II ENCONTRO GALAICO-PORTUGUÉS DE BIOMETRÍA,

con aplicación ás Ciencias da Saúde, á Ecoloxía e ás Ciencias do Medio Ambiente

SANTIAGO DE COMPOSTELA

30 Xuño - 2 Xullo 2016

http://biometria.sgapeio.es/

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

SPE
Sociedade Portuguesa de Estatística

II ENCONTRO GALAICO-PORTUGUÉS DE BIOMETRÍA

con aplicación ás Ciencias da Saúde, á Ecoloxía e ás Ciencias do Medio Ambiente

PROGRAMA
E RESUMOS

Santiago de Compostela

30 de Xuño – 2 de Xullo de 2016

ORGANIZAN:

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

Sociedade Portuguesa de Estatística (SPE)

Editoras: María José Ginzo Villamayor
Sonia Pértega Díaz
María Isolina Santiago Pérez

ISBN: 978-84-608-8178-0
Xuño 2016

PRESENTACIÓN

A Sociedade Galega de Estatística e Investigación de Operacións (SGAPEIO), e a Sociedade Portuguesa de Estatística (SPE) comprácense en presentar o II Encontro Galaico-Portugués de Biometría, con aplicación ás Ciencias da Saúde, á Ecoloxía e ás Ciencias do Medio Ambiente, o BIOAPP2016.

Trátase dunha nova iniciativa conxunta das dúas Sociedades, tralos Encontros celebrados en Braga (2013) e Vila Real (2014), todos eles co obxectivo común de difundir os últimos avances no desenvolvemento e aplicación de métodos estatísticos e matemáticos na Bioloxía, Medicina, Ecoloxía, Psicoloxía, Farmacoloxía, Agricultura, Medio Ambiente e outras ciencias da Vida. Nesta ocasión, será Santiago de Compostela a cidade anfitrioa, invitando aos socios e socias da SGAPEIO e da SPE, e a toda a comunidade científica en xeral, a facer unha peregrinaxe biométrica para compartir tres días de intenso programa científico e interesantes actividades sociais e culturais.

Contaremos coa presenza de recoñecidos profesionais da Estatística para impartir 4 sesións plenarias, 4 sesións invitadas e unha mesa redonda, ademais dun minicurso sobre análise de supervivencia. Tamén contamos coas numerosas contribucións dos participantes, que van presentar, neste Encontro, 57 comunicacións orais e 40 en formato de póster. Coa convocatoria de varios premios poderemos outorgar un recoñecemento científico e unha aportación económica aos mellores destes traballos.

Todo isto non sería posible sen a colaboración de diferentes institucións, organismos públicos e empresas, ás que agradecemos a súa contribución. A organización dun congreso supón sempre un desafío, que o comité organizador desde II Encontro Galaico-Portugués de Biometría afrontou con ilusión e moito traballo. Confiamos que os congresistas desfruten do Encontro e dunha cidade que os acolle cos brazos abertos.

Santiago de Compostela, xuño de 2016

O Comité Organizador

COMITÉ ORGANIZADOR

GALICIA

Presidenta:

Mª Isolina Santiago Pérez.
Dirección Xeral de Saúde Pública (Consellería de Sanidade).

Vocais:

Mª Esther López Vizcaíno.
Instituto Galego de Estatística.

Mª José Ginzo Villamayor.
Universidade de Santiago de Compostela.

Balbina Virginia Casas Méndez.
Universidade de Santiago de Compostela.

César Sánchez Sellero.
Universidade de Santiago de Compostela.

Graciela Estévez Pérez.
Universidade da Coruña.

Mª Carmen Iglesias Pérez.
Universidade de Vigo.

Sonia Pértega Díaz.
Instituto de Investigación Biomédica (INIBIC). Complexo Hospitalario Universitario da Coruña (CHUAC).

PORTUGAL

Presidenta:

Patrícia de Zea Bermudez.
Universidade de Lisboa.

Vocais:

Irene Oliveira.
Universidade de Trás-os-Montes e Alto Douro.

Giovani Silva.
Universidade de Lisboa.

Dora Pratas.
Universidade Nova de Lisboa.

COMITÉ CIENTÍFICO

GALICIA

Presidenta:

Carmen Cadarso Suárez.
Universidade de Santiago de Compostela.

Vocais:

Mª Esther López Vizcaíno.
Instituto Galego de Estatística.

Wenceslao González Manteiga.
Universidade de Santiago de Compostela.

Rosa Crujeiras Casais.
Universidade de Santiago de Compostela.

José Antonio Vilar Fernández.
Universidade da Coruña.

Ricardo Cao Abad.
Universidade da Coruña.

Javier Roca Pardiñas.
Universidade de Vigo.

PORTUGAL

Presidenta:

Antónia Turkman.
Universidade de Lisboa.

Vocais:

Raquel Menezes.
Universidade do Minho.

Pedro Oliveira.
Universidade do Porto.

Russell Alpizar Jara.
Universidade de Évora.

Carlos Daniel Paulino.
Universidade de Lisboa.

Maria Eduarda Silva.
Universidade do Porto

INFORMACIÓN XERAL

SEDE

Facultade de Matemáticas

Universidade de Santiago de Compostela

C/ Lope Gómez de Marzoa s/n

15782, Santiago de Compostela

DATAS

30 de xuño, 1 e 2 de xullo de 2016

ACCESO WIFI NA SEDE

SSID: biometria2016

Usuario: biometria2016

Contrasinal: 15TT%&r

CERTIFICADOS

Todos os certificados se enviarán en formato dixital por correo electrónico unha vez rematado o Encontro.

UBICACIÓNS NA FACULTADE

Minicurso: Aula 1 (nivel 1).

Actos de apertura e clausura: Aula Magna.

Sesións plenarias e invitadas: Aula Magna.

Mesa redonda: Aula Magna.

Comunicacións orais: Aulas 7, 8, 9 (nivel 4).

Sesión de posters: corredor nivel 3.

Cafés: corredor nivel 3.

RECEPCIÓN DE BENVIDA

Xoves, 30 de xuño ás 21:00 h.

Pazo de Fonseca

Praza de Fonseca, 2

15704, Santiago de Compostela

COMIDAS DO ENCONTRO

Restaurante Monte da Condesa

Rúa de Xoaquín Díaz de Rábago, s/n

15782, Santiago de Compostela

CEA DO ENCONTRO

Venres, 1 de xullo ás 21:00 h.

Hostal dos Reis Católicos

Rúa de San Francisco, 1

15704, Santiago de Compostela

ÍNDICE

Programa 2
Minicurso 17
Sesións plenarias 19
Sesións invitadas 24
Mesa redonda 29
Comunicacións orais 32
Premio Ciencias da Saúde 32
Métodos estatísticos en Ciencias do Medio Ambiente 57
Modelos mixtos e EEG 79
Premio Ecoloxía-Medio Ambiente 100
Modelos de regresión distribucional aditivos estructurados e GAM 122
Métodos estatísticos en Medicina I 141
Métodos estatísticos en Agricultura e Bioloxía 159
Epidemioloxía I 166
Métodos estatísticos en Medicina II 177
Epidemioloxía II 193
Métodos estatísticos en Ecoloxía 209
Modelización conxunta 220
Comunicacións en poster 233
Autores 363

Programa

PROGRAMA

Xoves 30 de xuño

9:30-11:30	**Minicurso**: *Survival Analysis* **-** Luis Filipe Meira
11:30-12:00	Pausa-café
12:00-14:00	**Minicurso**: *Survival Analysis* **-** Luis Filipe Meira
14:00-15:30	Xantar
15:00-15:30	Acreditación
15:30-16:00	Acto de apertura
16:00-17:00	**Sesión plenaria I**: *Structured Additive Conditional Copula Regression Models* **-** Thomas Kneib
17:00-18:30	Comunicacións orais. Sesións I, II. Sesión Premio Ciencias da Saúde
19:00-20:30	Paseo turístico e matemático pola zona monumental
21:00-21:45	Recepción no Pazo de Fonseca e pincho de benvida

Venres 1 de xullo

9:00-10:30	Comunicacións orais. Sesións III, IV. Sesión Premio Ecoloxía-Medio Ambiente
10:30-11:10	**Sesión invitada I**: *Spatio-temporal species distribution modelling* - David Conesa
11:10-11:30	Pausa-café
11:30-12:30	Sesión de posters
12:30-13:30	**Sesión plenaria II**: *Adding spatial information to traditional statistical inference* - Daniela Cocchi
13:30-15:00	Xantar
15:00-15:40	**Sesión invitada II**: *What is the dose of radiation received? Statistical advances in Biodosimetry* **-** Pere Puig
15:40-17:00	Comunicacións orais. Sesións V, VI, VII
17:00-17:30	Pausa-café
17:30-18:30	**Sesión plenaria III**: *A characterization of missingness at random in a generalized shared-parameter joint modeling framework for longitudinal and time-to-event data, and sensitivity analysis* **-** Geert Molenberghs
19:00-20:00	Visita ao CESGA
21:00	Cea do Encontro e entrega de premios

Sábado 2 de xullo

10:20-11:00	**Sesión invitada III:** *Statistical challenges in air pollution health risk assessment* **-** Michela Cameletti
11:00-11:30	Pausa-café
11:30-12:30	Comunicacións orais. Sesións VIII, IX, X
12:30-13:30	**Sesión plenaria IV:** *Algumas aplicações de equações diferenciais estocásticas na modelação de fenómenos biológicos em ambiente aleatório* **-** Carlos Braumann
13:30-15:00	Xantar
15:00-15:40	**Sesión invitada IV**: *La agricultura del siglo XXI: nuevos retos para la biometría* **-** José Luis G. Andújar
15:40-17:00	**Mesa redonda**: *El papel de la Estadística en las CC. de la Salud, la Ecología y el Medioambiente: Reflexiones sobre la situación actual y nuevos retos*. Moderador: José Antonio Vilar Fernández Participantes: David Conesa, José Miguel Pereira, Luzia Gonçalves, Xurxo Mariño
17:00-17:30	Acto de clausura

Xoves 30 de xuño

9:30 – 11:30 e 12:00 – 14:00. Aula 1

Minicurso

ANÁLISE DE SUPERVIVENCIA.
Luís F. Meira Machado. Universidade do Minho.

16:00 – 17:00. Aula Magna

Sesión Plenaria I

Moderadora: **Carmen Cadarso Suárez**.

STRUCTURED ADDITIVE CONDITIONAL COPULA REGRESSION MODELS.
Thomas Kneib. University of Göttingen.

17:00 – 18:30. Aula Magna

Sesión Premio Ciencias da Saúde

Moderador: **Russell Alpizar-Jara**.

P1.1. CLASSIFICAÇÃO HIERÁRQUICA COM DISTRIBUIÇÕES A POSTERIORI: UM ESTUDO DE SIMULAÇÃO EM PADRÕES TEMPORAIS DE VIH.
Diana Rocha, Sónia Gouveia, Carla Pinto, Manuel Scotto, João Nuno Tavares.

P1.2. AJUSTE DO PESO E GORDURA DE RATOS POR MEIO DO MODELO DE VON BERTALANFFY.
Magali T. R. Amaral, Danielle F. Vileigas e Antãnio Carlos Cicogna, Carlos R. Padovani.

P1.3. JOINT MODELLING OF MULTIVARIATE LONGITUDINAL AND SURVIVAL DATA. A TWO-STAGE MODEL BASED PROPOSAL WITH APPLICATION TO NEPHROLOGY RESEARCH.
Ipek Guler, Christel Faes, Carmen Cadarso-Suárez, Laetitia Teixeira, Anabela Rodrigues, Denisa Mendonça.

P1.4. A NEW METHOD FOR DETECTION OF RHYTHMIC SIGNALS IN OSCILLATORY SYSTEMS.
Yolanda Larriba, Cristina Rueda, Miguel A. Fernández and Shyamal D. Peddada.

P1.5. SELECCIÓN DE COVARIABLES SIGNIFICATIVAS EN LA INCIDENCIA DE UN MODELO DE CURACIÓN NO PARAMÉTRICO.
Ana López-Cheda, Mª Amalia Jácome y Ricardo Cao.

P1.6. PREDICCION ONLINE DE LA GRIPE.
M. Oviedo de la Fuente, M. Febrero Bande, L. Basile, N. Torner, A. Martínez.

17:00 – 18:30. Aula 7

Sesión de Orais I: Métodos estatísticos nas Ciencias do Medio Ambiente

Moderadora: **Mª José Lombardía Cortiña**.

I.1. TECNOLOGÍA LÁSER ESCÁNER 3D EN INVENTARIO FORESTAL.
Julia Armesto, Juan Picos y Xana Álvarez.

I.2. NUEVOS MÉTODOS DE ESTIMACIÓN DEL NÚMERO DE INCENDIOS FORESTALES EN GALICIA.
Miguel Boubeta, María José Lombardía, Manuel Marey y Domingo Morales.

I.3. A PERIODIC STATE SPACE MODEL TO MONTHLY LONG-TERM TEMPERATURE DATA.
Marco Costa and Magda Monteiro.

I.4. DYNAMIC ENVIRONMENTAL ASSESSMENT USING DEA WINDOW ANALYSIS FOR WASTEWATER TREATMENT PLANTS IN SPAIN.
Yago Lorenzo-Toja, Ian Vázquez-Rowe, Desirée Marín-Navarro, Rosa M. Crujeiras, María Teresa Moreira, Gumersindo Feijoo.

I.5. DYNAMIC FACTOR MODELS FOR BIVARIATE COUNT DATA: AN APPLICATION TO FIRE ACTIVITY.
Magda Monteiro, Isabel Pereira e Manuel G. Scotto.

I.6. ESTIMACIÓN DEL NIVEL DE RIESGO CONTAMINANTE DE LOS RESIDUOS PETROLÍFEROS APLICANDO TÉCNICAS DE CLASIFICACIÓN FDA.
Miguel Flores, Salvador Naya, Javier Tarrío-Saavedra, Luis Horna, Rolando Sáez y Ana Escobar.

17:00 – 18:30. Aula 8

Sesión de Orais II: Modelos mixtos e EEG

Moderador: **Dalton Francisco de Andrade**.

II.1. AJUSTE DE MODELOS MISTOS A DADOS LONGITUDINAIS PROBLEMÁTICOS.
Julio M. Singer, Carmen D.S. André, Francisco M.M. Rocha e Talita Zerbini.

II.2. MODELO MULTINÍVEL NA AVALIAÇÃO DO NÍVEL ÓSSEO MARGINAL PERI-IMPLANTAR EM FUNÇÃO DA HIGIENE BUCAL.
Omar Pereira, Mariana Tormena, Flávia Matarazzo, Maurício Guimarães Araújo, Isolde Previdelli.

II.3. A SUPERIORIDADE DO MODELO MULTINÍVEL PARA AVALIAR DADOS ODONTOLÓGICOS.
Tiago Peres da Silva Suguiura, Omar Cléo Neves Pereira, Waenya Fernandez de Carvalho, Isolde Previdelli.

II.4. EQUAÇÕES DE ESTIMAÇÃO GENERALIZADAS PARA MODELAGEM DE DADOS COMPORTAMENTAIS FARMACOLÓGICOS.
Matheus Henrique Dal Molin Ribeiro, Omar Pereira, Humberto Milani e Isolde Previdelli.

II.5. UMA APLICAÇÃO DOS POLINÔMIOS FRACIONÁRIOS EM MODELOS MISTOS.
Edijane Paredes Garcia e Luzia Aparecida Trinca.

Venres 1 de xullo

9:00 – 10:30. Aula Magna

Sesión Premio Ecoloxía-Medio Ambiente

Moderadora: **Rosa M. Crujeiras Casais**.

P2.1. MULTIVARIATE NEIGHBORHOOD PENALTIES FOR ANALYZING MUSSEL ABUNDANCE AT THE GALICIAN COAST.
Hauke Thaden, Nadja Klein, Thomas Kneib, María P. Pata and Carmen Cadarso-Suárez.

P2.2. MODELLING INTRA- AND INTER-DAY VARIABILITY OF NO_2 CONCENTRATIONS IN PORTUGAL.
Andreia Monteiro, Raquel Menezes and Maria Eduarda Silva.

P2.3. ESTIMATING BLAINVILLE'S BEAKED WHALES ABUNDANCE OVER TIME: IMPROVING DIVE COUNTING METHODS.
Patrícia A. Jorge, Jessica W. Shaffer, Tiago A. Marques, Helena Mouriño, Len Thomas, David J. Moretti, Nancy DiMarzio.

P2.4. CAPTANDO A ORIXE DOS PATRÓNS DE DISTRIBUCIÓN DE ABUNDANCIA DE ESPECIES DO PLANCTO MICROBIANO: UNHA ANÁLISE DA DISPERSIÓN E A OCUPACIÓN DE NICHOS.
Miguel Blanco Vázquez, Salvador Pueyo, Francisco G. Figueiras, Isabel Teixeira e Pedro Cermeño.

P2.5. UN TEST NO PARAMÉTRICO PARA COMPARAR EL PATRÓN ESPACIAL DE LOS INCENDIOS FORESTALES REGISTRADOS EN GALICIA.
Isabel Fuentes-Santos, Wenceslao González-Manteiga, Jorge Mateu y Manuel F. Marey-Pérez.

P2.6. POLÍTICAS DE PESCA SUSTENTÁVEIS ÓTIMAS COM ESFORÇO CONSTANTE VERSUS POLÍTICAS DE PESCA ÓTIMAS COM ESFORÇO VARIÁVEL: COMPARAÇÃO E APLICAÇÃO EM AMBIENTE ALEATÓRIO COM O MODELO DE GOMPERTZ.
Nuno M. Brites e Carlos A. Braumann.

9:00 – 10:30. Aula 7

Sesión de Orais III: Modelos de regresión distribucional aditivos estructurados e GAM

Moderadora: **Marta Sestelo**.

III.1. UTILIDAD DE LOS MODELOS DE REGRESIÓN DISTRIBUCIONAL ADITIVOS ESTRUCTURADOS EN LA TOMA DE DECISIONES CLÍNICAS. A PROPÓSITO DEL POTASIO.
Jenifer Espasandín-Domínguez, Carmen Cadarso-Suárez, Thomas Kneib, Balbina Casas-Méndez, Alfonso Javier Benitez-Estévez, Tegra Barreiro-Martínez, Francisco Gude.

III.2. COMPARING HIGHER-DIMENSIONAL TENSOR PRODUCT AND MARKOV RANDOM FIELDS WHEN MODELING A TRIVARIATE INTERACTION TERM IN A STRUCTURED ADDITIVE PREDICTOR USING BAYESX.
Elisa Duarte, Bruno de Sousa, Carmen Cadarso-Suárez, Thomas Kneib and Vítor Rodrigues.

III.3. STUDYING THE RELATIONSHIP BETWEEN REPRODUCTIVE LIFE SPAN AND MENARCHE AGE USING BAYESIAN MULTIVARIATE STRUCTURED ADDITIVE DISTRIBUTIONAL REGRESSION.
Elisa Duarte, Bruno de Sousa, Carmen Cadarso-Suárez, Nadja Klein, Thomas Kneib and Vítor Rodrigues.

III.4. EMPLEO DE GAMLSS PARA EL CÁLCULO DE PERCENTILES DE RENDIMIENTO EN UNA PRUEBA MÁXIMA DE 2000 METROS SOBRE ERGÓMETRO DE REMO.
Telmo Silva Alonso, María del Carmen Iglesias Pérez y José Luis García Soidán.

III.5. AUC-BASED MODEL SELECTION IN GENERALIZED ADDITIVE MODELS FOR THE PREDICTION OF NEW THERAPEUTIC TARGETS.
P Cacheiro, J Roca-Pardiñas, C Cadarso-Suárez, A Carracedo.

9:00 – 10:30. Aula 8

Sesión de Orais IV: Métodos estatísticos en Medicina I

Moderador: **Julio M. Singer**.

IV.1. PROPUESTA DE DESARROLLO DEL SCORE GLOBAL DE UN CUESTIONARIO CON UNA ESTRUCTURA UNIDIMENSIONAL ORDENADA MEDIANTE EL ANÁLISIS MCA.
Maider Mateo-Abad, Arantza Urkaregi, Jone Lázaro, Inmaculada Arostegui.

IV.2. TEORIA DA RESPOSTA AO ITEM: APLICAÇÕES EM ENFERMAGEM E SAÚDE.
José W. Pereira Borges, Thereza M. Magalhães Moreira, Dalton F. de Andrade.

IV.3. CONTRASTE DA IGUALDADE DUN GRAN NÚMERO DE DENSIDADES BAIXO DEPENDENCIA.
Marta Cousido Rocha, Jacobo de Uña Álvarez e Jeffrey D. Hart.

IV.4. TEST DE MULTIMODALIDAD PARA DETERMINAR QUÉ GENES PERMITEN DISTINGUIR GRUPOS DE PACIENTES ONCOLÓGICOS.
Jose Ameijeiras-Alonso, Rosa M. Crujeiras y Alberto Rodríguez-Casal.

IV.5. O PROBLEMA DE TRIAGEM BAYESIANO DO PONTO DE VISTA MÉDICO.
Sandra Ramos, Antónia Amaral Turkman e Marília Antunes.

IV.6. A NEW APPROACH TO PREDICT CHANGES IN PHYSICAL CONDITION WITH APPLICATION TO PREVENTIVE MEDICINE.
Marcos Matabuena Rodríguez.

10:30 – 11:10. Aula Magna

Sesión Invitada I

Moderadora: **Raquel Menezes**.

SPATIO-TEMPORAL SPECIES DISTRIBUTION MODELLING.
David Conesa. Universitat de València.

11:30 – 12:30. Corredor Nivel 3

Sesión de Posters

Moderadoras: **Graciela Estévez e Irene Oliveira**.

3. INFORME SOBRE A XESTIÓN E O ESTADO ECONÓMICO DO COOPERATIVISMO AGRARIO EN GALICIA.
Higinio Mougán-Bouzón, Mª José Cabaleiro Casal, Mª José Ginzo Villamayor, Ruth Rodríguez Ferreiros, David Vila López, Ricardo Francos Gendra, Sandra Rey Louzán.

5. STATISTICAL TECHNIQUES FOR SENSORIAL CHARACTERIZATION OF WINES FROM THREE DIFFERENT GRAPE VARIETIES OF VINHO VERDE REGION.
Catarina Marques, Alice Vilela, Elisete Correia.

6. ESCOLHA DE TRATAMENTOS OTIMIZADOS NA CONSTRUÇÃO DE FATORIAIS FRACIONÃRIOS.
Claudiney Nunes de Lima.

8. SAMPLE SIZE REQUIRED FOR ESTIMATING THE MEAN PARASITE ABUNDANCE.
Svitlana Shvydka, Volodimir Sarabeev, Vicente D. Estruch, Carmen Cadarso-Suárez.

9. SPATIAL MODELLING OF THE ABUNDANCE OF *Culex Pipiens* IN PORTUGAL - A BAYESIAN APPROACH.
P. de Zea Bermúdez, Marília Antunes, Maria da Conceição Proença, Maria Teresa Rebelo, Hugo Osório and Maria João Alves.

10. APLICACIÓN DE REDES BAYESIANAS ÁS RELACIÓNS ENTRE ESPECIES VEXETAIS DESPOIS DE INCENDIO FORESTAL E A SÚA SENSIBILIDADE Ó TAMAÑO DAS UNIDADES DE MOSTRAXE.
García-Duro J, Álvarez R, Basanta M, Casal M, Reyes O.

11. USO DE GRANDES MASAS DE DATOS PARA ESTUDOS EN ECOLOXÍA.
Antón Blanco Vilariño, Óscar Lado Baleato.

12. USING RIPLEY'S K FUNCTION TO ANALYSE SPATIAL TERMITE MOUNDS PATTERNS: A CASE STUDY.
Ana J. Righetto, Djair D.R. Frade, Luiz R. Nakamura, Thiago G. Ramires, Ezequiel A.L. Bautista and Paulo J. Ribeiro Junior.

13. A INFLUÊNCIA DO ÍNDICE DE OSCILAÇÃO DO ATLÂNTICO NORTE SOBRE AS ADMISSÕES HOSPITALARES POR MEIO DOENÇAS DO APARELHO CIRCULATÓRIO EM LISBOA.
Giovani L. Silva, Ricardo Almendra, Paula Santana, João Vasconcelos, Fábio Gonçalves e Tércio Ambrizzi.

14. ANÁLISE COMPARATIVA DE RESULTADOS DE ESPETROMETRIA GAMA PORTÁTIL EM AFLORAMENTOS ROCHOSOS.
Flávio Dias, Marco Lima, Jorge Sanjurjo-Sánchez e Carlos Alves.

15. DETECCIÓN DE ATÍPICOS ESPACIALES: UNA APLICACIÓN A DATOS DE BIOMONITORIZACIÓN.
Andrea Meilán-Vila y Rosa M. Crujeiras.

16. INFLUENCIA DE LAS VARIABLES DASOMÉTRICAS EN LA PROBABILIDAD DE INCENDIOS PARA LA ESPECIE *Pinus Pinaster* EN GALICIA.
Ginzo-Villamayor, M.J., Iglesias-Allones, J.L. y Marey-Perez, M.F.

17. MODELAÇÃO ESPACIAL APLICADA A DADOS DE PRECIPITAÇÃO EXTREMA NA ILHA DA MADEIRA.
Prata Gomes, D., Gouveia-Reis, D. e Neves, M.M.

18. MODELLING ENVIRONMENTAL MONITORING DATA COMING FROM DIFFERENT SURVEYS.
L. Margalho, R. Menezes and I. Sousa.

19. SEARCHING FOR SOIL POLLUTION SOURCES BY MEANS OF GEOGRAPHICALLY WEIGHTED PRINCIPAL COMPONENTS ANALYSIS (GWPCA).
Javier Roca-Pardiñas, Celestino Ordóñez, Tomás R. Cotos-Yáñez and Rubén Pérez-Álvarez.

20. STRUCTURAL TIME SERIES MODELING: AN APPLICATION TO ENVIRONMENTAL VARIABLES.
Marco Costa, Olexandr Baturin and A. Manuela Gonçalves.

21. UNA APROXIMACIÓN PARA DETERMINAR GRUPOS DE MÚLTIPLES CURVAS DE REGRESIÓN. APLICACIÓN AL ESTUDIO DE LA GEOMETRÍA DE UN TÚNEL.
Nora M. Villanueva, Marta Sestelo, Celestino Ordóñez y Javier Roca-Pardiñas.

22. USO DE MÉTODOS BIOMÉTRICOS EN TAXONOMÍA VEGETAL: EL CASO DE LA SUBTRIBU *Anthoxanthinae* (*Poaceae*).
Lema-Suárez I., Sahuquillo E., Estévez G. y Pimentel M.

23. TÉCNICAS NO PARAMÉTRICAS EN EL RIESGO SÍSMICO: ESTIMACIÓN DE PROBABILIDADES DE MAGNITUDES DE TERREMOTOS EN EL SUR DE LA PENÍNSULA IBÉRICA.
Quintela-del-Río, Alejandro y Francisco-Fernández, Mario.

