7	0.00099	0.00299
60	5	7	0.00099	0.00299
65	3	8	0.00089	0.00281
70	3	8	0.00089	0.00281
75	3	10	0.00109	0.00317
80	1	8	0.00069	0.00244
85	1	9	0.00078	0.00263
90	1	6	0.00049	0.00206

Tabela 2: Intervalos de confiança para a probabilidade de falha do sistema estrutural

Neste estudo, a carga máxima é aplicada à estrutura compósita para cada valor do ângulo da camada, a, e é usada a simulação de Monte Carlo juntamente com metodologia Bayesiana, para confirmar o índice de fiabilidade fixado, $\beta_a = 3$. Na Tabela 2 são apresentados os intervalos de confiança (95%) para p_f. A probabilidade de falha $p_f = \Phi(-3) = 0.00135$, onde Φ representa a função distribuição da normal padronizada está dentro do intervalo de confiança, mostrando a concordância entre os resultados obtidos usando o método de Lind Hasofer e os obtidos usando simulação de Monte Carlo juntamete com metodologia bayesiana.

7. CONCLUSÃO

Dependendo de um determinado índice de fiabilidade pré-fixado imposto às estruturas compósitas é obtida a carga máxima admissível pela estrutura em função da orientação das camadas resolvendo o problema inverso associado.

Os resultados obtidos para a carga máxima, $\bar{L}(\beta_a)$, foram validados usando o método de Monte Carlo juntamente com metodologia bayesiana na estimação da probabilidade de falha.

A validação dos resultados, demonstra que a metodologia proposta é apropriada para aferir a probabilidade de falha de estruturas compósitas.

REFERÊNCIAS

António, C. and Hoffbauer, L.(2009) An approach for reliability-based robust design optimisation of angle-ply composites. Composite Structures, 90, 53–59.

Der Kiureghian, A. (2005) First and Second-Order Reliability Methods. Engineering Design Reliability Handbook. CRC Press.

Guérin, F., Barreau, M., Charki, A., Todoskoff, A. (2007) Bayesian estimation of failure probability in mechanical systems using Monte Carlo simulation. Quality Technology Quantitative Management, 4, 51–70.

Hasofer, A.M. and Lind N.C. (1974) Exact and Invariant Second-Moment Code Format. Journal of the Engineering Mechanics Division, 100, 111–121.

Melchers, R.E. (1999) Structural Reliability Analysis and Prediction. Wiley.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

DISEÑO DE UNA TÉCNICA DE MONITORIZACIÓN

Miguel E. Vázquez-Méndez[1], Lino J. Alvarez-Vázquez[2], Gerardo Casal[1] y Aurea Martínez[2]

[1]Departamento de Matemática Aplicada, Universidade de Santiago de Compostela. Escola Politécnica Superior, R/ Benigno Ledo, 27002 Lugo (Spain).

[2]Departamento de Matemática Aplicada II. EI Telecomunicaciones. Universidad de Vigo. 36310 Vigo (Spain).

RESUMEN

En este trabajo abordamos el diseño de una técnica de monitorización para un indicador conocido definido en un determinado intervalo (presión arterial de un paciente a lo largo del tiempo, concentración de un contaminante a lo largo de un río, etc.). El objetivo es localizar los puntos en los que debemos medir el indicador (puntos de muestreo), para que esas medidas proporcionen una visión global del comportamiento del indicador en todo el intervalo. Formulamos el problema como un problema de optimización multiobjetivo con variables enteras y variables reales y, asumiendo ciertas preferencias en la toma de decisiones, proponemos un algoritmo completo para obtener una solución que proporcione una técnica de monitorización *satisfactoria*. Finalmente, ilustramos la bondad del método propuesto, aplicándolo a dos ejemplos sencillos.

Palabras y frases clave: Técnicas de Monitorización. Optimización Multiobjetivo.

1. INTRODUCCIÓN

Entendemos por monitorización la medición sistemática y planificada de los indicadores de calidad de un proceso bajo estudio. Un aspecto importante en el diseño de técnicas de monitorización es la ubicación, espacial o temporal, de los puntos en los que se miden esos indicadores, puntos de muestreo (Gilbert, 1987). Por ejemplo, si el indicador que se quiere medir es una función ρ definida sobre un intervalo $[a, b]$, ya sea temporal (presión arterial o concentración de un medicamento en la sangre de un paciente) o espacial (concentración de un contaminante en un río), los puntos de muestreo deben ofrecer medidas representativas de lo que ocurre en todo el intervalo. En ciertas ocasiones no se tiene información a priori sobre la función ρ y la elección de esos puntos de muestreo debe hacerse utilizando técnicas que combinan aspectos aleatorios con diversas estrategias diseñadas por el experimentador. Esas técnicas (técnicas de muestreo) buscan alcanzar la información necesaria para poder inferir el comportamiento del indicador en todo el intervalo (ver, por ejemplo, Thompson, 2012). En otras situaciones, la función ρ puede estimarse utilizando modelos experimentales y/o numéricos suficientemente contrastados (en muchos casos obtenidos por inferencia estadística a partir de la información obtenida sobre ella). En estas situaciones, cuando se dispone de una estimación fiable del indicador ρ, las mediciones que se van a realizar en los puntos de muestreo servirán como medidas de control de calidad o diagnóstico y, en ese sentido, los valores que se obtengan deben ser medidas que ofrezcan una visión global del comportamiento de ρ en todo el intervalo.