24. EVOLUCIÓN ESPACIAL DOS SUICIDIOS EN GALICIA.
G Naveira Barbeito, MI Santiago Pérez e A Malvar Pintos.

25. INCIDENCIA DE CÁNCER EN PACIENTES VIH ¿ES MAYOR EN PACIENTES COINFECTADOS POR EL VIRUS DE LA HEPATITIS C?
S Pértega Díaz, MT Seoane Pillado, I Rodriguez Osorio, A Castro, H Meijide, E Poveda, A Mena de Cea.

26. INCIDENCIA DE LESIONES MEDULARES EN LA COMUNIDAD AUTÓNOMA GALLEGA EN 1995-2014.
V Balboa Barreiro, A Montoto Marqués, ME Ferreiro Velasco, S Salvador de la Barrera, A Rodríguez Sotillo, S Pértega Díaz y MT Seoane Pillado.

27. INFLUÊNCIAS DAS VARIAÇÕES CLIMÁTICAS NA INCIDÊNCIA DE CASOS DE DENGUE NA CIDADE DE GOIÂNIA, BRASIL.
Antônio Neco, Raquel Menezes e Susana Faria.

28. SAMPLE SIZE CALCULATORS FOR ESTIMATING ANTIMALARIAL SEROCONVERSION RATE AND ITS PUTATIVE CHANGES OVER TIME.
Nuno Sepúlveda, Carlos Daniel Paulino and Chris Drakeley.

30. ANÁLISE DE TEMPO ATÉ RECIDIVA EM CANCRO DA MAMA.
Liliana Coelho, Ana Borges, Luís Castro e Inês Sousa.

31. ASOCIACIÓN DE LOS HAPLOGRUPOS MITOCONDRIALES CON LA PROGRESIÓN RADIOGRÁFICA DE LA ARTROSIS DE RODILLA.
I Rego Pérez, S Pértega Díaz, M Fernández Moreno, A Soto Hermida, ME Vázquez Mosquera, E Cortés Pereira, S Relaño, N Oreiro Villar, C Fernández López, FJ Blanco.

32. CRIB vs SNAPPE II - AVALIAÇÃO ESPACIAL DO SEU DESEMPENHO ATRAVÉS DO COMP2ROC.
Maria Filipa Mourão, Ana Cristina Braga.

33. DISFUNÇÃO CARDÍACA EM RATOS: UMA ANÁLISE DE BAYESINA.
Magali T. R. Amaral, Maria C, Evangelista, Antônio C, Cicogna, Carlos R. Padovani.

34. EFECTO DEL LLENADO DE LA VEJIGA EN TRATAMIENTOS GINECOLÓGICOS DE BRAQUITERAPIA DE ALTA TASA.
Z Martín-Rodríguez, E Ambroa-Rey, MJ Ginzo-Villamayor, E Gónzalez-Patiño, J Mosquera-Seoane y M Pombar-Cameán.

35. IMPACT OF WORSENING RENAL FUNCTION IN SURVIVAL IN CRITICAL PATIENTS.
A. Baluja, M. López Ratón, P. Otero Castro, A. Iglesias, J. Álvarez Escudero y C. Cadarso Suárez.

36. INTERFAZ PARA EL CÁLCULO AUTOMÁTICO DE BLINDAJES EN BRAQUITERAPIA.
Z Martín-Rodríguez, I Taboada-Torres, M Giménez-Insua, J Guiu-Souto, M García-Sánchez, C Otero-Martínez, V Luna-Vega, M Pombar-Cameán y Mª José Ginzo-Villamayor.

37. MODELOS DE REGRESIÓN DISTRIBUCIONAL BIVARIANTE. APLICACIONES EN DIABETES.
Jenifer Espasandín-Domínguez, Carmen Cadarso-Suárez, Thomas Kneib, Nadja Klein, Francisco Gude.

38. EFECTO DE LOS IRRIGANTES DENTINARIOS EN LA CEMENTACIÓN ADHESIVA DE POSTES INTRARRADICULARES: ESTUDIO IN VITRO.
Sandra García Varela, María José Ginzo Villamayor, Benjamín Martín Biedma, João Carlos Tomás Ramos, Pedro Manuel Rivas Lombardero.

39. ESTIMAÇÃO DA DEGRADAÇÃO DE BIOPOLÍMEROS PARA USO ODONTOLÓGICO ATRAVÉS DE MODELOS DE CRESCIMENTO.
Yaroslava Robles-Bykbaev, Salvador Naya, Javier Tarrío-Saavedra.

40. PREBED: SOFTWARE DE EVALUACIÓN ODONTOLÓGICA.
Lozano, D., Martínez, I., Limeres J., Roca, J., Diz P. y Cadarso C.

41. A GENERAL CLASS OF EXPONENTIATED SINH CAUCHY REGRESSION MODELS.
Ramires, T.G and Ortega, E.M.M.

42. DADOS DE ESPETROMETRIA GAMA PORTÁTIL E TIPOLOGIA DE ESPAÇOS INTERIORES.
Marco Lima, Jorge Sanjurjo-Sánchez e Carlos Alves.

44. DINÂMICA POPULACIONAL MENDELIANA.
Maria de Fátima Brilhante, Dinis Pestana, Paulo Semblano e Fernando Sequeira.

45. MODELOS EXTREMAIS EM DINÂMICA DE POPULAÇÕES.
Maria de Fátima Brilhante, Maria Ivette Gomes e Dinis Pestana.

46. SCREENING AND VALIDATION IN PERSSILAA - THE PERSONALIZED ICT SERVICE FOR INDEPENDENT LIVING AND ACTIVE AGEING.
Joana Fernandes, Antónia Turkman, Marília Antunes, Feridun Turkman, Stephanie Jansen-Kosterink, Javier Solana, Enrique Gómez, Amelia Rauter and Miriam Vollenbroek.

12:30 – 13:30. Aula Magna

Sesión Plenaria II

Moderadora: **Rosa M. Crujeiras Casais**.

ADDING SPATIAL INFORMATION TO TRADITIONAL STATISTICAL INFERENCE.
Daniela Cocchi. Università di Bologna.

15:00 – 15:40. Aula Magna

Sesión Invitada II

Moderadora: **Carmen Cadarso Suárez**.

WHAT IS THE DOSE OF RADIATION RECEIVED? STATISTICAL ADVANCES IN BIODOSIMETRY.
Pere Puig. Universitat Autònoma de Barcelona.

15:40 – 17:00. Aula 7

Sesión de Orais V: Métodos estatísticos en Agricultura e Bioloxía

Moderadora: **Manuela Neves**.

V.1. MODELO NÃO-LINEAR MISTO APLICADO À EVAPOTRANSPIRAÇÃO DA ALFACE.
Omar Pereira, Altair Bertonha e Isolde Previdelli.

V.2. OPTIMIZACIÓN DEL ESFUERZO TAXONÓMICO MEDIANTE LA APLICACIÓN DE TÉCNICAS ESTADÍSTICAS.
Andrea Freijeiro, Carola Gómez-Rodríguez y Andrés Baselga.

V.3. AN INTEGRATED POPULATION DYNAMICS MODEL FOR WILD TURKEYS.
Russell Alpizar-Jara.

15:40 – 17:00. Aula 8

Sesión de Orais VI: Epidemioloxía I

Moderador: **Carlos L. Iglesias Patiño**.

VI.1. CURRENT STATISTICAL APPROACHES TO ANTIBODY DATA ANALYSIS OF INFECTIOUS DISEASES ON THE BRINK OF ELIMINATION.
Nuno Sepúlveda, Gillian Stresman, Michael White and Chris Drakeley.

VI.2. USO DOS MODELOS IDADE-PERÍODO-COHORTE NO ESTUDO DA MORTALIDADE E INCIDENCIA POR INFARTO AGUDO DE MIOCARDIO EN GALICIA.
G Naveira Barbeito, M Pérez Ríos e A Malvar Pintos.

VI.3. ANALYSIS OF PANCREATIC CANCER IN SPANISH PROVINCES COMBINING BAYESIAN APC MODELS AND DISEASE MAPPING APPROACHES.
Etxeberria J., López-Abente G., Riebler A., Goicoa T. and Ugarte MD.

VI.4. MODELAÇÃO ESPAÇO-TEMPORAL DE DADOS DE CRIMES EM PORTUGAL.
Conceição Ribeiro, Paula Pereira.

15:40 – 17:00. Aula 9

Sesión de Orais VII: Métodos estatísticos en Medicina II

Moderador: **Pedro Oliveira**.

VII.1. PROFILING PROTEÓMICO DE ISOLADOS DE *ESCHERICHIA COLI* E *ENTEROCOCCUS SPP*. ATRAVÉS DA ESPECTROMETRIA DE MASSA.
Tiago Santos, Irene Oliveira, José Luis Capelo, Hugo M. Santos, Catarina Marinho, Alexandre Gonçalves, José Eduardo Araújo, Patrícia Poeta, Gilberto Igrejas.

VII.2. ESTIMAÇÃO DE EQUAÇÕES DE REFERÊNCIA PARA PARÂMETROS ESPIROMÉTRICOS PARA A POPULAÇÃO DE VILA FRANCA DE XIRA.
Carina Silva, Sandra Carvalho, Lisete Sousa.

VII.3. MODELING LONG-MEMORY AND VOLATILITY IN HEAT RATE VARIABILITY.
Argentina Leite, Ana Paula Rocha, Maria Eduarda Silva.

VII.4. PERFORMANCE OF ADAPTIVE DESIGNS IN DOSE FINDING EXPERIMENTS.
Nancy Flournoy, José Moler, Fernando Plo.

VII.5. ECONOMICAL EVALUATION OF COMPUTED TOMOGRAPHY ANGIOGRAPHY (CTA) VERSUS CONVENTIONAL ANGIOGRAPHY (CA) TO DIAGNOSE CORONARY ISCHEMIA.
Reyes-Santías, Francisco, Cadarso-Suárez, Carmen, Espasandín, Jenifer.

17:30 – 18:30. Aula Magna

Sesión Plenaria III

Moderador: **Wenceslao González Manteiga**.

A CHARACTERIZATION OF MISSINGNESS AT RANDOM IN A GENERALIZED SHARED-PARAMETER JOINT MODELING FRAMEWORK FOR LONGITUDINAL AND TIME-TO-EVENT DATA, AND SENSITIVITY ANALYSIS.

Geert Molenberghs. Hasselt University.

Sábado 2 de xullo

10:20 – 11:00. Aula Magna

Sesión Invitada III

Moderadora: **Mª José Ginzo Villamayor**.

STATISTICAL CHALLENGES IN AIR POLLUTION HEALTH RISK ASSESSMENT.

Michela Cameletti. Università di Bergamo.

11:30 – 12:30. Aula 7

Sesión de Orais VIII: Epidemioloxía II

Moderadora: **Sonia Pértega Díaz**.

VIII.1. ADJUSTED WALD CONFIDENCE INTERVALS IN THE META-ANALYSIS OF PREVALENCE PROPORTIONS.
Vera Afreixo, Sara Escudeiro and Adelaide Freitas.

VIII.2. MODELOS DE REGRESSÃO LOGÍSTICA SOB DELINEAMENTOS AMOSTRAIS COMPLEXOS.
Isabel Natário, Sheyla Rodrigues Cassy.

VIII.3. DEMOGAL: PAQUETE PARA PROYECTAR FLUJOS Y STOCKS DE POBLACIÓN.
Calvo Ocampo, Mª Esther, Iglesias Patiño, Carlos, Gómez Duaso, Juan, López Vizcaíno, Mª Esther, Martín Vila, María y Silveira Calviño, Solmary.

VIII.4. ESPERANZAS DE VIDA SEN LIMITACIÓNS EN GALICIA.
Iglesias Patiño, C.L., López Vizcaíno, M.E., Martín Vila, M., Calvo Ocampo, E. e Santiago Pérez, M.I.

11:30 – 12:30. Aula 8

Sesión de Orais IX: Métodos estatísticos en Ecoloxía

Moderador: **Miguel Blanco Vázquez**.

IX.1. UM MODELO PARA A ABUNDÂNCIA DE PEIXES COM DEPENDÊNCIA ESPACIAL BASEADA NO NÚMERO DE CONFLUÊNCIAS ENTRE RIACHOS.
Júlio César Pereira, Maurício Cetra e Martin Ridout.

IX.2. UN NUEVO MÉTODO PARA SEPARAR COMPONENTES DE LA DISIMILITUD ENTRE MÚLTIPLES COMUNIDADES BIOLÓGICAS BASADA EN DATOS DE ABUNDANCIA: VARIACIÓN EQUILIBRADA VS. GRADIENTES DE ABUNDANCIA.
Andrés Baselga.

IX.3. OS PATRÓNS NA VARIACIÓN LATITUDINAL DA DIVERSIDADE BETA EN VERTEBRADOS SON SEMELLANTES ENTRE CLADOS E SUXIREN CAUSAS HISTÓRICAS COMÚNS.
Adrián Castro-Insua, Carola Gómez-Rodríguez e Andrés Baselga.

IX.4. PREDICCIÓN DEL PATRÓN DE ASENTAMIENTO LARVARIO EN LAS RÍAS GALLEGAS CON GAM PARA DATOS REALES Y FUNCIONALES.
Isabel Fuentes-Santos, Uxío Labarta, X. Antón Álvarez-Salgado y Mª José Fernández-Reiriz.

11:30 – 12:30. Aula 9

Sesión de Orais X: Modelización conxunta

Moderador: **Carlos Daniel Paulino**.

X.1. MODELAGEM CONJUNTA DE DADOS LONGITUDINAIS E SOBREVIVÊNCIA EM PACIENTES COM SEPSE.
Omar Pereira, Marcos Peres, Cristina Kuroda, Elza Kimura, Isolde Previdelli.

X.2. MODELAÇÃO CONJUNTA DE DADOS LONGITUDINAIS E DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE RISCOS COMPETITIVOS: AVALIAÇÃO DE UM PROGRAMA DE DIÁLISE PERITONEAL.
Laetitia Teixeira, Ipek Guler, Carmen Cadarso-Suarez, Inês Sousa, Anabela Rodrigues e Denisa Mendonça.

X.3. JOINT MODELLING OF LONGITUDINAL AND SURVIVAL DATA ON BREAST CANCER.
Ana Borges, Inês Sousa e Luís Castro.

12:30 – 13:30. Aula Magna

Sesión Plenaria IV

Moderadora: **Mª Antónia Turkman**.

ALGUMAS APLICAÇÕES DE EQUAÇÕES DIFERENCIAIS ESTOCÁSTICAS NA MODELAÇÃO DE FENÓMENOS BIOLÓGICOS EM AMBIENTE ALEATÓRIO.

Carlos Braumann. Universidade de Évora.

15:00 – 15:40. Aula Magna

Sesión Invitada IV

Moderador: **Ricardo Cao Abad**.

LA AGRICULTURA DEL SIGLO XXI: NUEVOS RETOS PARA LA BIOMETRÍA.

José Luis G. Andújar. Consejo Superior de Investigaciones Científicas.

15:40 – 17:00. Aula Magna

Mesa Redonda

Moderador: **José Antonio Vilar Fernández**.

EL PAPEL DE LA ESTADÍSTICA EN LAS CC. DE LA SALUD, LA ECOLOGÍA Y EL MEDIOAMBIENTE: REFLEXIONES SOBRE LA SITUACIÓN ACTUAL Y NUEVOS RETOS.

David Conesa. Universitat de València.
José Miguel Pereira. Universidade de Lisboa.
Luzia Gonçalves. Universidade Nova de Lisboa.
Xurxo Mariño. Universidade da Coruña.

Minicurso

II Galician-Portuguese Meeting of Biometry,
Santiago de Compostela, 30 June – 2 July 2016

MINI-COURSE ON SURVIVAL ANALYSIS

Luis Filipe Meira Machado[1]

[1]Universidade do Minho

Survival analysis describes the various methods for modelling the times from an initiating event to some terminal event. Survival analysis is prominently used in the biomedical sciences, but they are also used in many other fields such as the social sciences, engineering and economics.

A special feature of survival data is the presence of incomplete observations. Among other reasons, this may occur because the event of interest is not observed before the end of the study or because the subject leaves the study before an event occurs. This phenomenon is called censoring and it requires special statistical methods.

This course introduces general statistical concepts and methods for the analysis of failure time data. Different types of censored data will be introduced and techniques for estimating the survival function will be illustrated. The use of the log-rank test as a technique for testing equality of survival functions is also discussed. Finally, regression models based on the Cox proportional hazards model will also be studied. The course requires participants to use the R statistical package to analyze survival analysis data.

AIMS

- To provide an introduction to survival analysis;
- To illustrate how the methods can be implemented using R, a free software environment for statistical computing and graphics.

CONTENTS

- An overview of survival analysis methods
- Censoring
- Survival and hazard functions
- Estimation of the survival function: Kaplan-Meier
- The Log Rank test
- Introduction to the Cox Proportional Hazards (PH) model
- The Cox model with time-dependent covariates

Sesións plenarias

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

STRUCTURED ADDITIVE CONDITIONAL COPULA REGRESSION MODELS

Thomas Kneib[1]

[1]University of Göttingen

ABSTRACT

While most regression models focus on explaining aspects of one single response variable alone, interest in modern statistical applications has recently shifted towards simultaneously studying of multiple response variables as well as their dependence structure. A particularly useful tool for pursuing such an analysis are copula-based regression models since they enable the separation of the marginal response distributions and the dependence structure summarised in a specific copula model.

However, so far copula-based regression models have mostly been relying on two-step approaches where the marginal distributions are determined first whereas the copula structure is studied in a second step after plugging in the estimated marginal distributions. Moreover, the parameters of the copula are mostly treated as a constant not related to covariates and most regression specifications for the marginals are restricted to purely linear predictors. We therefore propose simultaneous Bayesian inference for both the marginal distributions and the copula using computationally efficient Markov chain Monte Carlo simulation techniques. In addition, we replace the commonly used linear predictor by a generic structured additive predictor comprising for example nonlinear effects of continuous covariates, spatial effects or random effects and also allow to make the copula parameters covariate-dependent. To facilitate Bayesian inference, we construct proposal densities for a Metropolis Hastings algorithm relying on quadratic approximations to the full conditionals of regression coefficients, thereby avoiding manual tuning and thus providing an attractive adaptive algorithm. The flexibility of Bayesian conditional copula regression models is illustrated in an application on childhood undernutrition.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

ADDING SPATIAL INFORMATION TO TRADITIONAL STATISTICAL INFERENCE

Daniela Cocchi[1]

[1]Università di Bologna

ABSTRACT

Several features of traditional statistical inference can be strengthened by a proper use of spatial information. Two examples are illustrated. The first one is about finite population inference; the talk deepens the concepts of spatial information, geography and prediction in this context. The presentation illustrates a design based individual predictor and how the well-known geographically weighted regression technique can be seen as a version of the Generalized Regression estimation (GREG), at present a very popular tool in model-assisted design-based inference in finite populations. An estimator for the population total is developed, whose statistical properties are derived. A Monte Carlo simulation study highlights the advantages of the proposal. The second example presents a relatively recent research field, which aims at accounting for space in entropy measures, as a generalization when the spatial location of the events is relevant to the purposes of the analysis. The main limitation is that all indices are computed conditional on a chosen distance. Space is exogenous, therefore the bivariate properties cannot be exploited. The inclusion of spatial components in entropy indices induces to investigate the characteristics of the quantity known as conditional entropy in order to include space as a second dimension in the analysis. This way, previously proposed indices such as the contagion index and the co-occurrence based entropy can be extended to a bivariate context, in a setting where the probabilistic meaning of all components is well defined. As a direct consequence, we also obtain an index satisfying the additivity property, as the global conditional entropy is a sum of local entropy measures.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

A CHARACTERIZATION OF MISSINGNESS AT RANDOM IN A GENERALIZED SHARED-PARAMETER JOINT MODELING FRAMEWORK FOR LONGITUDINAL AND TIME-TO-EVENT DATA, AND SENSITIVITY ANALYSIS

Geert Molenberghs[1]

[1]Universiteit Hasselt

ABSTRACT

Rubin (1976) classified missing data as MCAR, MAR, and MNAR. Given that models for missing data often make unverifiable assumptions about the missing value mechanism, sensitivity analysis is needed. In joint modeling of time-to-event and longitudinal data, similar issues arise. The longitudinal covariate may be measured with error, its values are likewise only available at the specific time points at that the patient appears at the clinic for longitudinal measurements, and the time-to-event may also be censored.

Undeniably, there is a strong connection between the missing data and the joint longitudinal and time-to-event settings, the theme of this work. We build an extended shared random effects joint model, similar in spirit to that of Creemers et al (2011). An added layer of complexity is that data can be coarsened in various ways: the longitudinal sequence can be incomplete; the time-to-event outcome can be censored; both of these can occur simultaneously. Within the extended framework, we provide a characterization of MAR, consistent to the one in the missing data setting, and juxtapose it with more conventional joint models. This opens routes for sensitivity analysis.

References

[1] Njeru Njagi, E., Molenberghs, G., Kenward, M.G., Verbeke, G., and Rizopoulos, D. (2014). A Characterization of Missingness at Random in a Generalized Shared-parameter Joint modelling Framework for Longitudinal and Time-to-Event Data, and Sensitivity Analysis. *Biometrical Journal*, 56, 1001-1015.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

ALGUMAS APLICAÇÕES DE EQUAÇÕES DIFERENCIAIS ESTOCÁSTICAS NA MODELAÇÃO DE FENÓMENOS BIOLÓGICOS EM AMBIENTE ALEATÓRIO

Carlos Alberto Dos Santos Braumann[1]

[1]Universidade de Évora

RESUMO

Alguns fenómenos biológicos têm uma dinâmica que é afetada por flutuações aleatórias das condições ambientais, podendo assim ser descrita por uma equação diferencial estocástica (EDE) com um termo que descreve a taxa média de variação e outro que descreve as perturbações nela causadas pelas flutuações ambientais.

Nesta sessão vamos apresentar a aplicação destes modelos em várias áreas, o que inclui também resultados obtidos em colaboração com vários coautores como Patrícia A. Filipe, Clara Carlos e Nuno M. Brites:

a) Modelos de crescimento de populações animais, com incidência no comportamento qualitativo de modelos gerais quanto à extinção da população e à existência de densidade estacionária; uso de modelos específicos aproximados e os efeitos nas previsões dos erros cometidos nas aproximações.

b) Modelos de pesca, com estudo da abordagem de usar políticas de pesca sustentáveis e aplicáveis que consistem em usar um esforço de pesca constante e otimizar, em regime estacionário, o lucro esperado por unidade de tempo; comparação com as políticas inaplicáveis de esforço variável que otimizam o lucro esperado acumulado descontado.

c) Modelos de crescimento individual de um animal com aplicação aos bovinos mertolengos e determinação da idade de abate que otimiza o lucro esperado; extensão da metodologia não paramétrica de estimação para o caso de dados de múltiplas trajetórias (múltiplos animais).

AGRADECIMENTOS

Carlos A. Braumann é membro do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT).

Sesións invitadas

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

SPATIO-TEMPORAL SPECIES DISTRIBUTION MODELLING

David Valentín Conesa Guillén[1]

[1]Universitat de València

ABSTRACT

Methodological approaches for modelling species distribution patterns have been lately proposed. Our approach makes use of hierarchical Bayesian models along with geographical and environmental characteristics at each sampling location. Maps of predicted probabilities of presence (or the abundance of the species) can then be generated using Bayesian kriging. Our interest here is to describe how to use the integrated nested Laplace approximation jointly with the Stochastic Partial Differential Equation approach to perform fast inference and prediction in such complex models. Moreover, our intention is to show that many extensions can also be easily implemented within this framework. In particular we will present practical examples involving preferential sampling, misalignment issues, spatio-temporal structures, etc.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

WHAT IS THE DOSE OF RADIATION RECEIVED? STATISTICAL ADVANCES IN BIODOSIMETRY

Pere Puig Casado[1]

[1]Universitat Autònoma de Barcelona

ABSTRACT

In the event of a radiation accident, biological dosimetry is essential for the timely determination of the radiation dose received by an exposed individual. The dose is estimated quantifying the amount of damage induced by radiation at a cellular level, for instance by counting the number of chromosome aberrations like dicentrics or micronuclei. The theory of count data distributions is essential for this aim. In this talk we introduce the standard statistical methodology for dose estimation described in the manual of the International Atomic Energy Agency (IAEA, 2011), and we also summarize the recent research led by our team. We present some models based on compound Poisson processes that are suitable to describe high-LET radiation exposures like those presented in the accident of Fukushima, models based on weighted Poisson distributions to integrate low and high doses and some Bayesian-type methods for Poisson inverse regression which have been implemented in the R package RADIR.

References

[1] Higueras M, Puig P, Ainsbury, EA, Rothkamm, K (2015). *A new Inverse Regression Model Applied to Radiation Biodosimetry.* Proceedings of the Royal Society A, 471:2174 , DOI: 10.1098/rspa.2014.0588

[2] IAEA (2011). *Cytogenetic Dosimetry: Applications in Preparedness for and Response to Radiation Emergencies.* International Atomic Energy Agency: Vienna.

[3] Moriña, D, Higueras M, Puig P, Ainsbury, EA, Rothkamm, K (2015). radir package: an R implementation for cytogenetic biodosimetry dose estimation. *Journal of Radiological Protection*, 35, p. 557–569.

[4] Puig, P., Barquinero, F. (2011). An Application of Compound Poisson Modelling to Biological Dosimetry. *Proceedings of the Royal Society A*, 467, 897-910.

[5] Pujol M, Barquinero JF, Puig P, Puig R, Caballín MR, Barrios L (2014). A New Model of Biodosimetry to Integrate Low and High Doses, *PlosOne*, 1-19, DOI: 10.1371/journal.pone.0114137.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

STATISTICAL CHALLENGES IN AIR POLLUTION HEALTH RISK ASSESSMENT

Michela Cameletti[1]

[1]Università di Bergamo

ABSTRACT

Air pollution is both an environmental and a social problem, as it leads to a multitude of adverse on human health, ecosystems and the climate. Starting from the late 1980s a substantial literature has emerged on health effects from exposure to air pollution, which escalated at the end of 2013 with a study by the International Agency for Research on Cancer which found sufficient evidence to classify air pollution as a leading environmental cause of cancer deaths. From the statistical point of view, studying the association between health and air pollution poses multiple challenges. First, data regarding pollutant concentration and health outcomes (i.e. mortality/morbidity data) are spatially misaligned, giving rise to the so-called change of support problem. Secondly, air pollution monitoring networks can be subject to selection bias if monitoring sites are not randomly distributed but are located in the more polluted areas, an issue referred to as preferential sampling. Finally, air pollution exposure estimate in ecological studies may suffer from potential confounding and measurement error. In this work we consider all these challenging issues while implementing a spatio-temporal model for air pollution health risk assessment, both through a simulation study and a real case application regarding drug prescriptions in England.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

LA AGRICULTURA DEL SIGLO XXI: NUEVOS RETOS PARA LA BIOMETRÍA

José Luis González Andújar[1]

[1]Consejo Superior de Investigaciones Científicas

RESUMEN

La interrelación entre la agricultura y la Biometría se remonta a principios del siglo XX. La agricultura ha jugado un papel impulsor en el desarrollo de la biometría. Se puede afirmar en este sentido que la agricultura ha jugado el mismo papel como motor del desarrollo de la estadística, que la física en relación con las matemáticas. La agricultura ha cambiado substancialmente en los últimos cien años. La agricultura del siglo XXI se enfrenta al reto de alimentar a una población creciente de seres humanos dentro de un contexto de cambio climático, escasez de tierra cultivable y pérdida de la biodiversidad agrícola. Conseguir el incremento sostenible de la producción y el uso óptimo de los recursos agrarios requiere, en muchos casos, la necesidad de plantear nuevos enfoques estadísticos que permitirán la evolución de la agricultura y la biometría. Se presenta una breve visión de la relación histórica entre biometría y agricultura y posibles perspectivas de evolución común de ambas materias.