En este trabajo partimos de esta última premisa, el indicador ρ ya ha sido estimado previamente y es por tanto conocido. Buscamos entonces una estrategia de monitorización que nos permita, a partir de las medidas de ρ en unos determinados puntos, conocer su comportamiento en todo el intervalo. Pretendemos además hacer eso con el menor gasto posible, lo que nos lleva (sección 2) a formular el problema como un problema de optimización multiobjetivo, con variables reales y enteras. Asumiendo ciertas preferencias en la toma de decisiones, proponemos (sección 3) un algoritmo completo para obtener una solución que proporcione una técnica de monitorización *satisfactoria*. Finalmente (sección 4), ilustramos la bondad del método propuesto, aplicándolo a dos ejemplos sencillos.

2. FORMULACIÓN MATEMÁTICA

Supongamos que tenemos una función ρ definida sobre un intervalo $[a, b]$. Queremos diseñar una estrategia de monitorización que nos permita, a partir de las medidas de ρ en unos determinados puntos, conocer su comportamiento en todo el intervalo $[a, b]$. Para ello proponemos dividir el intervalo $[a, b]$ en un número ($N \in \mathbb{N}$, a determinar) de subintervalos ($[a_{i-1}, a_i]$, $i = 1, \ldots, N$, a determinar) y en cada subintervalo medir el indicador ρ en un punto ($p_i \in [a_i, a_{i+1}]$, también a determinar), con la intención de que la medida en ese punto nos de una idea del valor de ρ en todo el subintervalo. Atendiendo a esto, observamos que:

- El punto p_i debe ser tal que el valor de ρ en ese punto sea una buena aproximación de la media de ρ en el subintervalo correspondiente, esto es, debe verificarse que

$$d_i = (\rho(p_i) - \bar{\rho}_i)^2 \longrightarrow 0 \tag{1}$$

  donde $\bar{\rho}_i = \dfrac{\int_{a_{i-1}}^{a_i} \rho(x)dx}{a_i - a_{i-1}}$ es la media de ρ en el intervalo $[a_{i-1}, a_i]$.

- La media de ρ en cada subintervalo debe ser *representativa*, esto es, las desviaciones de la media deben ser lo más pequeñas posibles:

$$\sigma_i = \sqrt{\int_{a_{i-1}}^{a_i} (\rho(x) - \bar{\rho}_i)^2 dx} \longrightarrow 0 \tag{2}$$

- El objetivo es diseñar el mejor sistema de monitorización con el menor gasto posible y, en ese sentido, suponemos conocida una función creciente f que nos da el coste económico en función del número de mediciones.

De este modo, si definimos $a_0 = a$, $a_N = b$ y denotamos por $\delta > 0$ la longitud mínima de cualquier intervalo de muestreo, el problema consiste en elegir $N \in \mathbb{N}$, $a = (a_1, \ldots, a_{N-1}) \in \mathbb{R}^{N-1}$ y $p = (p_1, \ldots, p_N) \in \mathbb{R}^N$, verificando

$$a_i - a_{i-1} \geq \delta \quad i = 1, \ldots, N, \tag{3}$$
$$p_i \in [a_{i-1}, a_i] \quad i = 1, \ldots, N, \tag{4}$$

y de manera que se alcancen los siguientes objetivos:

1. Se minimice el coste económico, dado por

$$f : N \in \mathbb{N} \longrightarrow f(N) \in [0, +\infty) \tag{5}$$

2. Se obtengan medias representativas en todos y cada uno de los intervalos. Como ya hemos señalado, eso pasa por minimizar el funcional

$$J_\sigma(a) = q||\sigma(a)||_\infty + (1 - q)||\sigma(a)||_2, \tag{6}$$

   donde $\sigma(a) = (\sigma_1, \ldots, \sigma_N)$ está definida por (2), $q \in [0, 1]$ es un parámetro de peso y $||.||_\infty$ y $||.||_2$ denotan, respectivamente, la norma del máximo y la norma euclideana en $\mathbb{R}^N$.

3. En todos los subintervalos debe existir un punto (punto de muestreo) en el que se *capture* bien la media. Igual que antes, que eso ocurra pasa por minimizar el funcional

$$J_d(a, p) = r||d(a, p)||_\infty + (1 - r)||d(a, p)||_2, \tag{7}$$

   donde $d(a, p) = (d_1, \ldots, d_N)$ está definida por (1) y $r \in [0, 1]$ es el correspondiente parámetro de peso entre las normas.

Desde un punto de vista formal, para cada $N \in \mathbb{N}$ se define

$$U_{ad}^N = \left\{(a,p) \in \mathbb{R}^{N-1} \times \mathbb{R}^N \text{ verificando (3) y (4)}\right\},$$

se toma $U_{ad} = \bigcup_{N\in\mathbb{N}} \left(\{N\} \times U_{ad}^N\right)$ y definiendo entonces

$$\mathbf{J} : (N,a,p) \in U_{ad} \longrightarrow \mathbf{J}(N,a,p) = (f(N), J_\sigma(a), J_d(a,p)) \in \mathbb{R}^3, \tag{8}$$

el problema se formula como el siguiente problema de optimización multiobjetivo:

$$\min_{(N,a,p)\in U_{ad}} \mathbf{J}(N,a,p) \tag{9}$$

Nota 1 *Si la función ρ está definida en un rectángulo $R \subset \mathbb{R}^n$, $n > 1$ (pensemos, por ejemplo, en que queremos controlar la contaminación de un río a lo largo del tiempo, tomando medidas puntuales en unos determinados instantes), la técnica que acabamos de describir para formular el problema puede extenderse sin dificultad alguna, y lo único que aumentará significativamente es la dimensión de la variable de control. Si ρ toma valores en $\mathbb{R}^M$, $M > 1$ (pensemos, por ejemplo, que la mencionada contaminación del río la queremos controlar midiendo la concentración de diferentes sustancias) la técnica para formular el problema sigue siendo válida, y en ese caso lo que aumentará significativamente será el número de funcionales objetivo. Por el contrario, si ρ toma valores en un espacio de dimensión infinito (pensemos, por ejemplo, que la contaminación del río la queremos controlar midiendo la concentración de una sustancia -o de varias- a lo largo del tiempo), la técnica es también válida, pero los funcionales objetivo deben redefinirse para adaptarse a esa situación (concretamente, las expresiones de d_i y σ_i, dadas respectivamente por (1) y (2), deben ser modificadas). Como ejemplo, en Alvarez-Vázquez et al., 2015b puede verse como hacer eso cuando se desea monitorizar la calidad del agua de un río.*

3. RESOLUCIÓN NUMÉRICA

Obviamente, la resolución del problema (9) pasa, previamente, por ser capaz de estimar, de algún modo, el indicador ρ. Con frecuencia, esa estimación suele hacerse a través de modelos experimentales y/o numéricos suficientemente contrastados y, en cualquier caso, nosotros vamos a suponer que la función ρ es conocida, al igual que lo es la función f, de manera tal que podemos evaluar sin dificultad alguna el funcional $\mathbf{J}$. En estas circunstancias, la búsqueda de una solución *satisfactoria* del problema (9) puede abordarse por diferentes métodos. Concretamente, para un problema como este, en el que sólo hay tres funcionales objetivo, el uso de *métodos a posteriori* (método de pesos, método de ϵ-restricciones, ...) pueden permitir graficar la superficie del frente Pareto, en la que el tomador de decisiones puede luego apoyarse para elegir la solución que más le satisface (ver, por ejemplo, Miettinen, 1999). Del mismo modo, la combinación de *métodos interactivos* puede ayudar a guiar al tomador de decisiones a la hora de elegir esa solución satisfactoria (como ejemplo, en Alvarez-Vázquez et al, 2015a puede verse como se han combinado los métodos STEM y VIA en un problema de contaminación atmosférica).

En este trabajo vamos a asumir que el número de mediciones que se puede realizar es bajo y que, además, el tomador de decisiones ya ha indicado que quiere un sistema de monitorización que cuente con el menor número posible de mediciones y que garantice que, en todos los segmentos, las medias sean *suficientemente representativas*. Atendiendo a esa preferencia proponemos proceder del siguiente modo:

Algoritmo 1

- *Paso 1. Para cada $N = 1, 2, \ldots$ resolver el problema*

$$\begin{cases} \min J_\sigma(a) \\ \textit{sujeto a } a_i - a_{i-1} \geq \delta, \quad i = 1, \ldots, N \end{cases} \tag{10}$$

y mostrar al tomador de decisiones los valores mínimos de J_σ que se obtienen en cada caso.

- *Paso 2. Pedir al tomador de decisiones que, en virtud de la información obtenida en el paso anterior, elija el número de mediciones a realizar (N) y establezca una cota máxima para el valor de J_σ, que denotaremos por σ_{max}.*

- *Paso 3. Resolver el problema*

$$\begin{cases} \text{mín } J_d(a,p) \\ \textit{sujeto a} \quad a_{i-1} \leq p_i \leq a_i \quad i=1,\ldots,N \\ \qquad\qquad a_i - a_{i-1} \geq \delta \quad i=1,\ldots,N \\ \qquad\qquad J_\sigma(a) \leq \sigma_{max} \end{cases} \tag{11}$$

  y mostrar la solución obtenida al tomador de decisiones.