Mesa redonda

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

EL PAPEL DE LA ESTADÍSTICA EN LAS CIENCIAS DE LA SALUD, LA ECOLOGÍA Y EL MEDIOAMBIENTE: REFLEXIONES SOBRE LA SITUACIÓN ACTUAL Y NUEVOS RETOS.

Oradores:

- ***David Valentín Conesa Guillén***
  Universidad de Valencia
  David.V.Conesa@uv.es
- ***Luzia Gonçalves***
  Universidade Nova de Lisboa
  luziag@ihmt.unl.pt
- ***José Miguel Pereira***
  Universidade de Lisboa
  jmocpereira@gmail.com
- ***Xurxo Mariño***
  Universidade da Coruña
  xurxo@udc.es

Moderador:

- ***José A. Vilar***

Universidade da Coruña
jose.vilarf@udc.es

RESUMEN

Actualmente ya nadie duda del papel relevante que la Estadística tiene en el desarrollo de un gran número de disciplinas, incluyendo desde luego los tres ámbitos científicos englobados en este **II Encuentro Galaico-Portugués de Biometría.** Este hecho irrefutable motiva el interés sobre cuál es la situación actual de la Estadística y de los estadísticos en el contexto de estas áreas, sobre cómo se desenvuelve esta interacción interdisciplinar, no solo al nivel de colaboración en tareas de investigación, sino también en aspectos como la formación específica requerida o los nuevos retos de aplicación de la estadística que se perciben como fundamentales a corto y a medio plazo.

Los oradores participantes en esta mesa redonda acreditan años de experiencia en la aplicación de técnicas estadísticas en alguna de estas tres áreas que nos ocupan y gozan de prestigio internacional en las mismas. En sus intervenciones ilustrarán a los asistentes acerca de tópicos de gran interés en los que han trabajado y que son de especial relevancia en estos ámbitos, haciéndonos partícipes de su valoración personal sobre el presente y el futuro de los equipos interdisciplinares incluyendo estadísticos.

El profesor David Conesa, en su condición de docente en el Máster de Bioestadística de la Universitat de Valencia y con una intensa actividad investigadora de estadístico desarrollada en los últimos años en muchas áreas de conocimiento relacionadas con las ciencias de la salud y la vida, nos dará su opinión sobre el papel de la Estadística en las CC. de la Salud, la Ecología y el Medioambiente. Discutirá acerca de la necesidad formativa de los graduados en ciencias aplicadas, particularizando en el tipo de formación que van buscando los estudiantes del Máster antes mencionado. Esta formación y los intereses que demuestran estos estudiantes le servirán para valorar qué ha encontrado cuando ha tenido que trabajar con profesionales de otras áreas que buscaban aplicar la Estadística. Cerrará su intervención con una reseña de hacia dónde vamos encaminados los estadísticos, la modelización estocástica.

En el contexto de las Ciencias de la Salud y las Ciencias Biomédicas, la profesora Luzia Gonçalves ilustrará aspectos importantes de la investigación en el área tropical, donde concurren especificidades que urge explorar. Los recientes casos de ébola, dengue o Zika dan idea de la fuerte presión para el registro de datos de calidad, de forma rápida y en tiempo real, en la búsqueda de una actuación eficaz. Los contextos de pobreza y/o las particularidades culturales, por ejemplo en algunos países africanos, convierten esta

recogida de datos en una tarea especialmente desafiante. Con el desarrollo de la investigación y de la enseñanza en África, la salud tropical ofrece nuevos retos y oportunidades a los estadísticos, de manera que la interdisciplinariedad está naturalmente asegurada en esta área.

En el ámbito del medioambiente, el Prof. José Miguel Pereira de la Universidad de Lisboa ha desarrollado gran parte de su investigación en detección remota de áreas quemadas y en pirogeografía, y nos proporcionará su visión sobre el papel de la estadística en este contexto. Finalmente, el Prof. Xurxo Mariño, investigador en neurobiología y con numerosas colaboraciones en el desarrollo de herramientas estadísticas para su aplicación a este campo, nos dará su visión personal incidiendo asimismo en los aspectos de divulgación de la Ciencia en los que ha publicado varios libros.

Además de estas intervenciones, es objetivo de la mesa establecer un debate enriquecedor abierto a todos los asistentes donde se traten, entre otras, cuestiones como las que siguen.

- ¿Está bien establecido y bien entendido el rol de un estadístico en un equipo interdisciplinar en estos campos? ¿Es este papel relevante o tal vez prescindible habida cuenta del nivel de formación en análisis de datos, cada vez más elevado, de investigadores de estas áreas?
- ¿En qué medida se integra un profesional de la Estadística en problemas y desafíos propios de estas áreas? ¿Dónde están las principales dificultades?
- ¿Qué consideración existe en los ámbitos de CC de la Salud/Medioambiente/Ecología acerca de la Estadística y de los estadísticos?
- ¿Está en ascenso el número de equipos de investigación en estas áreas que incluyen a estadísticos? ¿Qué medidas permitirían fomentar esta interdisciplinariedad?
- ¿Es “interesante” para un estadístico involucrarse en retos de investigación en estos campos?
- ¿Dónde publicar? ¿Tienen cabida contribuciones de alto nivel estadístico en revistas de impacto de estas áreas? ¿Cuál es el nivel de innovación que se requiere? ¿Están convenientemente valoradas estas publicaciones en el CV de un estadístico?
- ¿Son adecuadas las asignaturas de contenido estadístico en los diseños curriculares actuales de titulaciones de estos ámbitos? ¿Qué perfil formativo debe de tener un estadístico para involucrarse en un equipo de investigación en Medioambiente, Ecología o CC. de la Salud? ¿Son requeridos actualmente?

Comunicacións orais

Sesión: Premio Ciencias da Saúde

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Classificação hierárquica com distribuições *a posteriori*: Um estudo de simulação em padrões temporais de VIH

Diana Rocha[1], Sónia Gouveia[1,2], Carla Pinto[3,4], Manuel Scotto[5], João Nuno Tavares[3]

[1] Centro de Investigação e Desenvolvimento em Matemática e Aplicações - CIDMA, Univ de Aveiro
[2] Instituto de Engenharia Electrónica e Informática de Aveiro - IEETA, Universidade de Aveiro
[3] Centro de Matemática da Universidade do Porto - CMUP, Universidade do Porto
[4] Instituto Superior de Engenharia do Instituto Politécnico do Porto - ISEP
[5] CEMAT e Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa

RESUMO

O objetivo deste trabalho é avaliar diferentes abordagens para identificação de grupos de pacientes VIH com padrões temporais de evolução da doença similares. Foi considerado um sistema de equações diferenciais ordinárias para caracterizar a comportamento ao longo do tempo de um paciente VIH sob tratamento antiretroviral - TAR de longo prazo, com 5 parâmetros estimados a partir de metodologia Bayesiana. As distribuições *a posteriori* foram usadas para quantificar distâncias (univariadas) entre pacientes, através do valor médio da distribuição *a posteriori*, e considerando a distância entre as distribuições *a posteriori* para cada parâmetro. O resultado do agrupamento hierárquico obtido pelas duas abordagens sugere que o uso de uma distância que considere a distribuição *a posteriori* é preferível. Trabalho futuro irá considerar distâncias multivariadas em vez de distâncias univariadas.

Palavras e frases chave: Vírus da Imunodeficiência Humana (VIH), Estimação Bayesiana, Modelos Matemáticos, Classificação Hierárquica.

1. INTRODUÇÃO

Não existem estudos em Portugal que permitam a identificação de grupos de pacientes portadores de VIH, com padrões temporais de evolução da doença semelhantes. A identificação de grupos, além de ter relevância ao nível da epidemiologia da doença, pode permitir também definir terapêuticas mais ajustadas a cada (grupo de) paciente(s) semelhantes e assim conseguir atender simultaneamente à individualização do tratamento e à gestão mais eficiente dos recursos hospitalares. Por este motivo, há todo o interesse em desenvolver uma metodologia que permita fazer a identificação de grupos de pacientes com trajetórias similares. Neste trabalho considerou-se o seguinte sistema EDO (equações diferenciais ordinárias) para caracterizar a trajetória de cada paciente portador de VIH sob tratamentos TAR de longo prazo [2]

$$\begin{aligned} \frac{dT_U(t)}{dt} &= \lambda - \rho T_U(t) - \eta(t) T_U(t) V(t) \\ \frac{dT_I(t)}{dt} &= \eta(t) T_U(t) V(t) - \delta T_I(t) \\ \frac{dV(t)}{dt} &= N\delta T_I(t) - cV(t) \end{aligned} \tag{1}$$

com variáveis de estado $CD4 \equiv T(t) = T_U + T_I$, respetivamente a concentração de células CD4 não infetadas e infetadas, e V como a carga viral. Os parâmetros do sistema são λ (taxa de proliferação das células CD4), ρ e δ (respetivamente taxa de mortalidade de células CD4 não infetadas e infetadas), $\eta(t)$ (taxa de infeção, que varia ao longo do tempo e depende da eficácia do TAR), N (número médio de virions produzidos por célula infetada) e c (taxa de depuração de viriões livres). Este trabalho foi desenvolvido em 3 etapas. Primeiro, foram geradas 20 réplicas de dois padrões distintos de evolução temporal de $T(t)$ e $V(t)$ [2]. Segundo, para cada réplica, foram estimados os parâmetros do modelo. Finalmente, foram identificadas classes de pacientes por agrupamento hierárquico, tendo em conta que seria esperado existirem duas classes do dendrograma, cada uma das classes constituída por um dos padrões de evolução temporal considerado. O agrupamento hierárquico foi considerado com duas métricas de distância: para cada parâmetro do modelo, consideraram-se distâncias sobre o valor médio e distâncias baseadas na distribuição *a posteriori*. Os resultados das diferentes abordagens foram finalmente comparados por intermédio da dispersão intra e entre grupos.

2. MÉTODOS EXPERIMENTAIS

Neste trabalho foram geradas observações igualmente espaçadas das séries temporais $T(t) = T_U(t)+T_I(t)$ e $V(t)$, por resolução numérica do sistema 1 no intervalo de tempo $[0, 20]$(dias) através do método de Runge-Kutta. Foram consideradas as condições iniciais $T_U(0) = 600$, $T_I(0) = 30$, $V(0) = 105$, $\eta(t) = 9 * 10^{-5}(1 - 0{,}9\cos(\pi t/1000))$ e valores de $(\lambda, \rho, \delta, N, c)$ correspondentes a [2]. Foram também considerados dois conjuntos distintos para os parâmetros $(\lambda, \rho, \delta, N, c)$ referenciados na Tabela 1, os quais representam curvas que correspondem ao comportamento de indivíduos reais, sejam indivíduo 1 e 2 [2]. No final deste processo foram constituídos dois conjuntos de 19 observações não igualmente espaçadas no intervalo de tempo $[0, 20]$(dias), treze no primeiro dia de tratamento (internamento) e uma medição a cada dois dias, do dia 2 ao dia 14 do tratamento (ambulatório), um para cada indivíduo. Neste trabalho foram consideradas 10 réplicas de cada indivíduo, cada réplica foi constituída adicionando ao conjunto original de 19 observações um erro de 19 observações de CD4 disponíveis, assumindo uma distribuição Normal de média nula e variância ajustada de acordo com o facto de que as medições laboratoriais de CD4 estão sujeitas a um erro de aproximadamente 20 % em relação ao valor medido [3]. Para cada uma das 20 réplicas geradas, os parâmetros $(\lambda, \rho, \delta, N, c)$ foram estimados por intermédio do algoritmo Metropolis-Hastings com condições iniciais iguais às usadas para a geração dos dados. Consideraram-se as distribuições *a priori* Gamma, Normal e Wishart [1].

As métricas de distância univariadas para avaliar semelhança entre pacientes foram obtidas a partir das distribuições *a posteriori*. Para cada parâmetro, foi considerada a distância Euclideana entre a média do parâmetro obtida para a réplica i e para a réplica j, isto é a) $d^2_{ij} = (\bar{x_i} - \bar{x_j})^2$, e também a semelhança entre a distribuição acumulada *a posteriori* de cada parâmetro da réplica i e da réplica j, sejam $F_i(x)$ e $F_j(x)$, dada por b) $d^2_{ij} = \int_0^1 (F_i^{-1}(y) - F_j^{-1}(y))^2 y(1-y)dy \simeq \sum_{k=1}^{99} (\hat{F}_i^{-1}(y_k) - \hat{F}_j^{-1}(y_k))^2 y_k(1-y_k)$ com $y_k = k/100$. Para esta última distância são considerados diferentes pesos para cada percentil $F^{-1}(y)$, onde diferenças na mediana terão maior peso $(y = 0{,}5)$. Esta distância é estimada por intermédio da aproximação do integral por um somatório ponderado de áreas de trapézios, obtidos numa escala com 99 pontos igualmente espaçados num intervalo de $[0; 1]$.

A avaliação das métricas de distância para cada par de réplicas $i, j = 1, 2, \ldots, 20$ originou uma matriz 20×20 utilizada para obter um dendrograma por agrupamento/linkage sobre o valor médio do grupo. As abordagens de classificação foram comparadas pela variabilidade estimada intra e entre grupos.

3. ANÁLISE DE RESULTADOS

A Tabela 1 apresenta medidas descritivas sobre as distribuições *a posteriori* para cada parâmetro e réplica. Foi possível observar que a ordem de grandeza do erro de medição considerado não degradou de forma evidente o desempenho na estimação dos parâmetros do modelo.

Parâmetro	**Valor**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$log_{10}(\lambda)$	1,56	1,52 (2,41)	1,55 (2,38)	1,57 (2,40)	1,55 (2,40)	1,55 (2,41)	1,55 (2,43)	1,56 (2,40)	1,57 (2,41)	1,54 (2,43)	1,53 (2,41)
$log_{10}(\rho)$	−0,97	−0,97 (2,55)	−0,98 (2,57)	−0,95 (2,55)	−0,99 (2,57)	−0,94 (2,57)	−0,98 (2,56)	−0,94 (2,56)	−0,97 (2,54)	−1,00 (2,59)	−0,97 (2,55)
$log_{10}(\delta)$	−0,30	−0,29 (2,71)	−0,30 (2,72)	−0,30 (2,69)	−0,30 (2,71)	−0,30 (2,71)	−0,31 (2,71)	−0,30 (2,70)	−0,30 (2,72)	−0,30 (2,71)	−0,30 (2,68)
$log_{10}(c)$	0,48	0,45 (2,69)	0,49 (2,68)	0,46 (2,72)	0,50 (2,72)	0,49 (2,74)	0,50 (2,71)	0,48 (2,69)	0,46 (2,73)	0,49 (2,70)	0,49 (2,73)
$log_{10}(N)$	3,00	3,00 (2,55)	3,02 (2,55)	2,98 (2,56)	2,98 (2,57)	3,01 (2,56)	3,00 (2,55)	3,01 (2,58)	2,97 (2,56)	3,00 (2,55)	2,99 (2,57)
Parâmetro	**Valor**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
$log_{10}(\lambda)$	1,26	1,25 (2,41)	1,26 (2,40)	1,27 (2,42)	1,25 (2,41)	1,27 (2,42)	1,28 (2,39)	1,27 (2,39)	1,24 (2,43)	1,26 (2,41)	1,28 (2,41)
$log_{10}(\rho)$	−1,40	−1,38 (2,56)	−1,41 (2,57)	−1,37 (2,57)	−1,39 (2,55)	−1,40 (2,55)	−1,40 (2,56)	−1,40 (2,55)	−1,42 (2,59)	−1,39 (2,57)	−1,40 (2,54)
$log_{10}(\delta)$	−0,37	−0,39 (2,67)	−0,36 (2,72)	−0,40 (2,71)	−0,37 (2,70)	−0,38 (2,74)	−0,33 (2,72)	−0,34 (2,69)	−0,39 (2,70)	−0,39 (2,71)	−0,36 (2,68)
$log_{10}(c)$	0,58	0,54 (2,70)	0,56 (2,71)	0,59 (2,72)	0,58 (2,72)	0,59 (2,75)	0,59 (2,68)	0,58 (2,70)	0,57 (2,71)	0,59 (2,72)	0,57 (2,71)
$log_{10}(N)$	3,44	3,44 (2,58)	3,44 (2,56)	3,43 (2,57)	3,46 (2,59)	3,43 (2,59)	3,43 (2,55)	3,45 (2,55)	3,46 (2,56)	3,43 (2,54)	3,45 (2,56)

Tabela 1: Estimativas dos parâmetros do modelo EDO, média (desvio padrão), para as realizações do indivíduo 1 (1 a 10) e para as realizações do indivíduo 2 (11 a 20).

A Figura 1 apresenta os dendrogramas obtidos pelas métricas de distância a) e b) considerando cada um dos parâmetros do modelo. Para ambas as métricas, os parâmetros λ, ρ, N e c permitem genericamente uma discriminação adequada das 20 réplicas em 2 grupos distintos (réplicas de 1 a 10 do indivíduo 1 e réplicas de 11 a 20 do indivíduo 2). No entanto, para o caso do parâmetro δ, observa-se que na discriminação pela métrica a) existem dois elementos que se encontram no grupo que não o esperado (réplicas 16 e 17 no grupo das réplicas do indivíduo 1). Com a métrica b), apenas o elemento 16 se encontra no grupo que não o esperado. Como observado na tabela 1, estas classificações incorretas devem-se ao facto dos valores médios de δ para as réplicas 16 e 17 serem os mais baixos em relação aos restantes valores médios do mesmo grupo e mais próximos dos valores médios das réplicas do outro grupo. Assim sendo, os resultados sugerem que a métrica b) poderá menos sensível a grandes variações nas médias dos parâmetros.

Para comparar as métricas de distância a) e b), de forma mais quantitativa, foram calculados os coeficientes de variação para os valores da distância obtidos entre pares de réplicas. Para avaliar a variabilidade intra grupo, consideraram-se as distâncias entre réplicas da mesma classe, enquanto que para quantificar a variabilidade entre grupos, consideraram-se as distâncias entre réplicas de classes diferentes. Os resultados obtidos encontram-se na Tabela 2. Como é possível observar, a dispersão intra-grupo é menor para a métrica b), indicando que a métrica b) permite quantificar uma distância mais pequena quando as réplicas são efetivamente do mesmo grupo. A métrica intra-grupos foi igual ou ligeiramente menor para a métrica b), indicando uma pequena degradação na diferenciação dos dois grupos distintos sem, no entanto, alterar o desempenho na classificação.

4. CONCLUSÕES

Este trabalho considerou o agrupamento hierárquico de trajetórias temporais através de duas métricas univariadas: a) uma baseada nas diferenças entre as médias das distribuições *a posteriori* e b) outra baseada nas diferenças ponderadas entre os percentis das distribuições *a posteriori*, para cada parâmetro do modelo. Pelos resultados obtidos podemos verificar que em ambas as abordagens existe uma separação correta dos dois grupos de indíviduos (1 a 10 e de 11 a 20), excepto para o

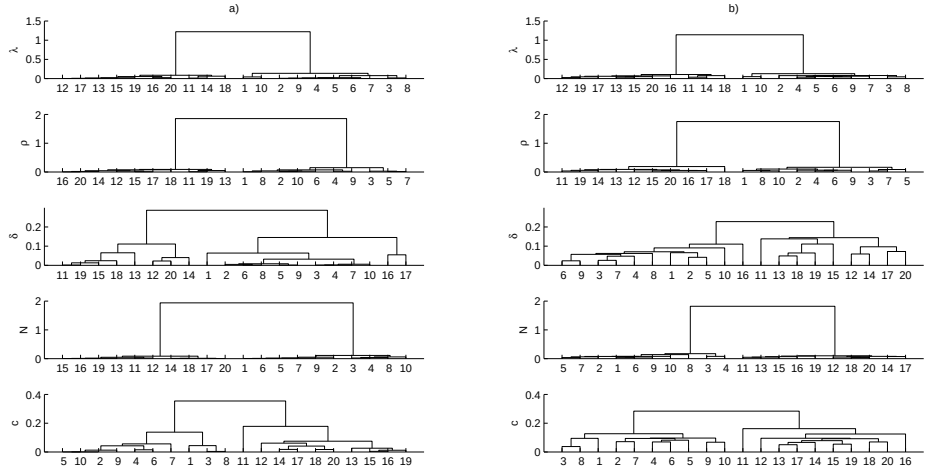


Figura 1: Dendrogramas obtidos pelas métricas de distância a) e b). Para mais informação, ver texto.

Parâmetro	**Intra-grupo (1 a 10)**		**Intra-grupo (11 a 20)**		**Entre-grupos**	
	a)	**b)**	**a)**	**b)**	**a)**	**b)**
λ	0.76	0.25	0.70	0.27	0.07	0.07
ρ	0.67	0.38	0.65	0.43	0.05	0.05
δ	0.82	0.26	0.68	0.33	0.32	0.26
c	0.73	0.21	0.77	0.25	0.25	0.20
N	0.68	0.41	0.68	0.27	0.04	0.04

Tabela 2: Coeficientes de variação intra e entre grupos, quantificados através dos valores da distância entre duas réplicas, respectivamente, da mesma classe e de classes diferentes. Para mais informação, ver texto.

parâmetro δ. No entanto, é também sugestivo de que uma métrica de distância multivariada, capaz de combinar adequadamente a informação distribucional de todos os parâmetros, poderá ainda melhorar o desempenho de classificação bem como diminuir a variabilidade intra-grupo.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia, FCT, através de fundos nacionais (MEC) e estruturais europeus (FEDER), no âmbito dos projetos UID/MAT/04106/2013 (CIDMA/UA), UID/CEC/00127/2013 (IEETA/UA) e UID/MAT/00144/2013 (CMUP/UP). Diana Rocha agradece o financiamento concedido através da bolsa de Doutoramento FCT (ref. SFRH/BD/107889/2015).

Referências

[1] Huang, Y., Liu, D., Wu, H. (2004). Hierarchical bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* 62, 413–423.

[2] Liang, H., Miao, H., Wu, H. (2010). Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. *The Annals of Applied Statistics* 4, 460–483.

[3] Whitby, L., Whitby, A.,Fletcher, M.,Helbert, M.,Reilly, J. T., Barnett, D. (2013). Comparison of Methodological Data Measurement Limits in CD41 T Lymphocyte Flow Cytometric Enumeration and Their Clinical Impact on HIV Management. *Cytometry Part B (Clinical Cytometry)* 84B, 248–254.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Ajuste do Peso e Gordura de Ratos por meio do modelo de Von Bertalanffy

Magali T. R. Amaral[1], Danielle F. Vileigas[2] e Antônio Carlos Cicogna [3], Carlos R. Padovani [4]

[1] Departamento de Bioestatística, Programa de pós-graduação em Biometria-UNESP, Universidade Estadual de Feira de Santana - UEFS.
[2,3] Departamento de Clínica Médica, Faculdade de Medicina de Botucatu-UNESP.
[4] Departamento de Bioestatística, Programa de pós-graduação em Biometria-UNESP.

RESUMO

O presente trabalho consiste em utilizar o modelo Von Bertalanffy para relacionar funcionalmente o peso e a gordura de ratos com estenose aórtica supravalar submetidos a dieta hipercalórica rica em ácidos graxos saturados sobre o metabolismo. Para a aplicação do procedimento estatístico considerou-se uma amostra de 25 ratos Wistar macho de 30 dias de idade. No contexto metodológico utilizou-se o algoritmo de simulação de Monte Carlo em Cadeia de Markov (Metropolis-Hastings) com abordagem bayesiana com densidades a priori não informativa de Jeffreys. O procedimento utilizado permitiu estimativas consistentes do peso dos ratos Wistar em função da gordura.

Palavras-chave: Modelo de Von Bertalanffy, Inferência Bayesiana, MCMC.

1. INTRODUÇÃO

A obesidade experimental por dietas hiperlipídicas tem sido utilizada como alternativa para o esclarecimento das causas e consequências da obesidade humana [5]. Estudos que envolve humanos obesos poderia responder muitas dúvidas em relação ao ganho de gordura e consequentemente do aparecimento de diversas doenças, entretanto essas pesquisas apresentam sérias limitações éticas e finaceiras.
O modelo experimental de Von Bertalanffy usual para ajuste de crescimento apresenta diversas dificuldades. Devido a não linearidade deste modelo, encontrar estimadores de máxima verossimilhança torna-se numericamente difícil frente a problemas de convergência. Neste sentido optou-se por uma abordagem bayesiana utilizando o algoritmo de Metropolis-Hastings (M-H) para fazer um sumário das densidades a posteriori, considerando estimadores de Monte Carlo para média, mediana, coeficiente de assimetria e curtose das densidades a posteriori. A análise bayesiana foi realizada contemplando densidades a priori não informativa de Jeffreys. O algoritmo de M-H usou como núcleo uma distribuição normal multivariada cujas médias e matriz de covariância são as modas e matriz de covariância da densidade a posteriori conjunta, calculadas numericamente usando-se o método de Laplace. A abordagem proposta é testada considerando-se uma amostra de dados de Ratos Wistar machos, com 30 dias de idade, provenientes do Biotério do Laboratório Experimental do Departamento da Clínica Médica, Faculdade de Medicina de Botucatu.

2.MODELO PROPOSTO

Para modelar crescimento existem diversas funções não lineares utilizadas, porém no presente estudo optou-se pelo modelo de Von Bertalanffy descrito a seguir:

$$P(g) = a(1 - ce^{-kg})^3 \tag{1}$$

em que $P(g)$ é a média do peso dos ratos em gramas, g é a gordura corporal em gramas, a é o parâmetro que representa o peso assintótico, c é uma constante que indica a proporção do crescimento assintótico a ser ganho, k é a taxa de crescimento da variável de interesse e determina a eficiência do crescimento dos ratos. É importante salientar que todos esses parâmetros sofrem efeitos aleatórios do tipo genético dos ratos que já podem ter tendência no aumento da gordura corporal, entretanto neste estudo tais efeitos não foram levados em consideração.