- *Paso 4. Preguntar al tomador de decisiones si la solución obtenida es satisfactoria. Si lo es, PARAR. Si no lo es, volver al Paso 2 y aumentar el número de mediciones a realizar (N) y/o disminuir la cota máxima para J_σ (σ_{max}).*

La aplicación de este algoritmo pasa por saber resolver, para los diferentes valores de N, los problemas (10) y (11). La dimensión del problema (10) es $N-1$ y, admitiendo que ese valor no es demasiado alto, en este trabajo proponemos resolver ese problema con el método de Nelder-Mead (ver Nelder & Mead, 1965), utilizando una *función penalty* para tratar las restricciones lineales. La dimensión del problema (11) es ya $2N-1$ y además las variables a y p juegan papeles muy diferentes, lo que desaconseja el uso directo de ese método, incluso para valores de N no muy grandes. Independientemente de que otro tipo de métodos de búsqueda directa o evolutivos puedan resultar útiles para resolver (11), teniendo en cuenta la experiencia obtenida por los autores en algún trabajo previo (ver Alvarez-Vázquez et al., 2006), describimos a continuación un método *ad-hoc* para resolver ese problema.

3.1 UN MÉTODO AD-HOC PARAR RESOLVER EL PROBLEMA (11)

El problema (11) depende de dos variables $a \in \mathbb{R}^{N-1}$ y $p \in \mathbb{R}^N$ que, no sólo tienen una interpretación física distinta, sino que también pueden ser tratadas de modo diferente a la hora de resolver el problema. En efecto, es fácil observar que para a dado (fijo), verificando (3) y $J_\sigma(a) \leq \sigma_{max}$, el valor de $p(a)$ verificando (4) que minimiza el valor de $J_d(a,p(a))$ puede obtenerse como la solución de N problemas de optimización unidimensional. Concretamente, para $i=1,\ldots,N$, basta tomar $p_i(a)$ como la solución del problema

$$\begin{cases} \text{mín } d_i(p_i) \\ \text{sujeto a} \quad p_i \in [a_{i-1}, a_i] \end{cases} \tag{12}$$

donde, obviamente, $d_i(p_i)$ está dado por (1).

De este modo, si denotamos $p(a) = (p_1(a),\ldots,p_N(a))$ y definimos $J(a) = J_d(a,p(a))$, el problema (11) se reduce a

$$\begin{cases} \text{mín } J(a) \\ \text{sujeto a} \quad a_i - a_{i-1} \geq \delta \quad i=1,\ldots,N \\ \qquad\qquad J_\sigma(a) \leq \sigma_{max} \end{cases} \tag{13}$$

de manera que, si a^* es la solución del problema (13), la solución del problema original (11) vendrá dada por $(a^*, p(a^*))$.

El problema (13) puede resolverse, igual que (10), con el método de Nelder-Mead, utilizando una *función penalty* para tratar ambas restricciones. Cada vez que se necesite evaluar la función $J(a)$, los problemas unidimensionales (12) pueden resolverse utilizando, por ejemplo, el método de la sección áurea (ver, por ejemplo, Bazaraa & Shetty, 1979).

4. EJEMPLOS

4.1 EJEMPLO 1

Para ilustrar la bondad del método propuesto, hemos consideramos el indicador $\rho(x)$ cuya gráfica se muestra en la Figura 1. En este ejemplo sencillo, es inmediato ver cual sería la división en intervalos que habría que hacer para una correcta monitorización del indicador y cual sería, en cada caso, la ubicación adecuada de las estaciones de muestreo. Los resultados obtenidos por el método descrito en la sección anterior se muestran en la tabla 1 y confirman que, efectivamente, en este ejemplo sencillo el método proporciona la respuesta deseada.

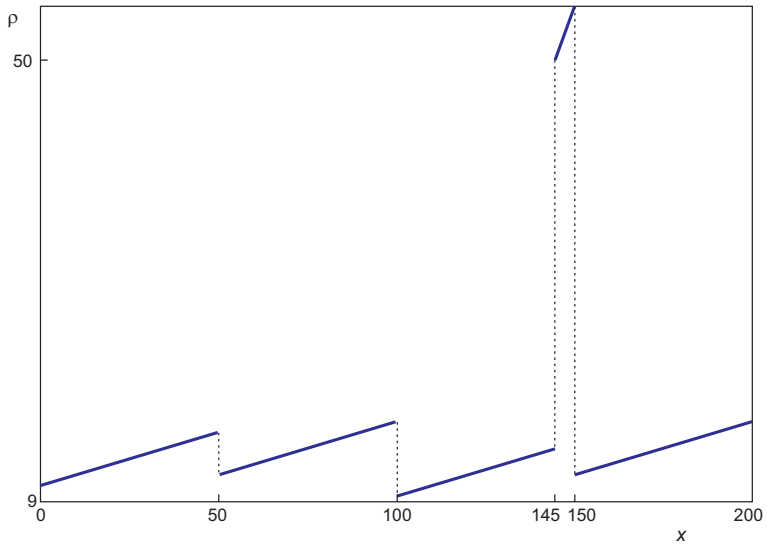


Figura 1: Función $\rho(x)$ considerada en el Ejemplo 1

N	a (óptimo)	$J_\sigma(a)$	$p(a)$ (óptimo)	$J_d(a, p(a))$
2	147.5	69.21	(31.14, 194.17)	0
3	(145,150)	20.64	(54.57, 147.5, 175)	0
4	(93.7,145,150)	15.28	(28.14, 128.08, 147.5, 175)	0
5	(50,100,145,150)	12.16	(25, 75, 122.5, 147.5, 175)	0

Tabla 1: Resultados obtenidos utilizando el método propuesto en la sección anterior para monitorizar el indicador mostrado en la Figura 1.

En este primer ejemplo, la división correcta en intervalos se obtiene sin más que resolver el problema (10). Para esa división los puntos de muestreo óptimos se obtienen también resolviendo directamente los correspondientes problemas (12), de manera que no es necesario resolver directamente el problema (11). Eso se debe a que, para los valores de a obtenidos al resolver (10), los puntos de muestreo óptimos que se obtienen (valores $p(a)$) son ya los mejores posibles ($d_i(p_i(a)) = 0$, $i = 1, \ldots, N$). Dicho de otro modo, para cada $N \in \mathbb{N}$, el frente-Pareto del problema multiobjetivo correspondiente está formado por un único punto. Obviamente, eso no es lo que suele pasar en la práctica, y en la siguiente sección se muestra otro ejemplo sencillo en el que eso ya no ocurre.

4.2 EJEMPLO 2

Consideramos el indicador $\rho(x)$ cuya gráfica se muestra en la Figura 2. Para $N = 2$, la solución que se obtiene al resolver el problema (10) es a =133.25, que se corresponde con un valor de $J_\sigma(a)$=443.39. Si para ese valor de a resolvemos los correspondientes problemas (12) obtenemos que los puntos óptimos son $p_1(a)$=64.99 y $p_2(a)$ =150. El primero de ellos es efectivamente óptimo,

en tanto que $d_1(p_1(a)) = 0$, esto es, la media se captura perfectamente, pero el segundo parece que podría mejorarse ya que $d_2(p_2(a))$=457.78 (la media en el intervalo (133.25, 200) no se captura correctamente). Para capturar mejor la media en el segundo intervalo es necesario mover el punto a y, consecuentemente, aumentar el valor de J_σ (empeorar la representatividad de esas medias). En estas circunstancias, sería cuando el tomador de decisiones debería aparecer (Paso 2 del algoritmo 1) y decirnos si aumenta el número de subintervalos y/o establece una cota máxima para J_σ (un valor mayor que el obtenido en el óptimo). Por ejemplo, si dejamos $N = 2$ y tomamos como cota $\sigma_{max} = 485$, los valores óptimos obtenidos al resolver el problema (11) son ya $a = 150$, $p(a)$ =(73.54, 175.0) que se corresponden con unos valores de $J_\sigma(a)$ =484.83 y $J_d(a, p) = 0$.

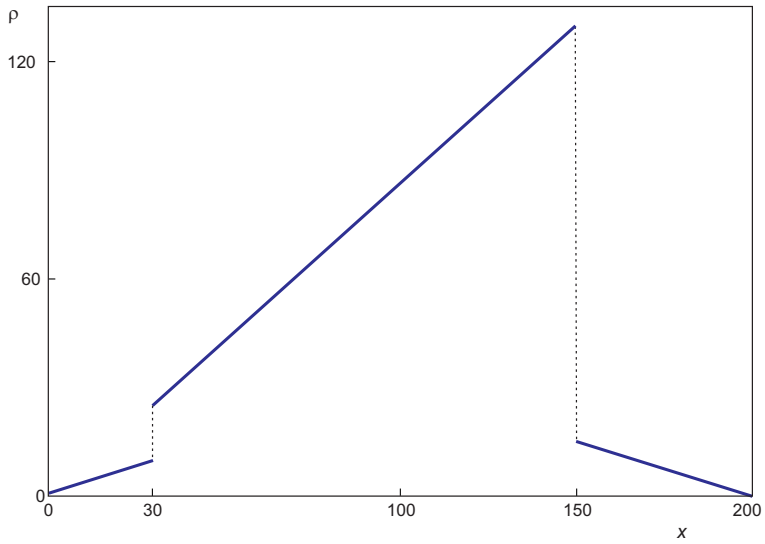


Figura 2: Función $\rho(x)$ considerada en el Ejemplo 2

REFERENCIAS

Alvarez-Vázquez, L.J., Martínez, A., Vázquez-Méndez, M.E., Vilar, M. (2006) Optimal location of sampling points for river pollution control. Math. Comput. Simulat., 71, 149–160.