3.ABORDAGEM BAYESIANA

A função de verossimilhança para o modelo (1) é construída considerando $y_i = P(g_i;\theta) + \epsilon_i$ para $i = 1, 2, \ldots, n$ e

$$f(y_i|\theta,\tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}} \exp\{-\frac{\tau}{2}(y_i - P(g_i;\theta)^2)\},$$

com ϵ_i iid $N(0, \tau^{-1})$, em que $\tau = \frac{1}{\sigma^2}$, $\sigma > 0$ e $\theta \epsilon R^p$.

Sejam $\underset{\sim}{X} = (t_1, \ldots, t_n)^T$, $\underset{\sim}{Y}^T = (y_1, \ldots, y_n)$ e $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_p)^T$, então a função de verossimilhança de $f(y_i|\theta,\tau)$ é dada por:

$$L(\underset{\sim}{\theta},\tau|\underset{\sim}{Y},\underset{\sim}{X}) \propto \prod_{i=1}^{n} \tau^{1/2} \exp\left\{-\frac{\tau}{2}(y_i - P(g_i;\theta)^2)\right\}.$$

Considerando as densidades a priori definidas como, $\pi_0(\underset{\sim}{\theta},\tau) = \pi_0(\underset{\sim}{\theta}|\tau)\pi_0(\tau)$, e assumindo independência entre $\underset{\sim}{\theta}$ e τ a priori não-informativa de Jeffreys para τ é dada por $\pi(\tau) \propto (1/\tau)$.
A posteriori é escrita como

$$\pi(\underset{\sim}{\theta},\tau|\underset{\sim}{Y},\underset{\sim}{X}) \propto \tau^{(n/2)-1} \exp\left\{ -\frac{\tau}{2}(\underset{\sim}{Y}^T\underset{\sim}{Y} + \psi(\underset{\sim}{\theta}))\right\}\pi(\underset{\sim}{\theta}|\tau)\pi(\tau),$$

em que $\psi(\underset{\sim}{\theta}) = -2\underset{\sim}{Y}^T P(\underset{\sim}{X};\underset{\sim}{\theta}) + P^T(\underset{\sim}{X};\underset{\sim}{\theta})P(\underset{\sim}{X};\underset{\sim}{\theta})$.
A forma usual do teorema de Bayes a menos de uma constante normalizadora, pode ser escrita como $\pi(\theta|Dados) \propto L(Dados|\theta)\pi(\theta)$.
Neste estudo foram adotadas densidades a priori não informativa de Jeffreys definida por uma densidade proporcional à informação de Fisher [1], ou seja:

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}}. \tag{2}$$

De forma geral, a matriz de informação de Fisher $I(\theta)$, para o modelo 1 pode ser calculada como

$$I(\theta) = -E\left\{\frac{\partial^2 l(x|\theta)}{\partial\theta_i\partial\theta_j}\right\} = -\frac{1}{\sigma^2}\sum_{g=1}^{n}\frac{\partial P(g,\theta)}{\partial\theta_i}\frac{\partial P(g,\theta)}{\partial\theta_j} \tag{3}$$

A distribuição a priori não informativa de Jeffreys para o modelo de Von Bertalanffy pode ser construída da seguinte forma:

$$\mathbf{I}(\theta) = \begin{pmatrix} -\frac{n}{\sigma^2 a^2} & \frac{3\sum_{i=1}^n \phi_{t_i}(c,k)}{\sigma^2 a} & -\frac{3c\sum_{i=1}^n \phi_{t_i}(c,k)t_i}{\sigma^2 a} \\ \frac{3\sum_{i=1}^n \phi_{t_i}(c,k)}{\sigma^2 a} & -\frac{9\sum_{i=1}^n \phi^2_{t_i}(c,k)}{\sigma^2} & \frac{9c\sum_{i=1}^n \phi^2_{t_i}(c,k)t_i}{\sigma^2} \\ -\frac{3c\sum_{i=1}^n \phi_{t_i}(c,k)t_i}{\sigma^2 a} & \frac{9c\sum_{i=1}^n \phi^2_{t_i}(c,k)t_i}{\sigma^2} & -\frac{9c^2\sum_{i=1}^n \phi^2_{t_i}(c,k)t_i}{\sigma^2} \end{pmatrix}$$

$$|I(\theta)| \propto \left| \left(-\frac{n}{\sigma^2 a^2} \right) \left(-\frac{9\sum_{i=1}^{n} \phi_{t_i}^2(c,k)}{\sigma^2} \right) \left(-\frac{9c^2 \sum_{i=1}^{n} \phi_{t_i}^2(c,k) t_i}{\sigma^2} \right) \right|^{\frac{1}{2}}.$$

$$\pi(\theta) \propto \frac{\sqrt{\sum_{i=1}^{n} \phi_{t_i}^2(c,k) \sum_{i=1}^{n} t_i \phi_{t_i}^2(c,k)}}{(a/c)}.$$

Para calcular os estimadores bayesianos dos parâmetros dos modelos, utiliza-se o algoritmo de simulação de Monte Carlo Cadeia de Markov (MCMC). Especificamente, o algoritmo da Metropolis-Hastings [2], com as densidades condicionais

$$\pi(\underset{\sim}{\theta}|\tau, \underset{\sim}{Y}, \underset{\sim}{X}) \propto \exp\left\{ -\frac{\tau}{2}\psi(\underset{\sim}{\theta}) \right\} \pi_0(\underset{\sim}{\theta}|\tau),$$

$$\pi(\tau|\underset{\sim}{\theta}, \underset{\sim}{Y}, \underset{\sim}{X}) \propto G(a, b(\underset{\sim}{\theta})),$$

em que $G(a, b(\underset{\sim}{\theta}))$ denota uma densidade gamma com parâmetros $a = \frac{n}{2}$ e $b(\theta)$ dado por $b(\theta) = \frac{1}{2}(\underset{\sim}{Y}^T \underset{\sim}{Y} + \psi(\underset{\sim}{\theta}))$.

3. RESULTADOS

Para análise foi selecionada uma amostra de 25 ratos, com peso entre 432 e 715 gramas e gordura variando entre 28 e 75 gramas. Uma amostra de tamanho 150.000 foi gerada pelo algoritmo de M-H. A convergência do algoritmo foi monitorada pelo critério de Geweke [3]. Após a convergência, 50 % das amostras são descartadas. Para eliminar a correlação entre os valores gerados, uma reamostragem foi feita selecionando entre os parâmetros gerados um a cada 15 valores.
As estatísticas a posteriori para o modelo Von Bertalanffy, obtida após a simulação, são apresentadas na Tabela 1 onde observa-se a curtose um pouco elevada para o parâmetro c e grande amplitude no intervalo de credibilidade para o parâmetro a. Nota-se também, pelo critério de Geweke, que houve convergência em todos os parâmetros e que os intervalos de credibilidade foram todos significantes.
Assim pelas estimativas dos parâmetros do modelo Von Bertalanffy, observa-se que o peso assintótico é no máximo 839.9 g, com taxa de crescimento de aproximadamente 0.013g. Isto é, os ratos podem ser representados pelo preditor expresso por $P(g) = 839{,}9(1 - 0{,}272e^{-0{,}013g})^3$ com um crescimento mediano aproximado de 328g.

Tabela 1: Sumário da densidade a posteriori do modelo de Von Bertalanffy usando o Metropolis-Hastingns com 44 % de aceitação.

Estatística a posteriori	a	c	k	τ
Moda	635.216	0.236	0.033	113.764
Média	838.677	0.272	0.017	161.984
Mediana	839.934	0.267	0.013	157.278
Desvio-padrão	133.023	0.060	0.010	49.821
Assimetria a posteriori	0.274	1.703	1.863	0.848
Curtose	0.499	7.773	4.232	1.343
Intervalo de credibilidade	(621.103, 1132.189)	(0.173, 0.419)	(0.006, 0.048)	(83.395, 280.370)
Critério de Geweke	0.344	-0.334	-0.592	-0.6100

A Figura 1 mostra a curva de crescimento ajustada com o modelo Von Bertalanffy.

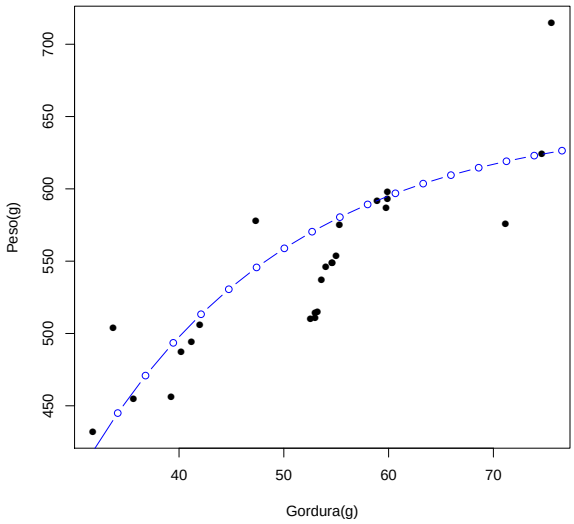


Figura 1: Curva ajustada pelo modelo de Von Bertalanffy.

4. CONCLUSÕES

O ajuste de curvas de crescimento na abordagem bayesiana mostrou-se de fácil aplicação e entendimento. Para fins de comparação, optou-se pelo uso de densidades a priori não informativas de Jeffreys. Com a metodologia aplicada constata-se que a curva de crescimento do modelo de Von Bertalanffy, foi adequado, pois forneceu excelentes estimativas significantes, indicando que a metodologia empregada pode ser aplicada para descrever o crescimento da gordura corporal em ratos.

AGRADECIMENTOS

A Pró-reitoria de Pós-graduação da Unesp (PROPG) pelo apoio financeiro.

REFERÊNCIAS

Referencias

[1] BOX, G. E. P; TIAO, G. C., *Bayesian inference in statistical analysis. J. Wiley, New York, NY, EUA, 1973.*

[2] CHIB, S.; GREENBERG, E., Understanding the Metropolis-Hastings Algorithm. *American Statistician, 49 (4), 327-335, 1995.*

[3] GEWEKE, J., *Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bayesian Statistics 4ed., JM Bernado, JO Berger, AP Dawid and AFM Smith). Clarendon Press, Oxford, UK, 1992.*

[4] JIM, A., *Bayesian Computation with R: Department of Mathematics and Statistics. 3ed., Springer, New York, NY, USA, 2008.*

[5] Campos, D.H.Salomé, *A influência da dieta hiperlipídica ricas em ácidos graxos saturados sobre o metabolismo, estrutura e função cardíaca de ratos com estenose aórtica supravalar.*

[6] Paulino, C.D., Amaral Turkman, A., Murteira, B. (2003). *Estatística Bayesiana.* Fundação Calouste Gulbenkian, Lisboa.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Joint modelling of multivariate longitudinal and survival data. A two-stage model based proposal with application to nephrology research.

Ipek Guler[1], Christel Faes[2] Carmen Cadarso-Suárez[1] Laetitia Teixeira[3] Anabela Rodrigues[4] Denisa Mendonça[3,5]

[1] Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), University of Santiago de Compostela, A Coruna, Spain
[2] Center for Statistics, Hasselt University, Belgium
[3] Institute of Biomedical Sciences Abel Salazar, University of Porto (ICBAS), University of Porto, Portugal
[4] Nephrology Department, Centro Hospitalar do Porto, Hospital Geral de Santo Antonio, Porto, Portugal
[5] Public Health Institute, University of Porto

Abstract

Many follow-up studies in biomedical research produce different types of outcomes including both longitudinal measurements and time-to-event outcomes. Often, it is of interest to study the association between them. The joint modeling approaches of a single longitudinal outcome and survival process have gained an increasing attention recently from both frequentist and Bayesian perspective. However, many studies collect several longitudinal biomarkers and instead of selecting one single longitudinal biomarker, the interest is often on the relationship between all these biomarkers and the survival process. Existent joint modeling approaches from the frequentist perspective are difficult to implement when the number of longitudinal biomarkers is large. We propose a two-stage model based approach for multivariate longitudinal and survival data. The proposed model is illustrated by a real data example from peritoneal dialysis program data, Nephrology Department, Centro Hospital do Porto, Hospital Geral de Santo Antonio, Porto, Portugal.

Keywords: joint modeling, multivariate longitudinal, survival, two-stage models

1. Introduction

Joint modeling approaches for longitudinal and survival data have gained an increasing attention in the literature over recent years. One of the common framework for these models is shared random effects models. The main idea behind this framework is to link the longitudinal and survival data with a random effect which underlines the conditional distributions of the both processes and maximize the full joint likelihood. However, this calculation becomes computationally complex in case of having a large number of longitudinal biomarkers.
For instance, peritoneal dialysis patients data present two different types of outcomes: (i) longitudinal outcomes, composed by clinical parameters measured at several time points and (ii) time-to-event outcome, composed by the follow-up time until the occurrence of an event of interest. We

can observe in Figure 2 the subject specific trends of the albumin, calcium and serum phosphor measurements for the alive and death patients, respectively. The albumin level is considered as an important clinical parameter for end-stage renal patients by clinicians and it is used to assess the health status of the patients in dialysis (Mehrotra, Duong et al., 2011). The aim of this application is to study the ability of these three longitudinal outcomes, albumin, calcium and serum phosphor, to predict the mortality.
To avoid the above mentioned complexities on joint likelihood calculation, we propose a two-stage model based approach to study the relationship between all these longitudinal biomarkers and their association with survival. The main idea of this two-stage model based proposal is coming from the initial approaches of simple longitudinal and time-to-event data (see Self and Pawitan (1992) and Tsiatis et al. (1995) among others) where the likelihood calculation is divided into two stages, separately for the single longitudinal outcome and survival process. Extending this idea, we can use the pairwise modeling approach to study the multivariate longitudinal data introduced by Fieuws and Verbeke (2005) and Fieuws et al. (2007) in the first stage. This approach takes into account the correlation structure of the longitudinal biomarkers assuming a multivariate normal distribution for each random effects. Furthermore, we can use the estimations of this model at time t in stage 2 to incorporate them into the survival model (See Figure 1).
The outline of the paper is as follows. We first introduce an earlier method to study the longitudinal and survival data which is an alternative way to deal with multivariate longitudinal data and survival. In the second section we describe our model proposal with the application. Finally we give the results including a comparison of our approach with the alternative method.

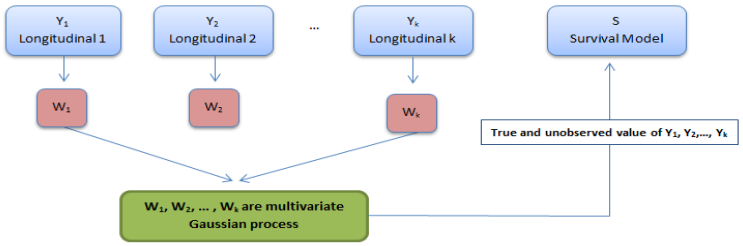


Figure 1: The main idea behind two-stage modelling

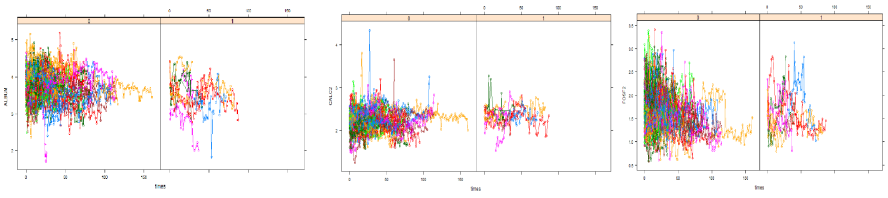


Figure 2: Subject specific trends over time of the longitudinal biomarkers, albumin, calcium and phosphor measurements, respectively

2. Time-dependent covariates (Model 1)

One of the earlier methods to study the longitudinal and survival data is an extension of the Cox Model introduced by Anderson and Gill (1982). This model allows incorporation of time dependent covariates into the survival model. We can use the latter approach to incorporate our three longitudinal biomarkers into the survival. However this model is focused on the survival process and studies independently the multivariate longitudinal biomarkers without taking into account the measurement error of each one and the correlation structure between them.
Let $h_0(t)$ is the baseline risk function at time t, the model can be written as follows:

$$h_i(t) = h_0(t)R_i(t)exp(\gamma_{1i} * age_i + \gamma_{2i} * gender_i + \alpha_1 * Album_i(t) + \alpha_2 * Calc_i(t) + \alpha_3 * Phosphor_i(t)) \quad (1)$$

where Album, Calc and Phosphor are the longitudinal biomarkers encoded using (start-stop) notation. These points indicate the time intervals of the recorded repeated measurements. $R_i(t)$ is a left continuous at risk process with $R_i(t) = 1$ if subject i is at risk time at time t and $R_i(t) = 0$ otherwise.

3. Two-stage model based proposal

This model proposal is composed by two stages: (1) A multivariate longitudinal model (2) A proportional hazard regression to study the survival.

Stage 1: Multivariate Longitudinal Model

In many situations, the multivariate longitudinal data includes correlated longitudinal biomarkers. In case of having these correlations, the association structure is important to take into account and joint modeling approach is needed. However, as mentioned before, when the number of longitudinal biomarkers increase the computational problems arise when calculating the joint likelihood function. To avoid these computational problems Fieuws and Verbeke (2005) introduced a pairwise approach of multivariate longitudinal data, instead of maximizing the likelihood of the full joint model. They fit all the possible pairs and obtain multiple maximum likelihood which will be averaged at the end to obtain one single estimate for each parameter. We use this latter approach in the first stage of the proposed model for the three longitudinal biomarkers included in the database: serum albumin, calcium and phosphor measurements.

$$Y_{ijk} = \beta_{0k} + \beta_{1k} * time_{ij} + \beta_{3k} age_i + \beta_{4k} gender_i + u_{0ik} + u_{1ik} * time_{ij} + \epsilon_{ijk}(t)$$

where Y_{ijk} is the longitudinal observation of ith patient at time j and k is the number of longitudinal biomarkers with k=3 (albumin, calcium and phosphor measurements).

In the pairwise fitting approach the log likelihood of the following form will be maximized separately $\sum_{i=1}^{N} l_{pi}(\theta_p)$ where $p = 1, ..., P$ and $P = k(k-1)/2$ is the total number of possible pairs and θ then be the vector combining each pair-specific parameter vectors θ_p Estimates for the elements in θ are obtained by maximizing each of the p likelihoods. (Fieuws and Verbeke, 2006; Fieuws, Verbeke and Molenberghs, 2007). Fitting all possible pairwise models is equivalent to maximizing a pseudo-likelihood (pl) function of the following form;

$$pl(\theta) = \sum_{i=1}^{N} l_{rsi}(Yri, Ysi|\theta_{r,s}) \quad (2)$$

where $r = 1, ..., k-1$ and $s = r+1, ..., k$. $\theta_{r,s}$ represents the vector of all parameters in each pair (r,s) of joint mixed model.
The asymptotic multivariate normal distribution for θ is given by;

$$\sqrt{N}(\theta - \theta) \approx MV(0, J^{-1}KJ^{-1})$$

where J is a block-diagonal matrix with diagonal blocks J_{pp} ,and where K is a symmetric matrix containing blocks K_{pq}. In the final step, estimates for the parameters can be calculated by taking averages over all pairs.

Stage 2: Survival Model

Coef (Std.Error)		
	Model 1	Model 2
gender	0.47(0.60)	-0.36(0.59)
age	0.04(0.02)	0.009(0.02)
α_1	-0.62(0.64)	-1.39(0.61)
α_2	1.23(1.68)	2.91(1.35)
α_3	0.23(0.72)	0.98(0.67)
logLik	-43.36	-52.01

Table 1: The results of Survival Model for Model 1 and Model 2 without a measurement error

We use the classical Cox proportional model (Cox, 1972) including the estimated values of longitudinal biomarkers obtained from the first stage.

$$h_i(t) = h_0 \exp(\gamma_{1i} * age_i + \gamma_{2i} * gender_i + \alpha_k m_{ik}(t)) \tag{3}$$

where $m_{ik}(t) = u_{0ik} + u_{1ik} * times$, h_0 is the baseline risk function, X_{it} are the fixed effects $m_{ik}(t)$ are the predictions of random intercept and slope for *kth* longitudinal biomarker at time (t), thus the α_k represents the association between this longitudinal biomarker and the patients‘ survival.

4. Results

The association between the random effects of the three longitudinal biomarkers from the multivariate longitudinal model is observed in Figure 3. Model 2 shows significant effect of the albumin and calcium measurements on survival. The association between the longitudinal biomarkers and survival are substantially larger in Model 2 than the Model 1. This is due to bias estimations of the Model 1 ignoring the measurement error in longitudinal data . Also the effect of longitudinal biomarkers become non-significant in Model 1.

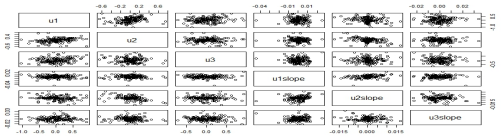


Figure 3: The correlation between the random effects u_{0ik} and u_{1ik} presented in Stage 1

Acknowledgements

This study developed under the project MTM2014-52975-C2-1-R:Ïnference in Structured Additive Regression (STAR) Models with Extensions to Multivariate Responses. Applications in Biomedicine". Cofinanced by the Ministry of Economy and Competitiveness (SPAIN) and by the European Regional Development Fund (FEDER).

Referencias

[1] Anderson, P.K., Gill, R.D. (1982). Cox‘s Regression Model for Counting Process: A Large Sample Study. *Annals of Statistics* 10, 1100 – 1120.

[2] Fieuws S, Verbeke G. (2005). Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles. *Biometrics* 62, 424–431.

[3] Fieuws S, Verbeke G, Molenberghs G. (2007). Random-effects models for multivariate repeated measures. *Stat Methods Med Res* 16, 387–397.

[4] Mehrotra, R., U. Duong, S. Jiwakanon, C. P. Kovesdy, J. Moran, J. D. Kopple and K. Kalantar-Zadeh (2011). Serum albumin as a predictor of mortality in peritoneal dialysis: comparisons with hemodialysis. *American Journal of Kidney Diseases* 58(3), 418-428.

[5] Vasconcelos, R. (1996). Self, S., Pawitan, Y.] (1992). Modeling a marker of disease progression and ofset of disease. *AIDS Epidemiology:Methodological Issues of Eds. Jewell, N.P., Dietz, K. Farewell, V.T. Boston* Birkhauser.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

A new method for detection of rhythmic signals in oscillatory systems

Yolanda Larriba[1], Cristina Rueda[1], Miguel A. Fernández[1] and Shyamal D. Peddada[2]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Spain
[2]Biostatistics Branch, NIEHS (NIH), Research Triangle Park, NC, USA

Abstract

The study of biological rhythms is receiving a lot of attention in the literature in recent years. At the core of this research lies the methodological problem of how to detect rhythmic signals in measured data. Night and day, or dark and light patterns impact on human health in many different ways. For this reason, researchers are studying the effect of sleep on the circadian clock in human body during various stages of life. Important components of this clock are the circadian genes which have rhythmic expression overtime with phases suitably matching the night and day. Consequently, the identification of rhythmic signals is a problem of considerable interest for biologists.

In this work, we develop a novel statistical procedure to detect rhythmic signals in oscillatory systems based on Order Restricted Inference (ORI). This methodology is tested both on simulations and on real data bases. Moreover the obtained results are compared with the most widely extended rhythmicity detection algorithms in literature.

Key words: Order Restricted Inference, Conditional Tests, Circadian Rhythms, Cyclic Signal, Oscillatory Systems

1. INTRODUCTION

Human health is influenced by rhythmicity patterns, like day and night, in many ways with circadian clock regulating metabolic and physiologic processes. There is abundant literature documenting that people are having, on average, less sleep during the night and this disruption or reduction in sleep is associated with numerous health outcomes including obesity. Among teenagers this may affect the production of their growth hormones and result in abnormal growth patterns. For these reasons, researchers are in studying the effect of sleep on the circadian clock in human body during various stages of life. Important components of this oscillatory system are the circadian genes which present rhythmic expression overtime according to night and day cycles. Circadian genes consist of two well differentiated compounds, the signal term that we denote by $\boldsymbol{\mu}$, and the random error term which is derived from the experimental noise. Thus, the research in oscillatory systems focuses on the detection of which of the signals they origin are rhythmic.
The identification of rhythmic genes in the circadian clock among several thousands of genes is not a simple problem due to the large variability in the data, the high number of data (several

thousands of genes) to be processed and the fact that in many of the available datasets the number of retrievable data for each gene is not high enough to properly fit mathematical models such as Fourier's models. This difficulty is reflected in the richness of the literature on this subject as well as in the wealth of methods and algorithms devoted to this task (e.g. Straume [2004], Hughes et al. [2010], Thaben and Westermark [2014]). These methods assess the fit between the expression pattern of each transcript and a series of curves, most often sinusoidal ones, with different period lengths and phases, and they estimate the period length and phase of each circadian transcript as that of its best-matched curve.

2. CONTRIBUTIONS

Circadian gene expressions data are observed during two periods of length 24 hours ($T = 24$) and usually the time sampling frequency is fixed to be 1 hour, i.e., there are 24 time points ($n = 24$) in each period. The general mathematical formulation of rhythmicity (periodicity) in the two periods $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ (from now on *property 1*) may be too rigid a model to be useful to detect interesting genes in practice. Real data evidence that the expression profiles of many genes are rhythmic but do not come from a sinusoidal signal, although quite often the signal has an unique maximum in each period corresponding to the moment where the gene is activated, and an unique minimum generating an up-down up pattern, (from now on *property 2*). Finally, in order to define alternative signal patterns that fit to the different gene expression data appearing in practice, we have distinguished four main classes of genes that correspond to the four signal patterns shown in Figure 2.
In this line, this work proposes a broader formulation for a rhythmic gene signal that we call *cyclic signal*, (see Definition 1), which includes as a particular case the sinusoidal signal. Figure 1 shows a typical cyclic pattern according to Definition 1.

Definition 1 *Cyclic signal.* $\boldsymbol{\mu}$ *is a cyclic signal* $\iff$ $\boldsymbol{\mu} \in \mathcal{C} = \bigcup_{L,U} C_{LU}$, *where* $L, U \in \{1, \ldots, n\}$ *and* $C_{LU} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mu_L \leq \mu_{L+1} \leq \cdots \leq \mu_U \geq \mu_{U+1} \geq \cdots \geq \mu_{L-1}\}$.