Alvarez-Vázquez, L.J., García-Chan, N., Martínez, A., Vázquez-Méndez, M.E. (2015a) An application of interactive multicriteria optimization to air pollution control. Optimization, 64:6, 1367–1380.

Alvarez-Vázquez, L.J., Casal G., Martínez, A., Vázquez-Méndez, M.E. (2015b) Mathematical formulation for designing sampling techniques. Application to the design of a river quality monitoring system. Preprint.

Bazaraa, M.S., Shetty, C.M. (1979) Nonlinear Programming. Theory and Algorithms. New York: John Wiley & Sons.

Gilbert, R.O. (1987) Statistical methods for environmental pollution monitoring. New York: Van Nostrand Reinhold.

Miettinen K.M. (1999) Nonlinear multiobjective optimization. Boston: Kluwer Academic Publishers.

Nelder J.A., Mead R. (1965) A simplex method for function minimization. Comput J., 7, 308-–313.

Thompson, S.K. (2012) Sampling. Wiley.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

Study and characterization of the multinomial values

Margarita Domènech, José Miguel Giménez and M. Albina Puente

Department of Applied Mathematics III and Engineering School of Manresa,
Technical University of Catalonia, Spain.

ABSTRACT

We investigate the conditions for the coefficients of multinomial values of cooperative games necessary and / or sufficient in order to satisfy some properties, including marginal contributions, balanced contributions, desirability relation and null player exclusion property.

Keywords: ITU cooperative game, probabilistic value, multinomial value, semivalue, regularity.

1. INTRODUCTION

Weber's general model for assessing cooperative games [6] is based on *probabilistic values*, a family of values axiomatically characterized by means of linearity, positivity, and the dummy player property. Every probabilistic value is defined by a set of *weighting coefficients* and allocates, to each player in each game of its domain, a convex sum of the marginal contributions of the player in the game. These allocations can be interpreted as a measure of players' bargaining relative strength. The most conspicuous member of this family (in fact, the inspiring one) is the *Shapley value* [5]. In the present paper we define and study a subfamily of probabilistic values that we call *multinomial (probabilistic) values*. Technically, their main characteristic is the *systematic generation of the weighting coefficients in terms of a few parameters (one parameter per player)*.

For more than a decade, our research group has been studying *semivalues*, a subfamily of probabilistic values introduced by Dubey et al. [1], characterized by anonymity, and including the Shapley value as the only *efficient* member. In the analysis of certain cooperative problems we have successfully used *binomial semivalues*, a monoparametric subfamily defined by Puente [4] that includes the Banzhaf value introduced by Owen [3].

From this experience, we feel that multinomial values (defined by n parameters, n being the number of players) offer a deal of flexibility clearly greater than binomial semivalues (defined by one parameter), and hence many more possibilities to introduce additional information when evaluating a game.

Probabilistic values provide tools to study not only games *in abstracto* (i.e. from a merely structural viewpoint) but also the influence of *players' personality* on the issue. They are assessment techniques that do not modify the game but only the criteria by which payoffs are allocated. In the multinomial case, the parameters are used to cope with different attitudes the players may hold when playing a given game, even if they are not individuals but countries, enterprizes, parties, trade unions, or collectivities of any other kind. We attach to parameter p_i the meaning of *generical tendency of player i to form coalitions*, assuming p_i and p_j independent of each other if $i \neq j$.

The aim of the paper is that providing a self–contained theory for multinomial values. We investigate the conditions for the coefficients of these values necessary and / or sufficient in order to satisfy some properties including marginal contributions, balanced contributions, desirability relation and null player exclusion property. The notions of regularity and hereditary value guaranty the validality of some of them. Some of these properties were studied by Felshenthal and Machover [2] for several power indices.

Two kind of results have been established. On one hand, general results for all probabilistic values. On the other hand, results not so general but interesting in the sense that they evidence the great influence that the weighting coefficients and the tendency profiles may exert on the validity of classical properties in this setup.

We also characterize: (i) the class of regular values within the class of probabilistic values; (ii) the (i, j)–symmetric values within the class of probabilistic values; and (iii) the class of solutions satisfying the balanced contributions property within the class of regular and hereditary regular probabilistic values.

Summing up, this paper presents multinomial values as a consistent alternative or complement to classical values (Shapley, Banzhaf). Tendency profiles can encompass a variety of situations arising from players' attitudes. Thus, multinomial values represent a wide generalization of binomial semivalues, whose monoparametric condition implies a quite limited capability of analysis of such situations. Of course, these situations cannot be analyzed, without modifying the game, by means of the classical values, which can be concerned only with the structure of the game.

REFERENCES

[1] Dubey, P., Neyman, A. and Weber, R.J. (1981): "Value theory without efficiency." *Mathematics of Operations Research* 6, 122–128.

[2] Felsenthal, D. S. and Machover, M. (1995): "Postulate and paradoxes of relative voting power –a critical re–appraisal." *Theory and Decision* 38, 195–229.

[3] Owen, G. (1975): "Multilinear extensions and the Banzhaf value." *Naval Research Logistics Quarterly* 22, 741–750.

[4] Puente, M.A. (2000): Contributions to the representability of simple games and to the calculus of solutions for this class of games (in Spanish). Ph.D. Thesis. Technical University of Catalonia, Spain.

[5] Shapley, L.S. (1953): "A value for n-person games." In: Contributions to the Theory of Games II (H.W. Kuhn and A.W. Tucker, eds.), Princeton University Press, 307–317.

[6] Weber, R.J. (1988): "Probabilistic values for games." In: The Shapley Value: Essays in Honor of Lloyd S. Shapley (A.E. Roth, ed.), Cambridge University Press, 101–119.

XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 y 24 de octubre de 2015

EL VALOR DE SHAPLEY PARA MODELOS COOPERATIVOS BI-OBJETIVO

Ana Meca [1], A. Kimms [2] y I. Kozeletskyi [2]

[1] I.U. Centro de Investigación Operativa (CIO); Universitdad Miguel Hernández (UMH)
[2] Mercator School of Management; Duisburg-Essen University, Germany

RESUMEN

En este trabajo presentamos un modelo cooperativo con un número finito de agentes y dos objetivos para cada agente: un objetivo común (utilidad transferible) y un objetivo individual (utilidad no transferible). Utilizando la teoría de juegos cooperativos con utilidad no transferible (juegos NTU), definimos una nueva clase de juegos NTU que se denomina Bi-Allocation Games.

En primer lugar, demostramos que siempre existe una forma óptima de repartir el beneficio generado conjuntamente entre todos los agentes (a través del objetivo común) y, al mismo tiempo, satisfacer el objetivo individual de cada uno de ellos, que sea justa y beneficie a todos. Proponemos el valor de Shapley como solución para los Bi-Allocation Games.

En segundo lugar, desarrollamos un procedimiento que permita obtener una buena aproximación de dicha solución, el valor de Shapley para BiA Games, y chequeamos el funcionamiento de dicho algoritmo utilizando un nuevo modelo cooperativo de producción lineal multio-bjetivo.

Palabras y frases clave: Modelos Cooperativos Bi-objetivo, NTU games, Bi-allocation games, Shapley value.

.

Autores:

Alberca, P. 275
Alvarez Vázquez, L.J. 392
Ameijeiras Alonso, J. 123
Aneiros Pérez, G. 121
António, C.C. 384
Araya Alpízar, C. 48
Arias Rodil, M. 114
Bécue Bertaut, M. 5, 219
Borrajo García, M.I. 176
Boubeta Martínez, M., ... 232
Calaza Díaz, L. 282
Cao Abad, R. 51
Carreira Figueiras, V. 301
Carpente Rodríguez., M.L. 15
Carvalho das Neves, J. 343
Casado Barrio, M.J. 78
Casal, G. 392
Casas Méndez, B. 98, 170
Castillo Páez, S. 182
Castro Iglesias, A. 332
Cavique, L. 7
Cerdeira Pena, A. 15
Coelho do Carmo, M.A. 230
Conde Amboage, M. 135
Cotos Yáñez, T.R. 364
Crujeiras Casais, R.M. . 123, 282, 292, 301
de Miguel Domínguez, J.C. 306
Diéguez Aranda, U. 114
Domènech, M. 398
Durval, M. 343
El Ghouch, A. 191
Esteban, M.D. 261
Faraldo Roca, P. 301
Febrero Bande, M. 206
Fernandes, M. 273
Fernández Casal, R. 182, 201
Fernández de Castro, B.M. 301
Fernández Llanes, A.B. 266
Fernández Montes, M.C. 301, 306
Ferreira, M. 350
Fiestras Janeiro, M.G 160, 174
Flores, M. 201
Fuentes Santos, I. 382
Gallardo, J.M. 174
García Jurado, I. 160
García Luque, J. 372
García Quiroga, M. 371
García Rodríguez, M.T. 334
García Soidán, P. 182, 309
Giménez, J.M. 161, 398
Ginzo Villamayor, M.J. 98, 292
Gómez Costa, I. 98
Gómez López, C. 309
Gómez Rúa, M. 33
González Díaz, J. 2, 103
González Manteiga, W. 135, 176, 364, 382
González Rueda, A. 103
González Saez, L. 330
González Santamaría, P. 326, 330
Guisado González, M. 306
Hervada Vidal, X. 341
Hoffbauer, L.N. 384
Iglesias Patiño, C. 234
Iglesias Pérez, M.C. 356
Infante, P. 230
Jácome Pumar, M.A. 51, 63
Jiménez Gamero, M.D. 186, 191
Jiménez Losada, A. 174
Kimms, A. 400
Kozeletskyi, I. 400
Liang, H.Y. 356
Lombardía Cortiña, M.J. 232, 262, 301
López Calviño, B. 326, 330, 332, 334, 338, 341
López de Ullibarri, I. 63
López García, M.M. 373
López Múñiz, J. 3
López Pereira, J.E. 301, 306
López Vizcaíno, M.E. 234, 262
López Cheda, A. 51
Lorenzo Freire, S. 15
Luaces Pazos, R. 319
Mendes, J.M. 230
Maestro Requena, M.G. 374
Malvar Pintos, A. 266
Marey Pérez, M.F. 232, 374, 382
Martínez Calvo, A. 301
Martínez Miranda, M.D. 176
Martínez, A. 392
Matías Fernández, J.M. 301