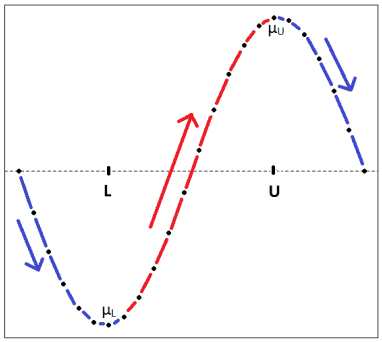


Figure 1: Profile of a Cyclic Signal $\boldsymbol{\mu}$

Two elements are remarkable in the definition of a cyclic signal. On one side, the two landmarks L and U, that determine the minimum and maximum in a period. On the other side, the restrictions among the components of $\boldsymbol{\mu}$, determining the up-down-up pattern. This order pattern points to the use of ORI methodology (see Robertson et al. [1988]). In addition, the properties defined above let us classify genes into one of the main classes of genes according to their signal patterns, see Figure 2. To be more precise, most typical periodic genes verify *properties 1* and *2* and belong to Class A. Genes in class B verify *property 1* but not *property 2*, i.e., they present for each period more than one local maximum or minimum. Class C includes the genes whose expression data variability, in both periods, is totally explained by the random error term. Finally, genes in Class D fail *property 1* but verify *property 2* at least in one period. From the biological interest point of view, classes A and B include periodic genes and C and D involve the non periodic ones.

(a) Class A: Cyclic

(b) Class B: Quasi Cyclic

(c) Class C: Flat

(d) Class D: Non homogeneous

Figure 2: Class gene patterns in the two periods

This work also presents the algorithm we propose to identify periodic genes, and classify them into the classes appearing in Figure 2. It consists of three basic steps: estimate the landmarks L and U; conduct a filtering step and then a classification step. The algorithm relies heavily on ORI methodology and involves a series of sequentially likelihood ratio conditional tests (see Menéndez et al. [1991], Fernández et al. [2012]). In addition, a false discovery rate (FDR) procedure is also considered since multiple testing problems are conducted.

3. RESULTS

To evaluate the performance of ORI to detect periodicity, different simulation studies have been conducted, including the simulation of an *artificial dataset* which imitates what occurs in circadian real data bases. This *artificial dataset* contains 40000 simulated genes. The proportion of periodic and non periodic genes are 30% and 70%, respectively. The first group of patterns in Table 1 simulates periodic genes and the second one does non periodic genes. We also compare the ORI results with those from the most widely used rhythmicity detection algorithms in literature, JTK_Cycle (JTK) and RAIN (see Hughes et al. [2010] and Thaben and Westermark [2014], respectively), which are non parametric algorithms to detect sinusoidal signals.

	ORI		JTK		RAIN	
Shape pattern	FP+	FN-	FP+	FN-	FP+	FN-
Cosine		0.000		0.000		0.000
Cosine Two		0.000		0.000		0.000
Cosine Peak		0.000		0.003		0.000
Sine Square		0.000		0.000		0.000
Asymmetric		0.001		0.973		0.652
Quasi Cyclic		0.001		1.000		0.687
Flat	0.052		0.000		0.018	
One Outlier	0.000		0.000		0.010	
Two Outliers	0.000		0.000		0.014	
Cosine Flat	0.012		0.504		0.900	
Flat Trend	0.008		0.102		0.572	
MEAN ERROR	**0.014**	**0.000**	**0.121**	**0.329**	**0.303**	**0.223**

Table 1: False positive and negative rates for each algorithm for the *artificial dataset*

From Table 1 we conclude that JTK presents the highest false negative rates for asymmetric and quasy cyclic shape patterns and that RAIN increases JTK false positive rates for non periodic patterns. In contrast with this, ORI outperforms the simulation results obtained by JTK and RAIN.

We have also analysed four real datasets including (each of them) 45101 circadian genes from mouse liver, pituitary and NIH3T3 cell lines; and 32321 circadian genes from U2OS human cell lines. In all four real datasets, RAIN is the approach which identifies more periodic genes and JTK is which does less, while ORI is between them, reiterating the results obtained in simulations. In addition, we found particularly interesting the case of NIH3T3 and U2OS cell lines, where as it is known the influence of noise is higher; but even so ORI still advantages over JTK.
Finally, from biological interest point of view, the relevant contribution is that ORI detects new periodic circadian genes. Figure 3 illustrates six of those new periodic circadian genes from the mouse liver tissue. In contrast with ORI, JTK and RAIN detect all of them as non periodic genes, despite that a simple inspection of them confirms a strong circadian rhythmiticity.

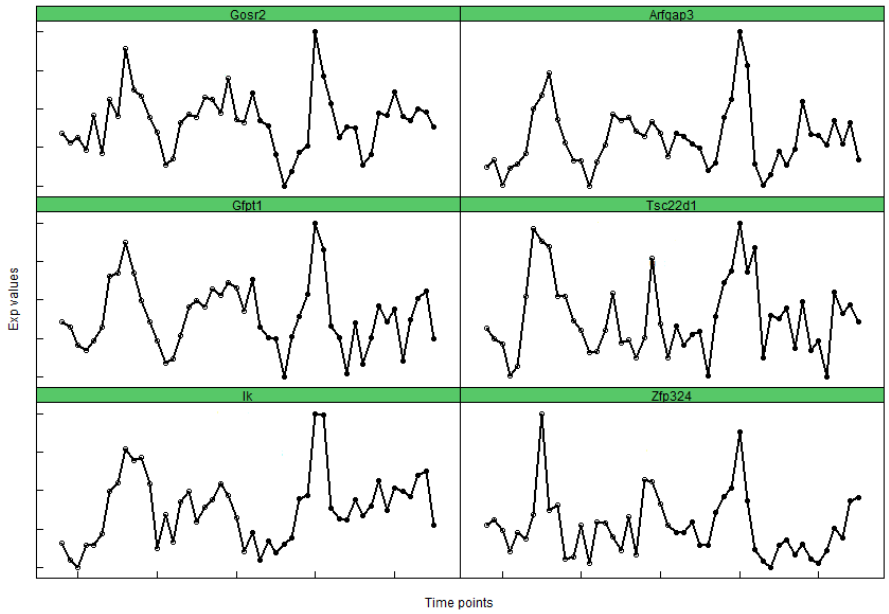


Figure 3: Six new periodic circadian genes detected by ORI from the mouse liver tissue

References

M.A. Fernández, C. Rueda, and S.D. Peddada. Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Research*, 40(7):2823–2832, 2012.

M.E. Hughes, J.B. Hogenesch, and K. Kornacker. JTK CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–380, 2010.

J.A. Menéndez, C. Rueda, and B. Salvador. Conditional test for testing a face of the tree order cone. *Communications in statistics. Simulation and computation*, 20(2-3), 1991.

T. Robertson, F.T. Wright, and R.L. Dykstra. *Order Restricted Statistical Inference*. John Wiley & Sons, 1988.

M. Straume. DNA microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning. In Ludwig Brand and Michael L. Johnson, editors, *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 149–166. Academic Press, 2004.

P.F. Thaben and P.O. Westermark. Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400, 2014.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Selección de covariables significativas en la incidencia de un modelo de curación no paramétrico

Ana López-Cheda[1], Mª Amalia Jácome [2] y Ricardo Cao[1]

[1]Grupo MODES, Dep. Matemáticas, INIBIC, Facultade de Informática, Universidade da Coruña
[2]Grupo MODES, Dep. Matemáticas, INIBIC, Facultade de Ciencias, Universidade da Coruña

RESUMEN

Para el modelo de curación de tipo mixtura completamente no paramétrico propuesto por López-Cheda et al (2016), se propone un test para la selección de covariables significativas en la probabilidad de cura, basado en el contraste de significación de Delgado and González-Manteiga (2001). Concretamente, se contrasta si dicha probabilidad es una constante. Finalmente, el buen comportamiento del método propuesto se muestra en un estudio de simulación con Monte Carlo, en el que la distribución del test es aproximada mediante bootstrap.

Palabras y frases clave: bootstrap, datos censurados, análisis de supervivencia

1. INTRODUCCIÓN

La efectividad de los actuales tratamientos contra el cáncer han provocado un aumento en la probabilidad de cura de un paciente o, al menos, de sobrevivir a largo plazo. Como consecuencia, surgen los llamados modelos de curación. Sea Y el tiempo hasta el suceso de interés. Los modelos de curación de tipo mixtura distinguen entre sujetos curados ($\nu = 1, Y = \infty$) y no curados ($\nu = 0, Y < \infty$) y, dado un vector de covariables $\mathbf{X}$, permiten estimar la proporción de pacientes curados (incidencia) $1 - p(\mathbf{x})$, donde $p(\mathbf{x}) = P(\nu = 0|\mathbf{X} = \mathbf{x})$ y la supervivencia en un tiempo determinado de los pacientes que no se curan (latencia), $S_0(t|\mathbf{x}) = P(Y > t|\mathbf{X} = \mathbf{x}, \nu = 0)$, independientemente. Sea $S(t|\mathbf{x}) = P(Y > t|\mathbf{X} = \mathbf{x})$ la función de supervivencia condicional de Y, el modelo de curación tipo mixtura asume que

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

Bajo el modelo de censura aleatoria y siendo C la variable censurante, el tiempo observado es $T = \text{mín}(Y, C)$ y $\delta = 1(Y \leq C)$ es la indicadora de no censura. Existen métodos completamente no paramétricos de estimación en este modelo que han sido propuestos recientemente por Xu and Peng (2014) y López-Cheda et al (2016).

2. CONTRASTE PARA LA SELECCIÓN DE COVARIABLES EXPLICATIVAS

En este trabajo se propone un test para la selección de covariables significativas en la incidencia. De forma general, sea $\mathbf{W} = (\mathbf{X}, \mathbf{Z}) = (X_1, ..., X_q, Z_1, .., Z_p)$ el conjunto de covariables. Se desea estudiar si la probabilidad de cura (incidencia) depende de las covariables $Z_1, ..., Z_p$:

$$H_0 : 1 - p(\mathbf{X}, \mathbf{Z}) = E\left(\nu|\mathbf{X}, \mathbf{Z}\right) \equiv E\left(\nu|\mathbf{X}\right) = 1 - p\left(\mathbf{X}\right).$$

La variable ν no es observable debido a la censura, puesto que se desconoce si un individuo censurado sería finalmente curado ($\nu = 1$) o no ($\nu = 0$). En su lugar, trabajaremos con η que verifica igualmente $E(\nu|\mathbf{X}) = E(\eta|\mathbf{X}) = 1 - p(\mathbf{X})$, definida como:

$$\eta = \frac{\nu(1 - \mathbf{1}(\delta = 0, T \leq \tau))}{1 - G(\tau|\mathbf{X})},$$

donde τ es un tiempo a partir del cual el sujeto se considera curado y G es la función de distribución condicional de C. En la práctica, τ se estima por $T^1_{\text{máx}} = \text{máx}_{i:\delta_i=1} T_i$. Además, se asume que la variable censurante C no depende de las covariables $\mathbf{X}$, de modo que $G(\tau|\mathbf{X})$ se estima por el estimador límite producto, $G^{KM}(T^1_{\text{máx}})$, definiendo las estimaciones de η de la siguiente manera:

- $\delta_i = 1$ (fallo) $\Rightarrow$ no curado ($\nu_i = 0$) $\Rightarrow \hat{\eta}_i = 0$
- $\delta_i = 0$ (censurado) y $T_i \leq \tau \Rightarrow \hat{\eta}_i = \frac{\nu_i(1-1)}{1-G^{KM}(T^1_{\text{máx}})} = 0$
- $\delta_i = 0$ (censurado) y $T_i > \tau \Rightarrow$ curado ($\nu_i = 1$) $\Rightarrow \hat{\eta}_i = \frac{1(1-0)}{1-\widehat{G}(\tau|\mathbf{X}_i)} = \frac{1}{1-G^{KM}(T^1_{\text{máx}})}$.

En este trabajo nos centraremos en el caso de una sola covariable, $\mathbf{W} = Z$, estudiando si la probabilidad de cura depende de dicha covariable:

$$H_0 : E(\nu|\mathbf{Z}) \equiv E(\nu) = 1 - p \ \text{ constante } \ \text{vs } H_1 : E(\nu|\mathbf{Z}) = 1 - p(Z)$$

a partir de las observaciones $\{(Z_i, \hat{\eta}_i), i = 1, \ldots, n\}$.
Siguiendo las ideas de Delgado and González-Manteiga (2001), el estadístico de contraste será un funcional del siguiente proceso:

$$T_n(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j\right)\right)\mathbf{1}(Z_i \leq z),$$

tal como el estadístico de Cramér-von Mises

$$C_n = \sum_{i=1}^{n} T_n^2(Z_i),$$

o el de Kolmogorov-Smirnov

$$K_n = \max_{i=1,\ldots,n} |n^{1/2} T_n(Z_i)|.$$

La distribución del test se aproxima mediante técnicas bootstrap, a partir de un remuestreo naive independiente. El procedimiento bootstrap es el siguiente:

1. Para $i = 1, 2, \ldots, n$, generar Z_i^* y $\hat{\eta}_i^*$ a partir de $(Z_1, \ldots, Z_n)$ y $(\hat{\eta}_1, \ldots, \hat{\eta}_n)$ de forma independiente, mediante muestreo aleatorio con reemplazamiento.

2. Con la remuestra bootstrap $\{(Z_i^*, \hat{\eta}_i^*), i = 1, \ldots, n\}$, calcular la versión bootstrap de T_n:

$$T_n^*(Z_k^*) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\eta}_i^* - \left(\frac{1}{n}\sum_{j=1}^{n}\hat{\eta}_j^*\right)\right)\mathbf{1}(Z_i^* \leq Z_k^*) \quad k = 1, \ldots, n$$

   y la correspondiente versión bootstrap del estadístico de Cramér-von Mises y Kolmogorov-Smirnov

$$C_n^* = \sum_{i=1}^{n} T_n^*(Z_i^*) \quad \text{y} \quad K_n^* = \max_{i=1,\ldots,n} |n^{1/2} T_n^*(Z_i^*)|.$$

3. Repetir B veces los pasos 1-2 para generar B valores de C_n^* y K_n^*. Definir el punto crítico d^* como el valor que está en la posición $(1 - \alpha)B$ del vector ordenado.

4. Comparar el valor del estadístico C_n (respectivamente K_n) obtenido con la muestra original con d^*, y rechazar la hipótesis nula si $C_n > d^*$ (respectivamente $K_n > d^*$).

5. Repetir m veces los pasos $1-4$. La potencia del test se aproxima como el porcentaje de rechazos del total de m.

3. ESTUDIO DE SIMULACIÓN

El objetivo del estudio de simulación es mostrar el buen comportamiento de la metodología propuesta. Consideramos dos modelos diferentes. En ambos, los tiempos de censura fueron generados a partir de una distribución exponencial con media $10/3$ y la covariable Z fue generada a partir de una distribución uniforme $U(-20, 20)$.

Modelo 1 La incidencia es $1 - p(x)$, donde

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

con $\beta_0 = 0{,}476$ y $\beta_1 = 0{,}358$ y la latencia es una función que aproximadamente cumple el modelo de riesgos proporcionales y que ha sido truncada:

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

donde $\tau_0 = 4{,}605$ y $\lambda\,(x) = \exp\left((x + 20)/40\right)$. Un $54\,\%$ de los pacientes son censurados y un $47\,\%$ son curados.

Modelo 2 La probabilidad de no cura es:

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

con $\beta_0 = 0{,}0476$, $\beta_1 = -0{,}2558$, $\beta_2 = -0{,}0027$ y $\beta_3 = 0{,}0020$ y la función de supervivencia de la población susceptible es

$$S_0(t|x) = \frac{1}{2}\left(\exp(-\alpha(x)t^5) + \exp(-100t^5)\right),$$

con

$$\alpha(x) = \frac{1}{5}\exp((x + 20)/40).$$

En este caso, el porcentaje de curados y censurados es ligeramente mayor que para el Modelo 1: cerca del $62\,\%$ de los individuos son censurados y alrededor del $53\,\%$ son curados.
Trabajamos con muestras de dos tamaños diferentes, $n = 100$ y $n = 200$ y simulamos $m = 1000$ trials con $B = 1000$ remuestras bootstrap. Bajo la hipótesis nula estudiamos cuatro valores diferentes de dicha constante: p=0.3, p=0.5, p=0.7 y p=0.8. Los resultados en la Tabla 1 fueron obtenidos bajo la hipótesis nula y los de la Tabla 2 bajo la alternativa. Se observa que bajo la hipótesis alternativa, las potencias son muy próximas (o incluso iguales) a 1. Por otra parte, bajo la hipótesis nula, los resultados se acercan al nivel de significación α=0.05 para los diferentes valores de p.

AGRADECIMIENTOS

La primera autora está financiada por el programa FPU del MECD con referencia FPU13/01371. Todos los autores agradecen al MINECO por la financiación parcial con la ayuda MTM2014-52876-R, al MICINN, por la ayuda MTM2011-22392, y a la Xunta de Galicia, por la ayuda GRC CN2012/130.

			p =0.3	p =0.5	p =0.7	p =0.8
Modelo 1	CvM	$n = 100$	0.059	0.048	0.047	0.037
		$n = 200$	0.042	0.045	0.050	0.045
	KS	$n = 100$	0.062	0.063	0.052	0.037
		$n = 200$	0.045	0.057	0.053	0.042
Modelo 2	CvM	$n = 100$	0.061	0.056	0.040	0.041
		$n = 200$	0.043	0.053	0.044	0.039
	KS	$n = 100$	0.070	0.066	0.045	0.040
		$n = 200$	0.053	0.061	0.048	0.040

Tabla 1: Potencia del test bajo la hipótesis nula (α =0.05).

		Modelo 1	Modelo 2
CvM	$n = 100$	0.999	0.733
	$n = 200$	1.000	0.974
KS	$n = 100$	0.999	0.731
	$n = 200$	1.000	0.979

Tabla 2: Potencia del test bajo la hipótesis alternativa.

Referencias

[1] Delgado M. A., González-Manteiga W. (2001). Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics* 29, 1469–1507.

[2] López-Cheda A., Cao R., Jácome M. A., Van Keilegom I. (2016). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Artículo enviado (2^a revisión).*

[3] Xu J., Peng Y. (2014). Nonparametric cure rate estimation with covariates. *The Canadian Journal of Statistics* 42, 1–17.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Predicción online de la gripe

M. Oviedo de la Fuente[1], M. Febrero Bande[1] , L. Basile[2], N. Torner[2], *A.* Martínez[2]

[1]Dpt. de Estadística e Investigación Operativa, Univ. de Santiago de Compostela.
[2]Subdirecció gral. de vigilància i resposta a emergències de salut pública, Agència de Salut Pública de Catalunya, Generalitat de Catalunya.

RESUMEN

Los modelos de regresión son ampliamente utilizados por profesionales que trabajan en medicina (datos biométricos, epidemiológicos), biología (datos ecológicos), medio ambiente (datos de sensores, imagenes satélite) e incluso en ciencias sociales (datos de redes sociales). Los modelos lineales generalizados (GLM) son un punto de partida debido a la gran flexibilidad para modelar variables pertenecientes a la familia exponencial. El objetivo de este trabajo es extender las ideas de los modelos GLM cuando las observaciones estén correladas (dependencia en el espacio y/o tiempo), las covariables tengan naturaleza funcional (o de alta dimensión) y su relación con la variable respuesta sea no lineal. La metodología propuesta se aplica a un estudio real en el que se realiza una predicción online de la gripe en Cataluña y se divulgan los resultados mediante la generaión automática de informes y una página web interactiva.

Palabras e frases clave: Predición, Gripe, Modelos, Regresión, Funcional

1. INTRODUCCIÓN

La gripe es una infección respiratoria aguda que año tras año representa una importante carga de enfermedad. Las epidemias anuales se asocian a un incremento de las hospitalizaciones y de la mortalidad en personas a riesgo para presentar complicaciones graves como neumonía, así como a un incremento en la demanda de servicios asistenciales en general.
La modelización geográfica de estos eventos juega un papel importante en la mitigación del impacto de las epidemias estacionales de gripe que, en coordinación con las redes de vigilancia centinela de la gripe, puede ayudar a la mejor planificación de los servicios. Además, las variables metereológicas como temperatura y humedad permiten predecir a tiempo real, con mayor precisión y ademas son fáciles de conseguir [1].
El objetivo de este estudio es predecir con dos semanas de antelación la intensidad de la actividad gripal en las diferentes regiones de Cataluña utilizando la mejor inofrmación disponible.

4. PREDICCIÓN ONLINE DE LA GRIPE

En el estudio la variable respuesta es la intensidad de la actividad gripal (casos por 100,000 habitantes) en la semana $n + h$ que denotaremos por $y_{n+h,s}$ done h es el horizonte de predicción y s las regiones de Cataluña. Nótese que para $h = 1$ tenemos la predicción a una semana vista y $h = 2$ dos semanas vista.

En una primera fase del estudio se realizó una prueba piloto para la predicción de la temporada 2014-2015. En dicha fase se utilizaron los datos históricos en la estimación del modelo y selección de variables, para ello se construyeron base de datos que recopilaban información de múltiples fuentes de datos que aglutinaban variables de diferentes naturaleza (univariante, multivariante o funcional) como:

1. La tasa gripal en Cataluña $y_{n,s}$ y cobertura de los centros declarantes $yc_{n,s}$ en el sistema de enfermedades de declaración de obligatoria, `http://grip.gencat.cat/ca`.

2. Tasas de incidencia y porcentaje de aislamientos del virus de la gripe en la red de médicos centinelas de Cataluña, programa PIDIRAC: $centi_{n,s}$ y $aisla_{n,s}$.

3. Se utiliza la información proporcionada por Servicio de Meteorológico de Cataluña `www.meteo.cat` para evaluar el efecto de las variables meteorológicas en la actividad gripal. La naturaleza de estas variables es continua por lo que utilizamos herramientas del análisis de datos funcionales en nuestra modelización. Esto nos permite evaluar determinadas condiciones climáticas, como por ejemplo si una ola de frío contribuye a un aumento de la intensidad de la actividad gripal. Se crean las siguientes covariables funcionales: $\mathcal{X} = \{Temp_{n-t/7,s}, Humedad_{n-t/7,s}, Radiacion(t)_{n-t/7,s}\}$, donde t toma valores entre $[-14, 0]$, es decir, los valores observados durante los últimos 15 días (2 semanas).

4. Indice de Google FluTrends en Cataluña $FluCat_{n,s}$, en la Comunidad Valenaciana $FluVal_{n,s}$, en Aragón $FluAra_{n,s}$ y el global de España $FluEsp_{n,s}$, `http://www.google.org/flutrends`[1].

5. Variables calendario: índice semana, año, indicadores de días festivos y vacaciones.

El modelo lineal generalizado (GLM) [2] permite analizar la relación entre una variable respuesta ($y_{n+h,s}$) y una o varias covariables multivariantes ($\mathbf{X}_{n,s} \in \mathbb{R}^p$). El modelo lineal generalizado funcional (FGLM) [3] extiende los GLM al caso funcional y por tanto permite incluir las variables funcionales $\mathcal{X} = \mathbf{X}(t) \in \mathcal{L}^2$, siendo $\mathcal{L}^2$ el espacio de funciones integrables sobre t:

$$g(y_{n+h,s}) = \alpha + \mathbf{X}_{n,s}\beta + \sum_{k=1}^{K} \left\langle \mathcal{X}_{n,s}^k, \beta_k(t) \right\rangle + \varepsilon_{n+h,s}, \quad \forall h = \{1, 2\}. \tag{1}$$

dónde g en la de función de enlace, α en el intercepto, $\beta \in \mathbb{R}^p$, $\beta(t) \in \mathcal{L}^2$ y ε_i es el término de error, $(\varepsilon_1, \ldots, \varepsilon_{\text{ffln}ch}) \sim \mathcal{N}_n(0, \Omega)$. En el estudio utilizamos la regressión de Poisson con $g = log()$ y la población de personas susceptibles de contraer el virus de la gripe como parámetro de $offset$.
Muchos de los autores que utilizan los modelos anteriores consideran que el error ε es un proceso homocedástico independiente, i.e. $\mathbb{E}[\varepsilon] = 0, \mathbb{V}\text{ar}[\varepsilon] = \sigma^2$, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$, es decir $\Omega = \sigma\mathbb{I}^2$. Sin embargo, los casos de gripes normalmente están asociados a periodos epidémicos y por tanto sus observaciones están correlacionadas en el tiempo. Es común también la aparición de brotes o un aumento elevado de los casos correspondiente al contagio del virus por dependencia espacial.
El modelo GLS (generalized least squares) [4] extiende el modelo lineal tanto datos dependientes en el tiempo como para datos dependientes en el espacio. En los modelos GLS las dependencia temporales son estimadas mediantes estructuras de correlación ARMA(p,q), por ejemplo en un AR(1): $\varepsilon_i = \theta\varepsilon_{i-1} + \varepsilon_i$ con $|\theta| < 1$, $\mathbb{E}[\varepsilon] = 0$, $\mathbb{V}\text{ar}[\varepsilon_i] = \tau^2$ y $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.
Se ajustó el modelo 1 con diferentes variables y estructuras de dependencia ARMA(p,q) en los errores. El modelo con menor error cuadrático medio de predicción corresponde al modelo con: $\mathbf{X} = \{y_{n,s}, yc_{n,s}, centi_{n,s}, aisla_{n,s}\}$, $\mathcal{X} = \{Temp_{n,s}\}$ y con errores modelados mediante un $ARMA(2, 0)$. Dicho modelo es utilizado de manera automática en la segunda fase: predicción online de la temporada 2015-2016. Por ejemplo, la figura 1 muestra la evolución de la tasa observada en Cataluña hasta la semana 9 del 2016 y la predicción en las semanas 10 y 11, mientras la figura 2 muestra el nivel de la actividad gripal por región sanitaria en la semana 10 del 2016 y la predicción en las semanas 11 y 12.
Los resultados del estudio se han divulgado en los siguientes formatos:

[1] El servicio dejó de suministrar datos en agosto de 2015

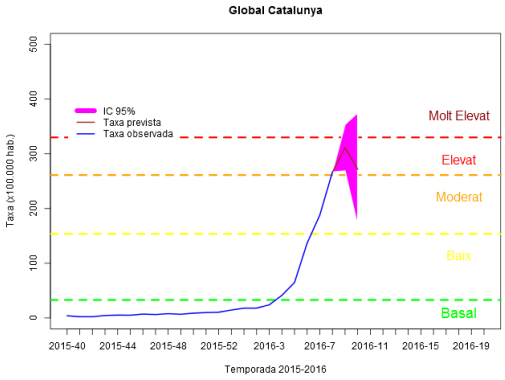


Figura 1: Tasa de incidencia de gripe en la semana 9 y su predicción en las semanas 10 y 11.

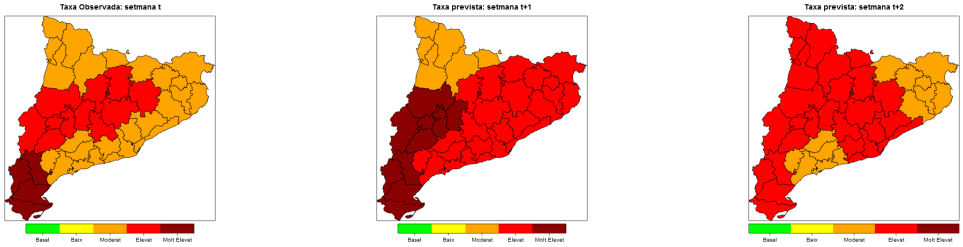


Figura 2: Nivel de la actividad (semana 9) por región sanitaria y su predicción (semanas 10 y 11).