Meca, A. 400
Meijide, H. 332
Meijide Vecino, M. 319
Mena, A. 332
Meneses Freire, A. 197
Monjardino, I.C. 242
Morales, D. 232, 261
Morrondo Pelayo, M.P. 301, 306
Mosquera Rodríguez, M.A. 174
Mouriño, H. 273, 343
Muñoz Ferreiro, M.N. 371, 372
Muñoz Murgui, F. 324
Naveira Barbeito, G. 266
Naya Fernández, S. 197, 201, 206, 301, 345
Neves Figueiredo, M.M. 41
Novoa Flores, G.I. 170
Oliveira, E. 247
Oliveira, P.E.. 40
Oviedo de la Fuente, M. 206
Pardo Fernández, J.C. 191
Sánchez Fernández P. 234
Pedreira, J. 332
Pereira Andrade, L.P. 306
Pérez Fernández de Córdoba, M. 103
Pérez Fontán, M. 338
Pérez González, A. 364
Pérez, A. 261
Pértega Díaz, S. .. 326, 330, 332, 334, 338, 341
Piñón Villar, M.C. 334
Pires Manso, J.R. 247
Pita Fernández, S. 326, 330, 334, 338, 341
Places, A.S. 15
Pollán Santamaría, M. 266
Poveda, E. 332
Prieto Aneiros, A. 211
Puente, M.A. 161, 398
Pujales Martínez, X.E. 84
Ramil Novo,L.A. 373
Raña Míguez, P. 121
Rebelo, M. 350
Riesco Muñoz, G. 220
Ríos Viqueira, J.R. 282
Rivas Martínez, G. 186
Robles Bykbaev, Y. 345
Rodríguez Camacho, E. 330
Rodríguez Casal, A. 123, 282
Rodríguez González, D. 306
Rodríguez Muíños, M.A. 341
Rodríguez Osorio, I. 332
Rodríguez Veiga, J. 98, 170
Rodríguez-Gigirey Pérez, M.A. 75
Rodríguez-Moldes Rey, C. 90
Romero Rodríguez, M.A. 371, 372, 373
Rosa, S. 273
Rueda, C. 262
Saavedra Nieves, A. 160
Saavedra Nieves, P. 382
Sánchez Bello, P. 319
Sánchez Carnero, N. 211
Sánchez Espigares, J.A. 219
Sánchez Sellero, C. 135
Santiago Pérez, M.I. 266
Santiso Blanco, J. 260
Santos, J. 7, 275
Segura Gisbert, J. 324
Seijo Bestilleiro, R. 326
Seoane Pillado, T. 326, 330, 332, 334, 338, 341
Sousa Fernández, X. 282
Suárez Quintas, S. 282
Tarrío Saavedra, J. 197, 201, 206, 211, 345
Tavares, D. 343
Trudeau C. 21
Vaamonde Liste, A. 319
Valcárcel Aguiar, B. 306
Valencia Toledo, A. 148
Van Keilegom, I. 51
Vázquez Ordériz, M.L. 371, 372, 373
Vázquez Pita, M. 372
Vázquez Méndez, M.E. 114, 392
Vecino Aguirre, A. 306
Vidal Puga, J. 21, 33, 148
Vieu, P. 4
Vilar Fernández, J. 121
Zaragoza, S. 206
Zubizarreta Aberdi, R. 266

ORGANIZA

Departamento de Estatística e Investigación Operativa da
Universidade de Santiago de Compostela e
Sociedade Galega para a Promoción da Estatística e da Investigación de Operacións

22, 23 e 24 de OUTUBRO 2015 Lugo

XII CONGRESO GALEGO DE ESTATÍSTICA E INVESTIGACIÓN DE OPERACIÓNS