- **Web Interactiva Shiny**

  Shiny es un paquete desarrollado por el grupo RStudio (`http://shiny.rstudio.com/`) que permite realizar interfaces web para la creación de informes interactivos y visualizaciones utilizando R. Estas aplicaciones permiten especificar, de forma interactiva, parámetros de entrada (inputs) utilizando controles deslizantes, menús desplegables, campos de texto, etc. Los resultados (outputs) que son generados en la web pueden se visualizarse como: gráficos, tablas, resúmenes, mapas, etc.

  Se ha realizado una aplicación Shiny que permite visualizar la predicción de la gripe en Cataluña en tres formatos (1) mapa de la predicción por comarcas (2) tabla con los valores previstos para un horizonte de dos semanas y (3) gráfico de la evolución de los valores previstos, `https://manueloviedo.shinyapps.io/gripRS/`.

- **Informe predictivo de la gripe usando R Markdown: dinámico y automático**

  R Markdown es un formato de edición que permite la creación de documentos dinámicos, presentaciones o informes utilizando R. Combina la sintaxis básica de markdown con trozos de código R integrados (llamados chunks). Cada chunk incluido en el documento se ejecuta al compilarlo y proporciona resultados que pueden incluirse en el documento final. Los documentos R Markdown son totalmente reproducibles (se pueden regenerar automáticamente siempre a partir del código R subyacente).

  Durante la temporada gripal 2015-2016 se esta ejecutando el modelo predictivo que sólo precisa de la actualización de las base de datos y cuyos resultados se divulgan periódicamente en la web del Departamento de Salud de Cataluña: informe predictivo.pdf

4. CONCLUSIONES

Esta aplicación permite monitorizar el estado actual de la gripe de forma online y obtener predicciones que ayuden a los servicios asistenciales a planificar su actuación en el territorio con suficiente antelación, proporcionando un ahorro en recursos (personal, económicos, equipamiento,...).

En el futuro se pretende incorporar más funtes de información útil para el modelo, también se planteará estimar estructuras de dependencia espacial a nivel de centro de declaración en lugar de región sanitaria o comarca.

AGRADECIMIENTOS

Xarxa sentinella de vigilància PIDIRAC 2014-2016. Este trabajo ha sido parcialmente financiado por el proyecto MTM2013-41383-P del Ministerio de Ciencia y Tecnología, España.

REFERENCIAS

Referencias

[1] Febrero-Bande, M., Oviedo de la Fuente, M. (2014). *Functional Regression Models with Temporal and/or Spatial Dependence*. Contributions in infinite–dimensional statistics and related topics.

[2] McCullagh, P., Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall.

[3] Müller, H. G., Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, 774–805.

[4] Kariya, T., Kurata, H. (2004). Generalized least squares. John Wiley & Sons.

Comunicacións orais

Sesión I: Métodos estatísticos en Ciencias do Medio Ambiente

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Tecnología láser escáner 3D en inventario forestal

Julia Armesto[1], Juan Picos[1] y Xana Alvarez[1]

[1] Escuela de Ingeniería Forestal, Campus Universitario A Xunqueira, CP 36005 (Pontevedra) Spain

RESUMEN

La toma de decisiones en el ámbito de la gestión forestal se sustenta habitualmente en información tomada mediante inventario de existencias. Hoy en día la mayor parte de los inventarios se realizan mediante mediciones con herramientas tradicionales, por lo que la precisión de los datos depende de la calidad y cantidad del trabajo de campo realizado. El láser escáner terrestre es una técnica de medida que proporciona coordenadas tridimensionales de millones de puntos de un entorno sin contacto con el objeto, con precisión milimétrica y en pocos segundos. Esto permite acometer la medición de variables dasométricas características de forma mucho más rápida, automática y precisa que por los métodos tradicionales. El empleo de esta técnica se generalizará en la gestión forestal cuando se introduzcan las mejoras oportunas en el software de tratamiento de nubes de puntos y se desarrollen protocolos operacionales adecuados para optimizar el uso de esta tecnología en esta aplicación. Este artículo analiza el potencial actual de la técnica láser escáner terrestre en la caracterización de variables dasométricas a través de casos prácticos.

Palabras y frases clave: láser escáner, LiDAR, inventario forestal.

1. INTRODUCCIÓN

El Láser Escáner Terrestre (TLS) es el resultado de trasponer la tecnología Light Detection and Ranging (LiDAR) a la obtención de medidas tridimensionales de escenarios terrestres. Son equipos de barrido masivo de puntos que permiten obtener miles o millones de puntos indiscrimandos de toda la escena. Se trata de una técnica que no requiere contacto con el objeto a medir, no es destructiva, invasiva ni selectiva.

Los instrumentos TLS se han empleado como instrumento topográfico desde su aparición en la década de los noventa en muy diversas aplicaciones. En [1] se muestra una aplicación a documentación y caracterización de elementos patrimoniales, [2] monitorización de taludes, [3] planificación urbanística, [4] modelado de edificios.

2. SISTEMA LÁSER ESCÁNER

Un TLS, denominado también LiDAR terrestre, opera midiendo la distancia recorrida por un haz láser y el intervalo angular preciso del mecanismo de deflexión del haz empleado. El resultado son coordenadas tridimensionales de los puntos tomados sobre la superficie de los objetos circundantes. El conjunto de puntos medidos se denomina nube de puntos.

Para medir la distancia entre el centro emisor del haz láser y el objeto a medir se pueden emplear básicamente dos técnicas: tiempo de vuelo y comparación de fase. Para la medición por comparación de fase se modula la amplitud de la onda y se mide la distancia a alta frecuencia. El sistema de tiempo de vuelo se basa en la medición del tiempo transcurrido entre la emisión y la recepción del haz. Típicamente, los equipos que incorporan este último sistema son más pesados y lentos, pero proporcionan más alcance en la

medida. Sea cual sea el principio de medida, los sistemas láser escáner permiten obtener nubes de puntos muy densas a tasas de medición muy elevadas (hasta un millón de puntos por segundo). En [5] se incluye una descripción completa de los principios de medición y las tipologías de instrumentos. Las características técnicas del instrumento Faro Focus 3D empleado en este artículo se detallan en la tabla 1.

Marca	FARO
Modelo	Focus3D
Tasa de emisión	976.000 puntos/s
Precisión	+-2mm/25m
Alcance	120m
Campo de visión (H/V)	305º/360º
Longitud de onda	905nm

Tabla 1: Características técnicas de Faro Focus 3D.

3. METODOLOGÍA

La medición mediante TLS de una parcela de inventario forestal para la caracterización de variables dasométricas incluye: planificación del proyecto de escaneo, toma de datos, registro de nubes de puntos, preproceso y medición de variables.

La planificación del proyecto de escaneo contempla los siguientes aspectos:

- Determinación de número y localización de los escaneos. A menor número de escaneos el tiempo de operación en campo es también menor; no obstante es preciso asegurar el solape entre nubes de puntos a fin de asegurar la cobertura total de la superficie de los fustes.
- Oclusiones. Tanto la radiación infrarroja como la visible, que son las longitudes de onda empleadas en los sistemas láser escáner, carecen de poder de penetración en la materia; por ello aquello que no es visible para el operario desde la posición de escaneo tampoco puede ser medido con el equipo. Cada objeto proyecta una sombra cuyo tamaño depende de la distancia al instrumento. En un entorno forestal es difícil evitar las oclusiones debido al patrón de distribución de pies en el terreno. Una planificación de campo inadecuada puede resultar en que queden pies sin medir o medidos de forma incompleta debido a las oclusiones. Se estima que desde el centro de la parcela hasta el 40% de los pies no son detectables [6]. El modo fin de minimizar oclusiones en el levantamiento de una parcela forestal es diseñar una adquisición multiestación, esto es, estacionar el equipo en el centro de la parcela y además en los límites de la misma, determinando el número de escaneos en función de la densidad de plantación.
- Resolución. La distancia entre puntos medidos es uno de los parámetros esenciales a configurar. A mayor resolución mayor número de puntos medidos sobre la superficie del objeto y por tanto con mayor fidelidad se conseguirá representar la geometría del mismo. El límite inferior viene impuesto por la divergencia del haz y la huella del mismo. En [7] se desaconseja un solapamiento entre haces sucesivos superior al 86%. El equipo FARO Focus3D tiene un haz de 3mm en la salida, y una divergencia de 0,19 mrad. Esto implica que a 10 m la resolución aconsejable es de 5,8 mm. Por ello se emplea el modo ¼ de la resolución total, que equivale a 6,136 mm a 10m.
- Sistema de referencia. Se pueden definir tres sistemas de coordenadas: 1) El sistema de coordenadas del equipo, que está centrado en su centro óptico. 2) El sistema de coordenadas de proyecto al que se habrán de trasponer todos los escaneos de un mismo levantamiento. 3) El sistema de coordenadas global es el que refiere el anterior a un sistema cartográfico oficial.

La toma de datos con FARO Focus3D conlleva unos 10 a 15 minutos por estación, incluyendo el desplazamiento, estacionamiento, nivelación del equipo, configuración de variables de escaneo y medición. Los datos de salida pueden contener las observaciones directas (distancia al objeto, ángulo horizontal, ángulo vertical) o indirectas (coordenadas X, Y, Z) de los puntos medidos. En el caso del sistema FARO Focus3D se obtienen las coordenadas cartesianas 3D estructuradas en forma de matriz. Además, se incluye una cuarta columna correspondiente a la intensidad de la señal. Esto puede ser útil para algunas aplicaciones, ya que la respuesta espectral de los objetos ante el haz incidente puede permitir en determinadas circunstancias aplicar técnicas de clasificación para el procesamiento automático de la nube de puntos. Finalmente, muchos sistemas TLS incorporan una cámara externa o interna de espectro óptico;

en el caso de FARO Focus3D la cámara es interna. El registro TLS-cámara permite relacionar la matriz de coordenadas con la matriz de color, asignando valores RGB a cada punto de la escena.

Para registrar las nubes de puntos de cada escaneo en el sistema de coordenadas de proyecto es preciso determinar la matriz de rotación-traslación a aplicar a los datos de cada escaneo y aplicar una transformación Helmert 3D a toda la nube. Para ello se pueden emplear dianas de registro o métodos iterativos de ajuste global nube a nube ([8], [9]). Los métodos de ajuste global son sensibles al movimiento de los objetos presentes en la escena; en entornos forestales las copas de los árboles no permanecen inmóviles en los sucesivos escaneos, ya que el viento puede introducir desplazamientos considerables. Es por ello que se descarta este método de ajuste en los casos de estudio. El ajuste por dianas consiste en el emplazamiento de dianas esféricas de 14,5 cm de diámetro y elevada reflectividad. Al menos deben ser visibles tres de ellas desde dos escaneos sucesivos. Se seleccionan los puntos correspondientes a la esfera y se estima el centro mediante ajuste paramétrico. Una vez determinados los centros se calcula la matriz de rotación-traslación y se aplica al resto de los puntos de la nube. De este modo se pueden conseguir errores de registro de pocos milímetros.

El preproceso de la nube de puntos incluye el filtrado de la nube de puntos mediante eliminación de puntos aislados y la homogeneización mediante una estructuras octree. Además se segmenta la nube para aislar la parte correspondiente a la parcela a medir, y se determina la paleta de color óptima para la interpretación. En el caso de estudio se empleó una paleta arco iris por valores de Z, alternándola con la visualización en pseudo color real (ver imagen 1).

La fase de medición de variables dasométricas se aborda aislando hileras de pies, y proyectando de forma ortogonal la nube de puntos de modo que se pueda seleccionar un punto en la base y otro en el ápice de la copa. De este modo se mide la altura de copa. Se rota la vista para comprobar que los puntos seleccionados son correctos. Para la medición del diámetro nominal se realiza una sección de la nube de puntos a 1,3m del suelo y se seleccionan dos puntos sobre la superficie del fuste perpendiculares al eje del mismo; a continuación se obtiene la vista en planta de la sección para verificar que los puntos seleccionados se corresponden con puntos distales de un diámetro.

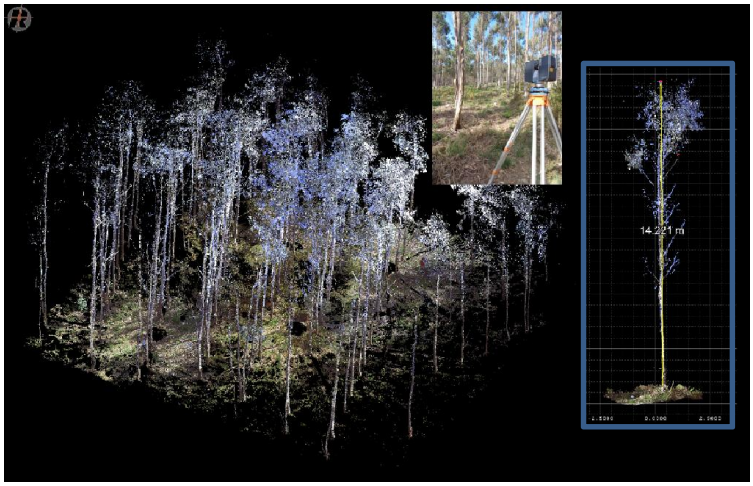


Figura 1. Nube de puntos TLS (vista en perspectiva), Faro Focus3D, y medición de altura de copa.

Se pueden presentar particularidades que introducen desviaciones en la metodología de medición. La imbricación entre copas dificulta la diferenciación entre las mismas y por tanto la medición de la altura de copa. La presencia de vegetación de sotobosque muy densa puede introducir oclusiones que impiden tomar puntos sobre la base del fuste o incluso sobre el fuste a la altura nominal.

Realizadas las medidas de las variables diámetro y altura para los pies de la parcela, se contrastan con las medidas de inventario tomadas con herramientas tradicionales. Para datos del monte Aboal, municipio de Mondariz, Pontecedra, poblado de *Eucalyptus globulus,* los coeficientes de correlación son muy elevados para la variable diámetro (ver figura 2). No así para la variable altura, donde se obtienen coeficientes de correlación R^2=0,5001. Esto puede deberse a que la medición de alturas mediante hipsómetro tiende a sobreestimar la lectura debido a la línea de visión tangente a copa.

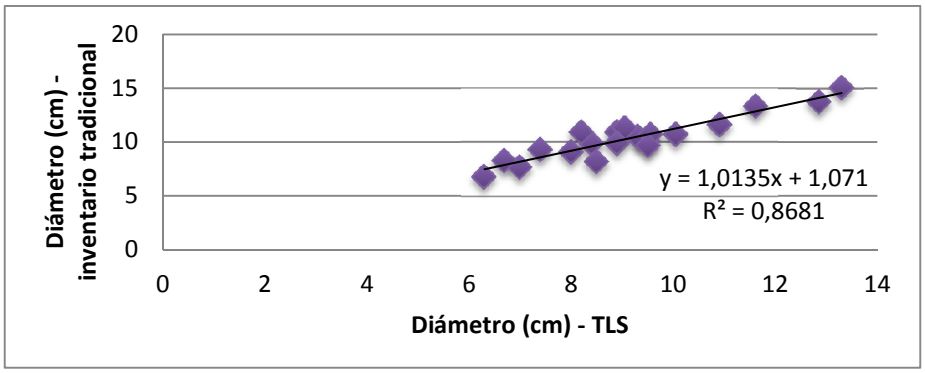


Figura 2. Correlación entre variables de inventario y TLS.

3. CONCLUSIONES

Este artículo describe una metodología operacional para aplicar la tecnología láser escáner terrestre a la medición de parcelas forestales y caracterización de variables dasométricas. El objeto es mostrar la aplicabilidad de estos instrumentos en la aplicación propuesta. Los TLS proporcionan información y detalle inaccesible para los sistemas LiDAR aéreos. Con un adecuado procesamiento es posible caracterizar variables dasométricas con precisión. La celeridad en la medida de millones de puntos y los avances en la tecnología en sí hacen que hoy en día constituya una alternativa de gran interés para la tecnificación de los procesos de inventario forestal y la reducción incertidumbres y errores.

REFERENCIAS

[1] Pieraccini, M., Dei, D., Betti, M., Bartoli, G., Tucci, G., Guardini, N. (2014). Dynamic identification of historic masonry towers through an expeditious and no-contact approach: Application to the "Torre del Mangia" in Siena (Italy). *Journal of Cultural Heritage*, 15, 275–282.

[2] Alejano, L. R. Ordóñez, C., Armesto, J., Rivas, T. (2010). Assessment of the instability hazard of a granite boulder. *Natural Hazards,* 53 (1), 77-95.

[3] Rodríguez-Cuenca B., García-Cortés, S., Ordóñez, C., Alonso, M. C. (2015). Automatic detection and classification of pole-like objects in urban point cloud data using an anomaly detection algorithm. *Remote Sensing*, 7, 12680–12703.

[4] Pu, S.; Vosselman, G. (2009). Knowledge based reconstruction of building models from terrestrial laser scanning data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64, 575–584.

[5] Vosselman, G.; Maas, H.G. (2010). Airborne and terrestrial laser scanning. Ed. Whittles.

[6] Liang, X.; Kankare, V.; Hyyppä, J.; Wang, Y.; Kukko, A.; Haggrén, H.; Yu, X.; Kaartinen, H.; Jaakola, A.; Guan, F.; Holopainen, M.; Vastaranta, M. Terrestrial laser scanning in forest inventories. ISPRS Journal of Photogrammetry and Remote Sensing 2016, 115, 63–70

[7] Lichti, D. (2006), Angular resolution of terrestrial laser scanners. *The Photogrammetric Record* 21 (114), 141- 160.

[8] Besl, P.J., McKay, N.D., 1992. A method for registration of 3-D shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 14 (2). Institute of Electrical and Electronic Engineers, 239–256.

[9] Gruen, A. and Akca, D. (2005). Least squares 3D surface and curve matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 59(3): 151-174.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

NUEVOS MÉTODOS DE ESTIMACIÓN DEL NÚMERO DE INCENDIOS FORESTALES EN GALICIA

Miguel Boubeta[1], María José Lombardía[1], Manuel Marey[2] y Domingo Morales[3]

[1]Departamento de Matemáticas, Universidade da Coruña.
[2]Departamento de Ingeniería Agroforestal, Universidade de Santiago de Compostela.
[3]Departamento de Estadística, Matemáticas e Informática, Universidad Miguel Hernández de Elche.

RESUMEN

Los incendios forestales suponen una de las principales causas de destrucción forestal, especialmente en Galicia donde el número de incendios y la superficie quemada en los meses de verano alcanzan valores elevados. Los modelos mixtos de Poisson con efectos correlados dados por un proceso SAR(1) ofrecen una alternativa competitiva a la par que flexible cuando la variable respuesta es de naturaleza discreta, tal y como el número de incendios por área forestal. En ellos se asume una estructura de dependencia entre dominios que comparten características similares. Bajo este contexto, se ha obtenido el estimador del método de los momentos (MM) de los parámetros del modelo y se ha obtenido el correspondiente empirical best predictor (EBP) del parámetro natural y de los efectos aleatorios. El comportamiento del estimador propuesto es analizado frente a otros estimadores tal y como el estimador plug-in o las correspondientes versiones de estos estimadores bajo el modelo de Poisson mixto con efectos de área independientes. Además, se analiza la incidencia de la matriz de proximidad en el modelo, usando para ello diferentes criterios. La precisión del estimador es evaluada a través de un algoritmo bootstrap paramétrico que estima el error cuadrático medio (MSE). Finalmente la metodología desarrollada se aplica al estudio del número de incendios en Galicia por área forestal.

Palabras e frases clave: Bootstrap, empirical best predictor, incendios forestales, método de los momentos.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

A Periodic State Space Model to Monthly Long-term Temperature Data

Marco Costa[1,2] and Magda Monteiro[1,2]

[1]ESTGA - School of Technology and Management of Águeda, University of Aveiro, Portugal
[2]CIDMA - Center for Research & Development in Mathematics and Applications, University of Aveiro, Portugal

ABSTRACT

This work presents a periodic state space model to model monthly temperature data. Additionally, some issues are discussed, as the parameter estimation or the Kalman filter recursions adapted to a periodic model. This framework is applied to monthly long-term temperature time series of Lisbon.

Key-words: State space model, Kalman filter, periodic data, temperature.

1. INTRODUCTION

The rise in global temperature has been an increasing concern of several authorities. According to the Intergovernmental Panel on Climate Change, the world's greenhouse gas emissions are still increasing and on the present path, the global temperature rise will substantial exceed the limit goal of two degrees Celsius that countries have agreed upon in order to avoid the most dangerous impacts of climate change. In the United Nations (UN) Framework on Climate Change, it was adopted the Paris Agreement, on 12 December 2015, at the UN Climate Change Conference in France where parties committed to take ambitious actions to keep global temperature rise below 2 degrees Celsius by the end of the century [9].
Climate change monitoring is performed with multidisciplinar approaches. However, the statistical analysis has been a key in this theme. Usually, monitoring climate change is performed based on regional or global temperature long-term data. Indeed, temperature data is considered in several studies of climate change or as the main variable under analysis or as a covariate on modeling other environmental variables, for instance in rainfall modeling. Thus, long-term temperature time series are, par excellence, the most appropriate data to study climate change, mainly in the global or local warming.
Indeed, modeling and foresight scenarios of environmental and climate variables are based on the selection of climate models using the analysis of other variables as rainfall and temperature [5]. On the other hand, the impact of changes in environmental variables changes, as the temperature or the precipitation, on environmental issues have been assessed and analyzed in order to explore the implications under a future climate [3].
The most applied statistical models in order to analyze climate time series, in particular of temperatures data sets, are the linear models and the usual time series models as ARIMA models [1, 2, 4]. These models are largely known as simple and they have well-known optimal properties under the usual assumptions (normality of errors, etc.). Some extensions and combinations of these models

have been proposed in environmental data analysis in order to improve models performance and to increase the predictions accuracy.
State space models have been largely considered an extension of the usual linear models which incorporate a versatile stochastic structure at the same time that integrate temporal dependence. The goal of this work is to analyze monthly long-term temperature time series of the Portuguese city of Lisbon. In [7] propose and implement a methodology for detection and correction of series of monthly temperatures for Lisbon, Oporto and Coimbra, providing the homogenization of these Portuguese long-term temperature data series (data in [8]). For instance, see Figure 1 related to Lisbon homogenized monthly temperature data.

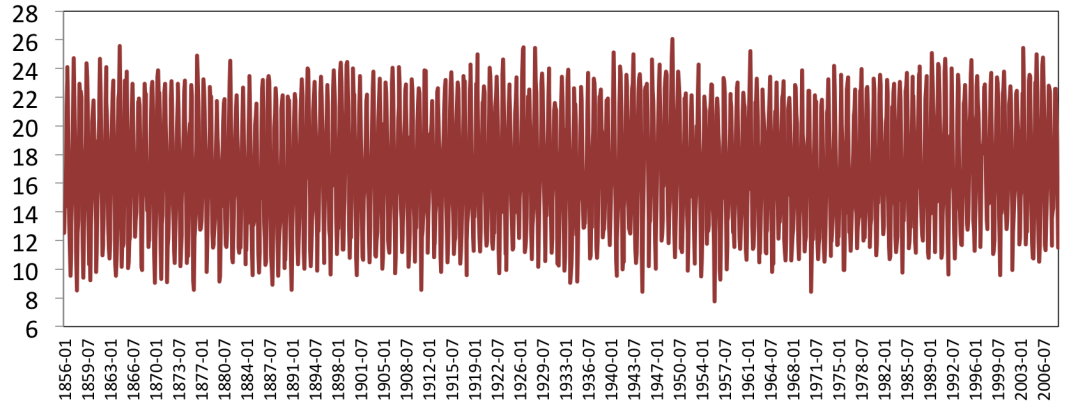


Figure 1: Lisbon homogenized monthly temperature data.

State space models have been largely applied in several areas of applied statistics. In particular, the linear state space models have desirable properties and they have a huge potential in time series modeling that incorporates latent processes.
Once a model is placed in the linear state space form, the most usual algorithm to predict the latent process, the state, is the Kalman filter algorithm. This algorithm is a procedure for computing, at each time t ($t = 1, 2, \ldots$), the optimal estimator of the state vector based on the available information until t and its success lies on the fact that is an online estimation procedure.
The main goal of the Kalman filter algorithm is to find predictions for the unobservable variables based on observable variables related to each other through a set of equations forming the state space model. Indeed, in the context of linear state space models, the Kalman filter produces the best linear unbiased estimators. When the errors and the initial state are Gaussian, the Kalman filter predictors are the best unbiased estimators in the sense of the minimum mean square error. However, the optimal properties only can be guaranteed when all model's parameters are known [6].

2. THE PERIODIC STATE SPACE MODEL - PSSM

In order to accommodate data properties as temporal correlation and the periodic behavior, a periodic state space model is proposed. This model is defined as

$$Y_{s,n} = [S(n-1)+s]X_{s,n} + D_{s,n}\beta + e_{s,n} \quad (1)$$

$$X_{s,n} = \mu_s + \phi_s(X_{s-1,n} - \mu_{s-1}) + \varepsilon_{s,n} \quad (2)$$

where

- s is the season of the year with $s = 1, 2, ..., S$;
- n is the year with $n = 1, 2, ..., N$;
- $Y_{s,n}$ represents the time series observation in the s^{th} season of the n^{th} year, that is, the $[S(n-1)+s]^{\text{th}}$ observation of the time series;

- $\beta = [\beta_1 \beta_2 ... \beta_S]'$ are unknown parameters representing the fixed effects in the model;
- $D_{s,n}$ is an $1 \times S$ matrix of known values, a design matrix;
- the error process $(e_{s,n})$ is a white noise disturbance, which is assumed to have $var(e_{s,n}) = \sigma_e^2$;
- μ_s is the mean of the process $(X_{s,n})$ for the s^{th} season, ϕ_s is the autoregressive parameter for season s and $\varepsilon_{s,n}$ is the white noise disturbance;
- the error process $(\varepsilon_{s,n})$ are distributed IID$(0, \sigma_s^2)$ and it is assumed to have an covariance defined by

$$cov(\varepsilon_{s,n}, \varepsilon_{s-i,n}) = \begin{cases} \sigma_{\varepsilon,s}^2 & i = 0 \\ 0 & i \neq 0 \text{ for } i = 1, 2, ..., S. \end{cases}$$

The first equation is referred as the observation equation and the second is called the state equation.

4. GAUSSIAN LIKELIHOOD ESTIMATION

Under the assumptions that the initial state, the state noise $\varepsilon_{s,n}$ and the observation noise $e_{s,n}$ are mutually independent and normally distributed, for a given realization $\mathbf{Y} = (Y_{1,1}, Y_{2,1}, ..., Y_{S,N})'$ the logarithm of the Gaussian likelihood function is computed as follows

$$\log \mathcal{L}(\Theta; \mathbf{Y}) = -\frac{NS}{2} \log(2\pi) - \frac{1}{2} \sum_{n=1}^{N} \sum_{s=1}^{S} \log(\Omega_{s,n}) - \frac{1}{2} \sum_{n=1}^{N} \sum_{s=1}^{S} \frac{\eta_{s,n}^2}{\Omega_{s,n}}$$

where $\Theta = (\beta, \sigma_e^2, \mu_1, ..., \mu_S, \phi_1, ..., \phi_S, \sigma_{1,\varepsilon}^2, ..., \sigma_{S,\varepsilon}^2)'$, $\eta_{s,n} = Y_{s,n} - \widehat{Y}_{s-1,n}$ is the innovation and $\Omega_{s,n}$ is the mean square error of the innovation. The forecast one-step-ahead of $Y_{s,n}$, $\widehat{Y}_{s-1,n}$, is obtained by the Kalman filter equations reformulated to the model (1)-(2). The maximum likelihood estimates are obtained through numerical procedures.

5. SOME RESULTS

The modeling procedure and the Kalman filter algorithm allow obtaining one-step-ahead forecasts and some overall results. In the first case, Figure 2 shows the one-step-ahead forecasts in the last decade of the period analyzed in Lisbon. It seems that the model has a good adjustment to the real data.

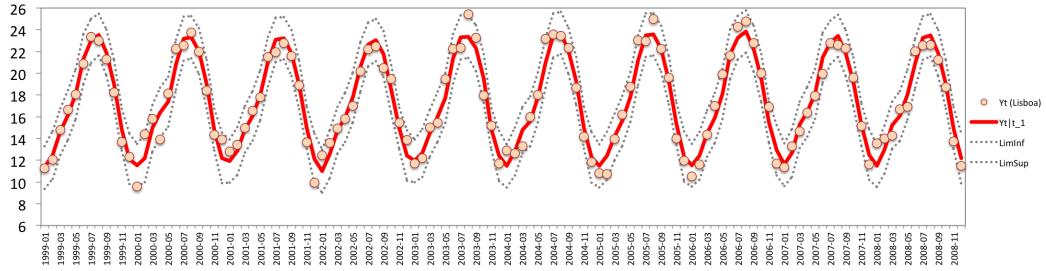


Figure 2: One-step-ahead forecasts in the last decade of the period analyzed in Lisbon.

On the other hand, the model's parameters allow to concluded that there is a global temperature rise. However, the months with the highest average rates of temperature increase by century are January (+0,9ºC), March (+1,1ºC), October (+0,8ºC) and December (+0,8ºC).

ACKNOWLEDGEMENTS

Authors were partially supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT– Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013.

References

[1] Alpuim, T., El-Shaarawi, A. (2009). Modeling monthly temperature data in Lisbon and Prague. *Environmetrics* 20 835–852.

[2] Bližňák, V., Valente, M.A., Bethke, J.E. (2015). Homogenization of time series from Portugal and its former colonies for the period from the late 19th to the early 21st century. *International Journal of Climatology* 35 2400–2418.

[3] Flannigan, M.D., Wotton, B. M., Marshall, G.A., Groot, W.J., Johnston, J., Jurko, N., Cantin, A.S. (2016). Fuel moisture sensitivity to temperature and precipitation: climate change implications *Climatic Change* 134 59–71.

[4] Freitas, C.R., Dedekind, M.O., Brill, B.E. (2015). A Reanalysis of Long-Term Surface Air Temperature Trends in New Zealand *Environmental Modeling & Assessment* 20 399-410.

[5] Guerreiro, S.B., Kilsby, C. G., Fowler, H.J. (2016). Rainfall in Iberian transnational basins: a drier future for the Douro, Tagus and Guadiana? *Climatic Change* DOI:10.1007/s10584-015-1575-z.

[6] Harvey, A.C. (1996). *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

[7] A.L. Morozova, M.A. Valente, Homogenization of Portuguese long-term temperature data series: Lisbon, Coimbra and Porto, *Earth Syst. Sci. Data* **4**: 187–213 (2012).

[8] A.L. Morozova, M.A. Valente, Homogenization of Portuguese long-term temperature data series: Lisbon, Coimbra and Porto, doi:10.1594/PANGAEA.785377 (2012).

[9] United Nations (2016). http://www.un.org/sustainabledevelopment/climate-change/ (accessed in January 22, 2016).

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Dynamic environmental assessment using DEA window analysis for wastewater treatment plants in Spain

Yago Lorenzo-Toja [1,*], Ian Vázquez-Rowe[1,2], Desirée Marín-Navarro[3], Rosa M. Crujeiras[4], María Teresa Moreira[1], Gumersindo Feijoo[1]

[1] Department of Chemical Engineering, Institute of Technology, University of Santiago de Compostela, 15782 Santiago de Compostela, Galicia, Spain
[2] Peruvian LCA Network, Department of Engineering, Pontificia Universidad Católica del Perú, 1801 Avenida Universitaria, San Miguel, Lima 32, Peru
[3] CETaqua, Water Technology Centre, 08940 Cornellà de Llobregat, Barcelona, Spain
[4] Department of Statistics and Operations Research, Faculty of Mathematics, University of Santiago de Compostela, Campus Sur s/n, C.P. 15782, Santiago de Compostela, Spain

ABSTRACT

The environmental analysis of numerous wastewater treatment plants (WWTPs) has been conducted following the Life Cycle Assessment (LCA) methodology in order to identify environmental impacts from a holistic perspective. However, most of these evaluations have focused either on a reduced number of WWTPs or on analysing a short time frame. In a recent study conducted on over 100 WWTPs in Spain, a combined LCA and Data Envelopment Analysis (DEA) study was performed, in which important differences in environmental impacts were observed between DMUs (i.e., individual WWTPs). Despite the depth of the study in terms of the number of WWTPs, it was limited to one year of assessment, preventing the study from identifying if efficiency values were linked to a specific trend of each DMU through time or to random interannual changes. Therefore, the current study applies a DEA window analysis to a four-year interval (2009- 2012) with the aim of understanding the interannual behaviour of WWTPs in terms of environmental sustainability. The results obtained suggest that regardless of the size of the WWTPs, efficient plants tend to maintain their performance through time. In addition, the average efficiencies tend to be fairly stable through the analysed period. However, a significant difference in efficiency values was identified between large and medium WWTPS, on the one hand, and smaller plants, on the other, that operate at lower relative efficiency values. In order to delve into the productivity variability across the four-year period the Malmquist productivity index (MPI) was calculated based on the DEA window efficiencies results. The decomposition of the MPI allowed unravelling the driving forces of the productivity improvement identified.

Keywords: Data Envelopment Analysis; eutrophication; Life Cycle Assessment; wastewater; windows analysis

1. MATERIALS AND METHODS

The present study aimed at overcoming the absence of a time-dependent analysis by means of an intensive data collection process throughout four years of assessment (2009-2012) for a considerable number of WWTPs. The initial sample was composed by operational and water quality data for 474 Spanish WWTPs; after a thorough screening process, the final sample was reduced to 47 due to data gaps and unreliable figures. Thereafter, they were divided taking into account the population equivalent (p.e.) served in three main size ranges (small from 0 to 20,000 p.e., medium from 20,000 to 50,000 p.e. and large above 50,000 p.e.), as proposed in Lorenzo-Toja et al. (2015).

Considering the available data, three different inputs were selected to be included in the DEA matrices based on their operational importance and environmental impact contribution: i) electricity use at the WWTPs; ii) consumption of chemicals; and, iii) sludge production. For inputs 1 and 3 the methodology applied was the direct computation of the inventory data. Nevertheless the in the case of the input 2, due to the different characteristics and amount of the chemicals used in WWTPs, a homogenization step was performed by means of the weighted ReCiPe endpoint life cycle impact assessment method. With regard to the output, a critical concern was considered: it should summarise the function of a WWTP, which from a environmental point of view is the removal of eutrophication potential (EP) (Larsen et al., 2007). Consequently, a new indicator named Eutrophication Net Environmental Indicator (ENEI) was developed for this study, to account not only for the net improvement in water quality due to the treatment, but also to bring forward the environmental benefits compared to the thresholds established by the legislation. Equation 1 represents the calculation approach followed for the calculation of the indicator:

$$\mathrm{ENEI} = \frac{\mathrm{EP_{influent}} - \mathrm{EP_{process}} - \mathrm{EP_{effluent(real)}}}{\mathrm{EP_{influent}} - \mathrm{EP_{effluent\ (theorical)}}} \quad \text{[Eq. 1]}$$

The DEA model chosen to evaluate WWTPS efficiency was the window analysis model, considering input orientation (i.e., minimize the use of raw materials) and constant return to scale (Charnes et al., 2013). While most of the DEA models deal with input/output flows in static conditions, this model allows the computation of dynamic settings to account for temporal variability in the use of material and energy flows, as well as in the generation of outputs products. Under this model each WWTP is considered an independent entity, named decision making unit (DMU), within each time period, allowing a comparison not only among facilities but also of the fluctuation in individual WWTP efficiencies along the period evaluated (Asmild et al., 2004). When conducting a window analysis through DEA, it is important to define the window length for the analysis, guaranteeing that it is large enough to have a sufficient sample size, but small enough to avoid unfair comparisons through the window. In this paper, two different window widths were selected: *p=1* and *p=4*; the former corresponds to a contemporaneous analysis (period by period), whereas the latter corresponds to a global intertemporal analysis (over the whole study period) (Tulkens and Eeckaut, 2006).

The non-parametric Malmquist Productivity Index (MPI) is a time-series analysis technique based on the application of distance functions. This scheme allows the performance comparison of DMUs over time by solving some DEA-type models; the MPI measures productivity changes over two periods based on the reference frontier under a multiple inputs/outputs framework. In the current study, the decomposition of the MPI in its two components: efficiency change (catching up) and technical change (frontier shift) will allow analysing the driving factors of the productivity change.

2. RESULTS AND DISCUSSION

Lorenzo-Toja et al. (2015) posed a final question of whether one single year of assessment for WWTPs was sufficient to identify their eco-efficiency using the LCA+DEA method. Therefore, the use of the window analysis model in the current study aimed at answering this specific research question. For this, the contemporaneous approach (i.e., *p=1*) was used. It appeared that WWTPs that operate efficiently in one year tend to operate at very high efficiencies in other periods. While there are differences between small, medium and large WWTPs, as well as important differences at an individual level, the average standard deviation for the matrices using *p=1* is in all cases below ±12%. To confirm these apparent trends that there are no significant differences between years in the different groups (see Figure 1), an adapted version of ANOVA for repeated measurements and the Friedman's test (the nonparametric par) were used. After analysing the p-values extracted, it can be concluded that in the case of medium and large clusters no significant differences in efficiency between years are found. Nevertheless, the high variability and heterogeneity in the small WWTPs group lead to contradictory results, showing differences between years (i.e. p-value=0.001).

Interestingly, the results were also run in order to identify if there were any significant differences between groups, fixing years. For this purpose, both ordinary ANOVA and Kruskal-Wallis tests were applied. In all cases (i.e., for all 4 years of analysis), differences were detected between the three groups, a result that supports the legitimacy of the performed descriptive grouping system in large, medium and small WWTPs. A statistical analysis was also performed in order to determine whether there were any

significant differences between the locations of WWTPs in terms of sensitive or non-sensitive water bodies. The ANOVA for repeated measurements test confirms (p-val=0.0276) that there are significant differences between the zones. This result gains importance in the sense that WWTPs in non-sensitive areas appear to be capable of attaining higher efficiency values repeatedly, whereas those in sensitive areas struggle to attain these higher values, especially for WWTPs below 100,000 p.e.

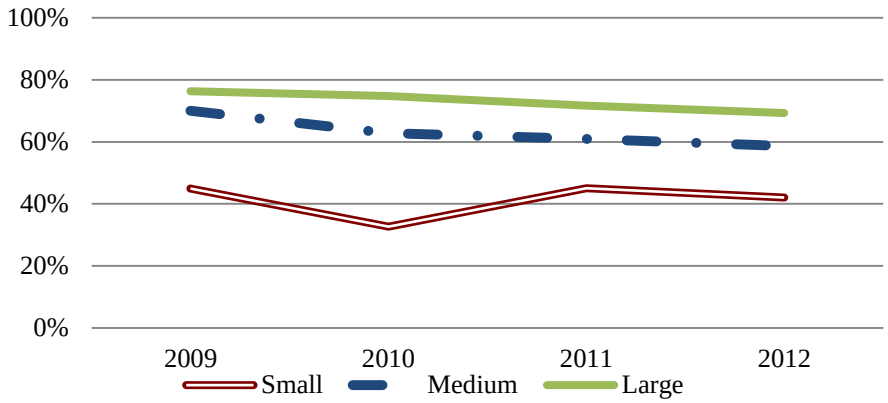


Figure 1. Average annual efficiency score (Φ) per analysed segment size.

An alternative to using the contemporaneous approach in window analysis, is to expand the window length that is being assessed to consider not only the entire set of DMUs for one single time period, but also the entire set of observations for the whole study period (Asmild et al., 2004). This perspective is based on the principle of moving averages, detecting performance trends of the WWTPs over time (Charnes et al., 2013). Therefore, in the current study the window was also elongated to the entire width of available data (i.e., 4 years) in order to determine whether any temporal trends could be identified. Given the lack of clear tendencies using the *p=4* window analysis, the MPI was used in order to analyse the productivity change in this period (Asmild et al., 2004; Färe et al., 1994). Based on this index, it was observed that the two matrices that presented similar trends in the window analysis, were capable of incrementing their productivity by 18.1% (small WWTPs) and 16.7% (large WWTPs), whereas medium WWTPs presented a decrease in their productivity of approximately 11%.

However, it should be noted that this productivity changes observed for the three matrices must be decomposed into the two sub-indices mentioned in section 1: technical change and "catching up". Hence, when the small and large WWTPs are compared in terms of these two indicators, it is evident that they actually present diverging tendencies. The former increase considerably their values through time in terms of efficiency change (+19.9%): however, the technical change only accounts for a minimal increase of +3.2%. This tendency suggests that technological improvement has been vague in this group of plants and, therefore, the efficiency frontier has remained stagnant through the analysed window. Interestingly, in line with the hypothesis suggested in Lorenzo-Toja et al. (2015), inefficiencies in these plants are attributed to a lack of continuous monitoring and supervision, although the tendency suggests that these operational inefficiencies are being reduced considerably through time.

In contrast, large WWTPs present a moderate reduction in "catching-up" (-6.8%), but a significant improvement in technical change (26.1%), which is probably linked to the implementation of more sophisticated and innovative technologies. Finally, medium WWTPs present a completely opposing behaviour to small plants, with a substantial change in the efficiency frontier (+23.7%), indicating technological innovation, and remarkable deterioration in terms of operational efficiency (-26.8%).

3. CONCLUSIONS

A revised analysis of a substantial sample of Spanish WWTPs was performed in order to understand their behaviour through time. Results demonstrated that an overwhelming number of WWTPs tend to maintain their efficiency standards through time without excessive variability as proved by means of the statistical

tests (i.e ANOVA and Friedman´s test). Hence, the use of a contemporaneous window analysis model proved that the previous use of an SBM model for one single year of operation constitutes a good proxy towards the accomplishment of efficiency standards for this type of production systems.

Having said this, a deeper analysis of the results, using an intertemporal window analysis and MPI, allows the identification of certain behavioural traits that may not be detected with an SBM model. For instance, the results under this perspective confirm that the lower average efficiencies identified in small WWTPs can be mainly attributed to operational shortcomings, whereas for medium and large plants technological improvements are a major carrying factor that accounts for changes in the efficient frontier.

In addition, the change in DEA output, shifting from the use of the NEB index to the ENEI indicator was performed with the intention of providing a more realistic measure of environmental improvements in WWTPs. This new methodological approach allowed an analysis on the influence of the discharge point and, thus, the different thresholds imposed by the legislation on WWTP efficiency. It has been concluded that for the sample considered, significant differences were found in WWTP efficiencies due to zone effect.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support of the AQUAENVEC project (LIFE10 ENV/ES/000520). Authors with affiliation to the University of Santiago de Compostela (Spain) belong to the Galician Competitive Research Group GRC 2013-032. Dr. Ian Vázquez-Rowe wishes to thank the Galician Government for financial support (I2C postdoctoral student grants programme). Rosa M. Crujeiras acknowledge the support of MTM2013-041383P and Interuniversity Attraction Poles Programme (IAP-network P7/06),

REFERENCES

Asmild, M., Paradi, J.C., Aggarwall, V., Schaffnit, C., 2004. Combining DEA window analysis with the Malmquist index approach in a study of the Canadian banking industry. J. Product. Anal. 21, 67–89. doi:10.1023/B:PROD.0000012453.91326.ec

Charnes, A., Cooper, W.W., Lewin, A.Y., Seiford, L.M., 2013. Data envelopment analysis: Theory, methodology, and applications. Springer Science & Business Media.

Färe, R., Grosskopf, S., Norris, M., Zhang, Z., 1994. Productivity growth, technical progress and efficiency change in industrialized countries. Am. Econ. Assoc. 84, 66–83. doi:10.1111/j.1467-8268.2004.00089.x

Larsen, H.F., Hauschild, M., Wenzel, H., Almemark, M., 2007. Homogeneous LCA methodology agreed by NEPTUNE and INNOWATECH, EU FP6 project, deliverable; No. 4.1.

Lorenzo-Toja, Y., Vázquez-Rowe, I., Chenel, S., Marín-Navarro, D., Moreira, M.T., Feijoo, G., 2015. Eco-efficiency analysis of Spanish WWTPs using the LCA+DEA method. Water Res. 68, 651–666. doi:10.1016/j.watres.2014.10.040

Tulkens, H., Eeckaut, P. Vanden, 2006. Nonparametric efficiency, progress and regress measures for panel data: Methodological aspects. Springer.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Dynamic fator Models for bivariate Count Data: an application to fire activity

Magda Monteiro[1], Isabel Pereira[2] e Manuel G. Scotto[3]

[1]ESTGA & CIDMA & Universidade de Aveiro
[2]Departamento de Matemática & CIDMA& Universidade de Aveiro
[3]Departamento de Matemática/IST &CEMAT & Universidade de Lisboa

ABSTRACT

The study of forest fire activity, in its several aspects, is essencial to understand the phenomenon and to prevent environmental public catastrophes. In this context the analysis of monthly number of fires along several years is one aspect to have into account in order to better comprehend this tematic. The goal of this work is to analyze the monthly number of forest fires in the neighboring districts of Aveiro and Coimbra, Portugal, through dynamic factor models for bivariate count series. We use a bayesian approach, through MCMC methods, to estimate the model parameters as well as to estimate the common latent factor to both series.

Keywords and key sentences: Dynamic latent variables; MCMC; bivarite Poisson distribution; wildfires.

1. INTRODUCTION

Wildefires are a very common phenomenon in Portugal and each year, a large amount of forest land is destroyed. In the years between 1996 and 2014, the average of fires were around 19 000 and the annual average of burned forest area was around 108 000 ha ([1]). The scale of these values is related with climate conditions, topography and also with a combination of various socio-economic factors including the large scale replacement of native Portuguese forest by fire-prone tree species (such as pine and eucalypt), which are more profitable, and the decline in traditional practices like grazing and forestry which helped to reduced the accumulation of flammable materials ([5],[4]). Is in this context that is extremely important to study forest fire activity in Portugal in its different aspects in which is included the study of time series that result from this activity such as monthly number of fires.
This work analizes the monthly number of forest fires in the neighboring districts of Aveiro and Coimbra, Portugal. Since we have two neighbour regions in the study, we can use bivariate models that allow potentially complex interdependence between series. As these data have a discrete nature we have to use appropriate multivariate models to describe them. The work in [2] presented a dynamic factor multivariate model to analyze Stock-Market Trading Activity. In this model, counts are assumed to be conditionally independently distributed with a discrete distribution whose averages are latent random variables which depend on a set of unknown factors which are autoregressive gaussian processes of order 1. In this work it will be used a bivariate dynamic model with one common latent factor and the temperature as an explanatory variable.

2. BIVARIATE DYNAMIC FACTOR MODEL

Consider a bivariate vector of counts $Y_t = (Y_{t,1}, Y_{t,2})$, recorded at time t (1- Aveiro, 2- Coimbra). In a dynamic multivariate count model the dynamics is introduced at the level of the latent factors. Therefore counts are assumed to be conditionally independently distributed with Poisson distribution

$$p(y_{t,j}|\lambda_{t,j}) = e^{-\lambda_{t,j}} \frac{\lambda_{t,j}^{y_{t,j}}}{y_{t,j}!},\ t = 1 \cdots, T;\ j = 1, 2 \tag{1}$$

whose means $\lambda_{t,j}$ are latent random variables. Having into account the sample autocorrelation functions in Figure , it is also also assumed that the logarithm of each component of the mean vector, $\boldsymbol{\lambda}_t = (\lambda_{t,1}, \lambda_{t,2})$, is a linear function of the type

$$\ln(\lambda_{t,j}) = \mu_{t,j} + \gamma_j Z_t + \beta_j X_{t,j},\ \ j = 1, 2, \tag{2}$$

where Z_t is an unknown random factor which is common to both components of the vector of counts, $X_{t,j}$ is the monthly average temperature in region j and $\mu_{t,j}$ is the intercept which depends on the month of the year, $\mu_{t,j} = \mu_{s,j} \times I(t = s + 12k)$, $s = 1, 2, \cdots, 12$ and some natural k. It was considered that $\gamma_1 = 1$ for normalization and to eliminate indeterminacies in the factor scale.

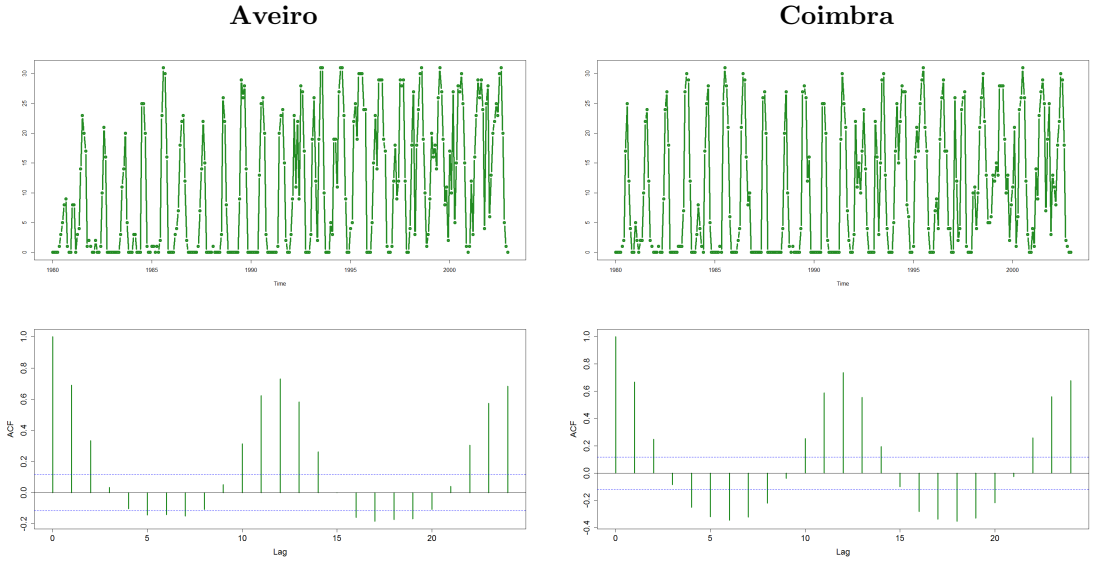


Figure 1: Monthly number of fires in Aveiro and Coimbra (top) and the respective sample autocorrelation functions (bottom).

The possibility of serial and cross-correlation in the counts is introduced through the structure of the latent factor Z_t. It is assumed that the factor follows an AR(1) Gaussian process, $Z_t|Z_{t-1} \sim N(\phi Z_{t-1}, \sigma^2)$.
The estimation of the model can be made either in a classical framework, using maximum likelihood estimation based uppon efficient importance sample, or through a bayesian approach, which will be considered here. Considering $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\boldsymbol{\theta}_1 = (\boldsymbol{\mu}, \boldsymbol{\beta})$ and $\boldsymbol{\theta}_2 = (\phi, \sigma)$, the vector of all parameters involved and $h(\boldsymbol{\theta})$ its prior distribution, and having into account that (Z_t) is an unknown Gaussian process, the prior distribution for the unknown components of the model can be written as $h(\boldsymbol{\theta}, \boldsymbol{z}) = h(\boldsymbol{\theta})h(\boldsymbol{z}|\boldsymbol{\theta}_2)$, so the posterior distribution of the unknown components can be obtained

$$h(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) = \frac{h(\boldsymbol{\theta}) \prod_{j=1}^{2} \prod_{t=1}^{T} f(z_1) f(z_t|z_{t-1}, \boldsymbol{\theta}_2) p(y_{t,j}|\lambda_{t,j})}{\int \cdots \int h(\boldsymbol{\theta}) \prod_{j=1}^{2} \prod_{t=1}^{T} f(z_1) f(z_t|z_{t-1}, \boldsymbol{\theta}) p(y_{t,j}|\lambda_{t,j}) d\boldsymbol{\theta} d\boldsymbol{z}}. \tag{3}$$

Since this distribution does not have a closed form, it will be use Markov Chain Monte Carlo methods (MCMC).

3. APPLICATION TO FIRE ACTIVITY

The bivariate dynamic factor model for counts introduced in Section 2 is used to analyze the dynamics and correlations of the monthly number of fires in Aveiro and Coimbra Counties in the period of january 1980 to December 2002. Although the data on the monthly of fires are known for a longer period of time, there are no available data about the temperature in the remaining period. The descriptive statistics are presented in Table 1 and there exists a positive correlation between series of counts. The estimation of the parameters of the model were made using a bayesian

	Aveiro	**Coimbra**
Mean	10.62	10.15
Median	5	6.5
standard dev.	11.49	10.66
Minimum	0	0
Maximum	31	31
correlation	0.74	

Table 1: Descriptive statistics for the number of fires

approach, through the use of MCMC methods with Metropolis-Hastings algorithm, using *Winbugs* software. Since we do not have prior information about parameters we used non-informative priors for them. We considered that parameters were independent, and for σ^2 we used the inverse Gamma with both parameters 0.001 and for the remaining parameters we used a normal prior distribution with a mean zero and a variance of 10000. A burnin period of 150000 were considered in order to the chain reaches the steady state (we use the package coda in R software), and a lag of 1000 to form a sample, with dimension 1000, from the posterior distribution. The results of the estimation procedure are presented in Table 2 and Figure 2 presents the predictions of the latent factor with 95% HPD intervals. The standardized Pearson residuals (see Table 3) do not present significant correlation but they present values of the means and standard deviations far from the expected values, i.e., the characteristics of mean zero and standard deviation 1. It was also analized the quantile quantile plot of randomized PIT residuals that revealed deviations from normality in the left tail. revealing that both models do not capture all the dynamics in the series. In order to improve the modeling results one can use other discrete distribution such as negative binomial, since dispersion index is equal to 1 in the Poisson distribution and also one can adjust the structure of the factor Z_t so it can capture all the dynamics in the series.

	Parameters						
City	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7
Aveiro	0.07134	0.9322	1.7120	1.6400	1.997	3.344	4.096
	(0.01878)	(0.02167)	(0.0166)	(0.0225)	(0.01933)	(0.0199)	(0.02052)
Coimbra	0.8402	2.0760	2.1910	2.1510	3.3700	4.786	5.4460
	(0.02167)	(0.02131)	(0.0249)	(0.01763)	(0.02338)	(0.02668)	(0.02951)
	μ_8	μ_9	μ_{10}	μ_{11}	μ_{12}	γ	β
Aveiro	4.239	3.842	2.269	-0.4684	-1.124	1	-0.0499
	(0.02066)	(0.01973)	(.01912)	(0.01898)	(0.01956)	–	(0.0005)
Coimbra	5.337	4.403	2.947	0.3264	-0.0.9631	0.9455	-0.1019
	(0.0293)	(0.0273)	(0.02455)	(0.02105)	(0.02349)	(0.00228)	(0.0012)
	factor parameters						
	ϕ	σ		**DIC**			
	0.8247	0.8585		2816.570			
	(0.0012)	(0.0020)					

Table 2: Parameter estimates (MC errors in brackets).

ACKNOWLEDGMENT

This work was supported in part by the Portuguese Foundation for Science and Technology (FCT-Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

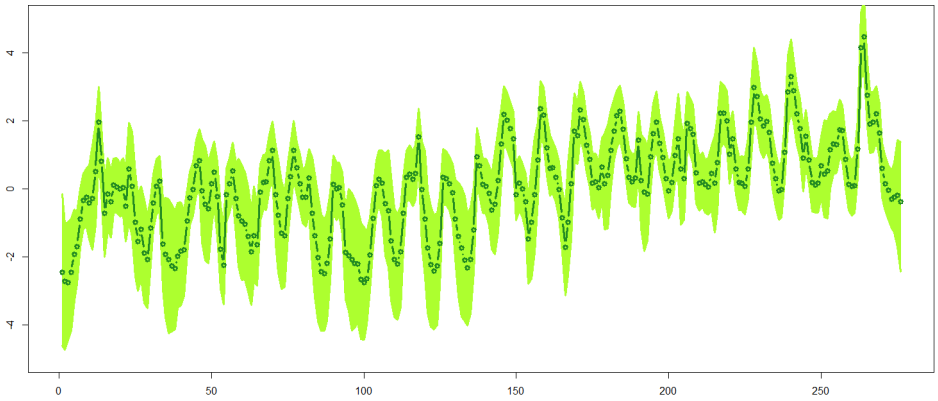


Figure 2: Predictions of factor Z and the respective 95% HPD interval.

Distribution	Region	Mean	Standard deviation	Portmanteu test p-value
Poisson	**Aveiro**	-0.105	0.8351	0.082
	Coimbra	-0.104	0.9276	0.239

Table 3: statistics of standardized Pearson residuals and Portmanteu test

References

[1] Instituto da Conservação da Natureza e das Florestas (2016). Estatística - dados sobre incêndios florestais.
http://www.icnf.pt/portal/florestas/dfci/inc/estat-sgif#tip (08/04/2016)

[2] Jung, R.C., Liesenfeld, R., Richard, J.F. (2011). Dynamic factor models for multivariate count data: an application to stock-market trading activity. *Journal of Business & Economic Statistics*, 29, 73–85.

[3] Paulino, C.D., Amaral Turkman, A., Murteira, B. (2003). *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa.

[4] Pereira, J.M.C., Carreiras, J.M.B., Silva, J.M.N., Vasconcelos, M.J., 2006. Alguns conceitos básicos sobre os fogos rurais em Portugal. In: Pereira, J.S., Pereira, J.M.C., Rego, F.C., Silva, J.M.N., Silva, T.P. (Eds.), Incêndios Florestais em Portugal: Caracterização, Impactes e Prevenção. ISAPress, Lisboa, pp. 133–161.

[5] Fernandes, Paulo (2007). Entender porque arde tanto a floresta em Portugal. In Árvores e Florestas de Portugal, Vol. 8 Proteger a Floresta - Incêndios, Pragas e Doenças. Chapter: 3, Publisher: Público, Comunicação Social, SA, & Fundação Luso-Americana para o Desenvolvimento, Editors: Silva J.S., pp.69–91.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

ESTIMACIÓN DEL NIVEL DE RIESGO CONTAMINANTE DE LOS RESIDUOS PETROLÍFEROS APLICANDO TÉCNICAS DE CLASIFICACIÓN FDA

Miguel Flores[1], Salvador Naya[2], Javier Tarrio-Saavedra[2], Luis Horna[1], Rolando Saez[3] y Ana Escobar[1]

[1]Departamento de Matemáticas, Facultad de Ciencias, Escuela Politécnica Nacional, miguel.flores@epn.edu.ec
[2]Departamento de Matemáticas, Escola Politécnica Superior, Universidade da Coruña
[1]Instituto Ciencias Básicas, Facultad de Ingeniería, Universidad Central del Ecuador

RESUMEN

En el proceso de extracción del petróleo, específicamente en el refinamiento e industrialización hidrocarburífera aparecen, como es conocido, múltiples desechos que son altamente contaminantes para la tierra, agua y aire. En este trabajo se estima el nivel de riesgo de estos desechos en zonas afectadas gracias a la aplicación de un modelo estadístico y a su implementación en un software que permite realizar una evaluación temprana, mediante un método semicuatitativo el nivel de riesgo (Bajo o Alto), por parte de los equipos humanos en los sitios cercanos a la afectación.

Se utilizó la técnica fluorescencia inducida por láser (LIF) la cual consiste en la espectroscopía que permite estudiar la estructura de las moléculas y la detección de las especies seleccionadas; los datos obtenidos mediante esta técnica se utilizan para el ajuste del Modelo Lineal Funcional Generalizado(MLFG), que posibilita clasificar el espectro generado a partir de la aplicación de la LIF en dos niveles de contaminación: Bajo y Alto.

Palabras y frases clave: Control de calidad, Fluorescencia inducida por láser, Modelo Lineal Funcional Generalizado, Clasificación, Contaminación, Hidrocarburos.

1. INTRODUCCIÓN

La filtración de petróleo (o de sus derivados), su transporte y su difusión-dispersión son procesos cuyo estudio es de vital importancia debido al gran impacto que tienen en la actividad humana y en el ambiente. La filtración de petróleo en suelos provoca un nivel de contaminación que es un problema muy complejo de evaluar, éste depende fundamentalmente de los siguientes elementos: tipo de suelo, porosidad, conductividad hidráulica, y las propiedades del petróleo tales como la densidad y la viscosidad. Por tal motivo, una tarea importante es determinar el estado del sistema en todo momento, y de forma prioritaria en el instante inicial, ya que permitiría la aplicación de medidas correctivas en el ecosistema de un modo más eficaz.
El petróleo está formado por una gran variedad de compuestos, algunos de los cuales producen fluorescencia cuando son iluminados con luz ultravioleta. La fluorescencia del petróleo es depende en gran medida de su composición química. Por tal motivo existen técnicas analíticas para la caracterización del petróleo crudo, en las que la intensidad y tiempo de vida de la fluorescencia se relacionan con la composición química y densidad (API) del petróleo.

Si juntamos una fuente de luz láser ultravioleta, un espectrómetro y petróleo, tendremos un sistema para detectar presencia de petróleo, tal como es el caso de las tierras contaminadas. La luz láser produce fluorescencia cuando hay petróleo en la tierra, la cual se detecta utilizando el espectrómetro. Como cada variedad de petróleo presenta un espectro de fluorescencia característico, las técnicas de fluorescencia se suelen utilizar para su identificación.

Los datos obtenidos por esta técnica son utilizados para ajustar un MLFG que permite resolver el problema de clasificación supervisada descrito. Este modelo asume que el contenido del elemento sólido permanece constante y que es indeformable. Es importante destacar que el modelo ya ha sido integrado en un software que interactúa con el espectrómetro de rayos láser, construido por el equipo de ingenieros del proyecto desarrollado por la Universidad Central del Ecuador.

Para el desarrollo del modelo se efectuaron 25 pruebas (con dos réplicas), las cuales se diferencian por el porcentaje (nivel) de gasolina en la muestra (ver tabla 1). Para este trabajo, si un espectro tiene un porcentaje de gasolina menor o igual al 10 % se clasifica en el grupo de baja contaminación; caso contrario se clasifica en el grupo de alta contaminación. El modelo permite considerar cualquier otro nivel de gasolina para discriminar entre espectros correspondientes a muestras con una baja o alta contaminación.

Especificación	Muestra	Nivel	Total	Especificación	Muestra	Nivel	Total
Extra	GE1	0,3	2	Extra	GE8	9,1	2
Extra	GE2	0,4	2	Alcohol	GA15	16,67	2
Extra	GE3	0,5	2	Extra	GE9	16,7	2
Alcohol	GA1	0,5	2	Extra	GE10	37,5	2
Alcohol	GA3	1,48	2	Alcohol	GA19	37,5	2
Extra	GE4	1,5	2	Extra	GE11	50	2
Extra	GE5	2,4	2	Alcohol	GA23	50	2
Alcohol	GA5	2,44	2	Extra	GE12	75	2
Extra	GE6	3,8	2	Extra	GE13	83,3	2
Alcohol	GA8	3,85	2	Alcohol	GA26	83,33	2
Extra	GE7	6,1	2	Extra	GE14	100	2
Alcohol	GA13	6,1	2	Alcohol	GA30	100	2
Alcohol	GA14	9,09	2				

Tabla 1: Muestras realizadas para el ajuste del MFLG y validación.

2. Modelo Lineal Funcional Generalizado (MFLG)

Se define a X como la variable funcional de interés, espectro generado a través de la técnica LIF, que toma valores en un espacio normado (o semi-normado) F, y se consideran como datos funcionales a los resultados de las 25 pruebas representadas como el conjunto $\{x_1, x_2, \ldots, x_n\}$ que provienen de n variables funcionales X_1, X_2, ...,X_n identicamente distribuidas como X.

Los datos funcionales se encuentran discretizados en un total de 3'648 puntos que se encuentran en el intervalo [176,39, 890,62]. Estos están represantos por el conjunto de puntos t_j.

En la Figura 1 se puede apreciar en color negro los espectros de nivel bajo y en color rojo los espectros de nivel alto. Se puede observar a partir de la figura que los espectros tienden a tener la misma forma sin embargo los espectros de nivel Alto tienen una mayor amplitud que los de nivel bajo y que esta va aumentando en relación al porcentaje de gasolina que hay en la muestra. Éste varia desde 16.67 % hasta 100 % (mayor que 10 %); por otro lado, la amplitud de los espectros de las muestras de nivel bajo van disminuyendo acorde al porcentaje de gasolina en la muestra. Un espectro es considerado con un nivel de contaminación bajo si el porcentaje de gasolina es menor que un 10 %, es decir va desde 3 % hasta 9.1 %.

Para la representación de los datos funcionales se empleó una base B-spline mediante el uso del paquete fda.usc del software estadístico R[6].

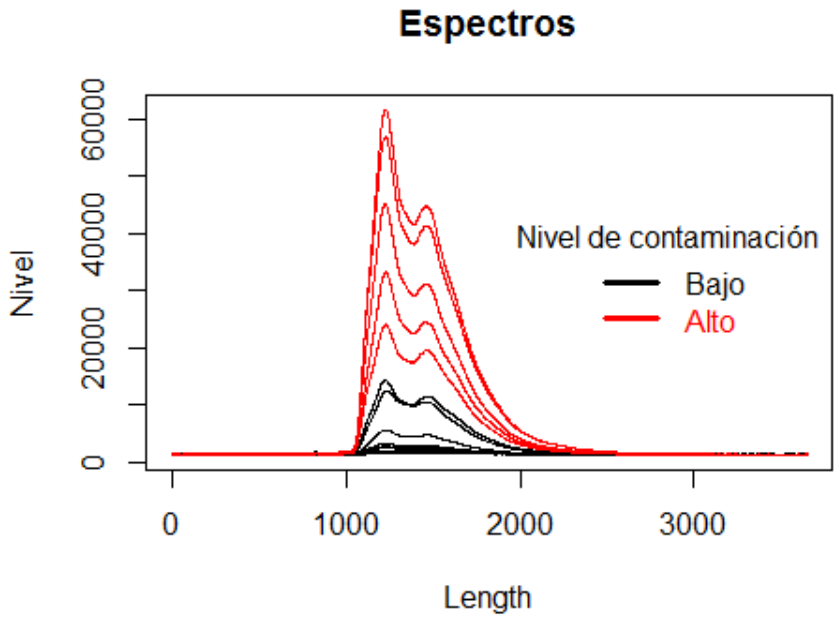


Figura 1: Espectros por nivel de contaminación.

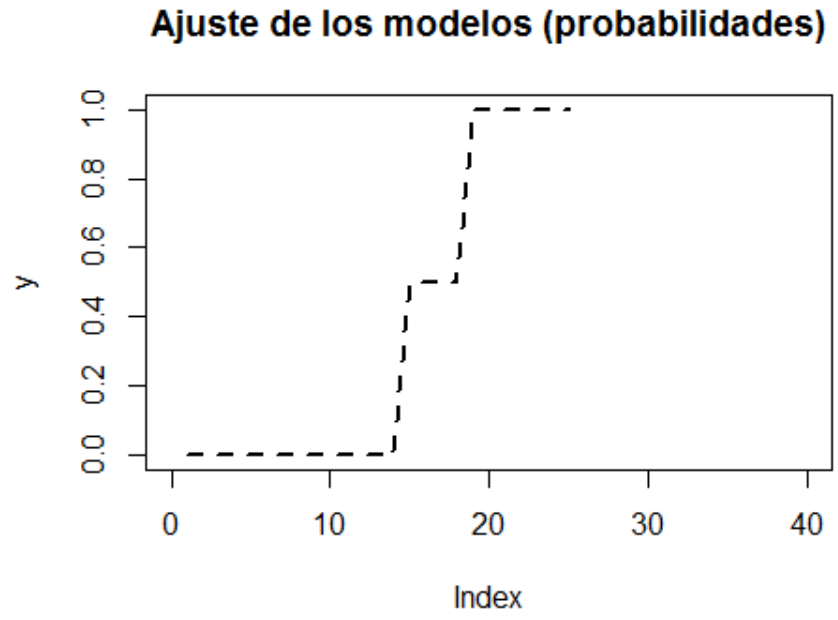


Figura 2: Ajuste del modelo: probabilidades estimadas.

Una vez representados los espectros a datos funcionales, se ajusta un Modelo Funcional Lineal Generalizado (MFLG) para estimar la probabilidad de que pertenezca a uno de los dos grupos. Para el ajuste y la implementación del modelo se utilizó la función "fregre.glm" del paquete fda.usc del software estadístico R.

El MFLG es también conocido en la literatura como Regresión Logística Funcional (FLR). El modelo explica la relación entre Y (respuesta binaria) y una covariable funcional $X(t)$ con representación en base de $X(t)$ y $\beta(t)$. π_i es la probabilidad de ocurrencia del evento $Y_i = 1$, que en este

caso corresponde a una alta contaminación, condicionada a la covariable $X_i(t)$, que se expresa de la siguiente manera:

$$y_i = \pi_i + \epsilon_i, i = 1, \ldots, n$$

$$\pi_i = P[Y = 1/x_i(t) : t \in T] = \frac{exp\{\int_T X_i(t)\beta(t)dt\}}{1 + exp\{\int_T X_i t\beta(t)dt\}} i = 1, \ldots, n$$

Donde ϵ_i son errores independientes con media cero.
Se define como covariable funcional el espectro denotado por: $X = X(t)$, y como variable escalar (binaria) de respuesta el tipo de contaminación denotada por Y (0= Baja contaminación, 1=Alta contaminación). En este caso, dado que el MFLG trabaja con una variable de respuesta binaria, este modelo proporciona una regla de clasificación para el tipo de contaminación (regla de Bayes). En la Figura 2 se presenta el ajuste del modelo GLFM.

3. CONCLUSIONES

Como se mencionó en la sección de introducción, cada muestra tiene dos réplicas, una de ellas es utilizada para la estimación del modelo y la otra para su la validación. En el caso de la muestra para la estimación se tiene que el porcentaje de espectros clasificados correctamente en cada grupo (Bajo y Alto) es del 99 %, mientras que para la muestra de validación el porcentaje de espectros que se han clasificado correctamente es de 99 %.
Se ha demostrado que la forma del espectro de fluorescencia láser está altamente relacionada con el contenido en gasolina de la muestra. Dada su naturaleza funcional, la aplicación de técnicas de clasificación supervisada FDA aportan una solución fiable para la identificación de un riesgo alto o bajo de contaminación en zonas potencialmente afectadas.
Para un trabajo futuro se desarrollará un modelo que permita estimar el porcentaje de gasolina en las muestras. Para esto se utilizará un modelo de regresión lineal funcional con variable de respuesta escalar porcentaje de gasolina y covariable funcional espectro generado a partir de la técnia LIF.

AGRADECIMIENTOS

Los autores agradecen el financiamiento otorgado por la Escuela Politécnica Nacional para la realización del proyecto PII-DM-002-2016: *"Análisis de datos funcionales en el control estadístico de calidad"*.

Referencias

[1] Celander, K., Freddricsson, B. Galle, S. y Svanberg (1988). Investigation of Laser-Induced Fluorescence with application to remote sensing of environmental parameters. Goteborg Institute of Physics Reports GIPR-149.

[2] Ferraty, F. y Vieu, P. (2006) Nonparametric Functional Data Analysis: Theory and Practice. Springer-Verlag, New York.

[3] Ferraty, F. y Romain, Y. (2010). The Oxford Handbook of Functional Data Analysis. Oxford University Press.

[4] González-Manteiga, W. y Vieu, P. (2007). Statistics for functional data. Computational Statistics and Data Analysis, 51, 4788-4792.

[5] Lakowicz, Joseph R. (2006). Principles of Fluorescence Spectroscopy (3ª ed.). USA Maryland: Springer Science+Business Media, LLC.

[6] Manuel Febrero-Bande and Manuel de la Fuente (2012),Statistical Computing in Functional Data Analysis: The R Package fda.usc, Journal of Statistical Software, 51.

[7] Oviedo de la Fuente, Manuel (2011), Utilities for Statistical Computing in Funcional Data Analysis: The R Package fda.usc trabajo de fin de Máster, Universidad Santiago de Compostela.

[8] Ramsay, J. O. y Silverman, B. W. (2005). Functional Data Analysis. Springer-Verlag, New York.

Comunicacións orais

Sesión II: Modelos mixtos e EEG

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Ajuste de modelos mistos a dados longitudinais problemáticos

Julio M. Singer[1], Carmen D.S. André[2], Francisco M.M. Rocha[3] e Talita Zerbini[4]

[1,2]Departamento de Estatística, Universidade de São Paulo, Brazil
[3]Escola Paulista de Política Economia e Negócios, Universidade Federal de São Paulo, Brazil
[4]Departamento de Patologia, Universidade de São Paulo, Brazil

RESUMO

Modelos não lineares são úteis para a análise de dados longitudinais, especialmente quando são obtidos como solução de equações diferenciais geradas a partir de características do fenômeno investigado. Vários métodos para o seu ajuste a dados obtidos de estudos experimentais estão disponíveis na literatura mas, em geral, os exemplos apresentados envolvem dados coletados de forma regular. O ajuste de tais modelos a dados problemáticos (*messy data*) pode provocar dificuldades de convergência dos algoritmos computacionais, particularmente quando os números de unidades amostrais e de observação são pequenos. Além disso, a literatura é pobre em técnicas de diagnóstico para avaliação da qualidade do ajuste. Nesse contexto, examinamos dados provenientes de um estudo piloto conduzido na Faculdade de Medicina da Universidade de São Paulo em que se propõe um novo método tomográfico para estimar o tempo desde a morte (intervalo *post mortem*). Com essa finalidade, a atenuação da densidade da concentração de hemácias no coração, conhecida por hipostase intracardíaca foi observada nos tórax de 21 corpos de pacientes hospitalizados para os quais se conhece o instante da morte. Para cada corpo, as imagens foram obtidas em diferentes instantes, não necessariamente em mesmo número, gerando um conjunto de dados problemáticos com natureza essencialmente não linear. Comparamos modelos de von Bertalanffy e modelos exponenciais linearizáveis além de modelos de regressão segmentada para representar a evolução da resposta ao longo do tempo desde a morte até a sua estabilização e usamos os resultados para estimar o intervalo *post mortem*. Discutimos as vantagens e desvantagens de cada enfoque e mostramos que os modelos adotados produzem conclusões similares, mas que o modelo exponencial linearizável é mais flexível, tira proveito de uma metodologia bem estabelecida e de algoritmos de ajuste bem eficientes além de ter várias técnicas de diagnóstico à disposição.

Palavras e frases chave: autópsia, intervalo *post mortem*, análise de resíduos.

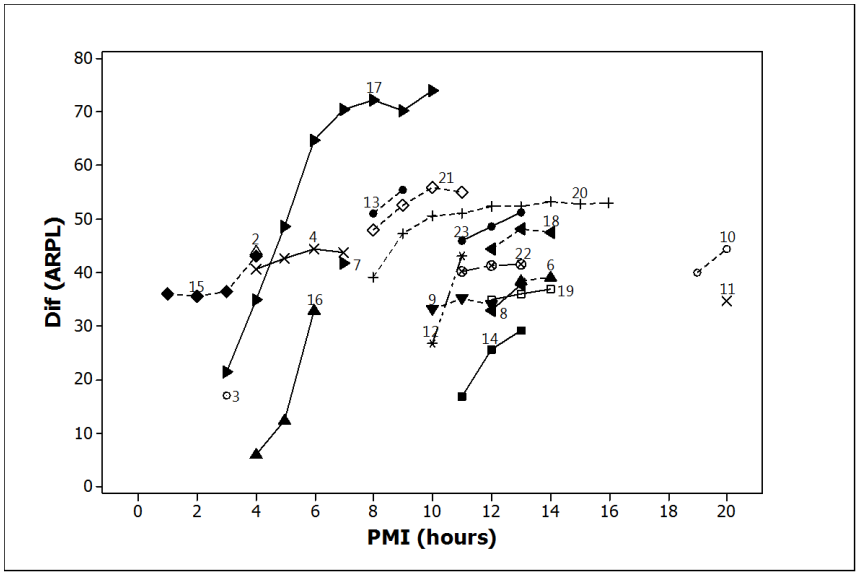
1. INTRODUÇÃO

O intervalo de tempo desde a morte, conhecido como intervalo *post mortem* (PMI), é fundamental em muitas circunstâncias, como em investigações criminais. Estimativas usuais obtidas de características físicas têm uma margem de erro que varia entre 2 e 4.5 horas dependendo do verdadeiro intervalo de tempo desde a morte. Muitas técnicas para melhorar a precisão dessas estimativas têm sido contempladas na literatura mas poucas têm sido utilizadas na prática. Nosso objetivo é discutir metodologia estatística para a análise de um estudo piloto cujo objetivo era utilizar uma série de imagens tomográficas da hipostase intracardíaca para estimar o PMI.

2. O ESTUDO

Os dados foram coletados de 21 corpos de pacientes que morreram sob cuidados médicos (o que garante o conhecimento do instante da morte) e foram encaminhados para a realização de uma autópsia. As imagens foram obtidas em intervalos de uma hora até um limite de 20 horas *post mortem* dependendo da disponibilidade do tomógrafo, gerando um conjunto de dados problemáticos (*missing data*). A diferença entre as medidas de atenuação obtidas do segmento anterior do átrio direito e do segmento posterior do átrio esquerdo [aqui rotulada por dif(ARPL)] em cada imagem foi utilizada como resposta. Um gráfico com a representação longitudinal das respostas [dif(ARPL)] de cada corpo está apresentado na Figura 1.

Figura 1: Perfis individuais das diferenças entre as medidas de atenuação obtidas do segmento anterior do átrio direito e do segmento posterior do átrio esquerdo [dif(ARPL)]



3. MODELOS

Como a hipostase tende a se estabilizar depois de um certo instante, é razoável adotar um modelo não linear com uma assíntota para representar a sua evolução temporal. Além disso, pode-se supor que o comportamento longitudinal da resposta dependa de características de cada unidade amostral (corpo). Um modelo com essa característica é o modelo de von Bertalanffy misto, expresso como

$$y_{ij} = (\gamma + g_i)\{1 - \exp[-(\delta + d_i)x_{ij}]\} + f_{ij},\ x_{ij} \geq 0 \tag{1}$$

em que y_{ij} denota a j-ésima resposta da i-ésima unidade amostral, x_{ij} denota o instante em que essa resposta foi observada, $\gamma > 0$ e $\delta > 0$ são parâmetros, os efeitos aleatórios $\mathbf{g}_i = (g_i, d_i)^\top$ seguem distribuições gaussianas independentes com vetor de médias $\mathbf{0}$ e matriz de covariâncias $\mathbf{G}$ e f_{ij} são erros aleatórios com distribuições gaussianas independentes de média zero e variância σ^2. Além disso, $\mathbf{g}_i$ e f_{ij} são independentes. O parâmetro γ representa a assíntota populacional, *i.e.*, o valor em que a resposta esperada se estabiliza e δ é um parâmetro populacional associado à velocidade com que a assíntota é atingida (também conhecido como parâmetro de crescimento intrínseco).
Outra alternativa é o modelo misto exponencial multiplicativo, expresso como

$$y_{ij} = (\alpha a_i)\exp[-(\beta + b_i)/x_{ij}]e_{ij},\ x_{ij} > 0 \tag{2}$$

com componentes definidos similarmente àqueles do modelo von Bertalanffy misto Note que (2) pode ser linearizado como

$$y^*_{ij} = \alpha^* + a^*_i + (\beta^* + b_i)x^*_{ij} + e^*_{ij},\ x^*_{ij} > 0 \tag{3}$$

em que $y_{ij}^* = \log(y_{ij})$, $\alpha^* = \log(\alpha)$, $\beta^* = -\beta$, $a_i^* = \log(a_i)$, $e_{ij}^* = \log(e_{ij})$ and $x_{ij}^* = 1/x_{ij}$. Supondo que os efeitos aleatórios $\mathbf{b}_i^* = (a_i^*, b_i)^\top$ sigam distribuições gaussianas independentes com vetor de médias $\mathbf{0}$ e matriz de covariâncias $\mathbf{G}^*$ e que e_{ij}^* sejam erros aleatórios com distribuições gaussianas independentes de média 0 e variância τ^2, (3) pode ser classificado como uma modelo linear misto usual.
Alternativamente, a evolução de dif(ARPL) pode ser representada por um modelo de regressão segmentada misto como

$$y_{ij} = I\{x_{ij} \leq x_0\}[(\rho + r_i)x_{ij}] + (1 - I\{x_{ij} \leq x_0\})[(\rho + r_i)x_0] + h_{ij},\ x_{ij} \geq 0 \tag{4}$$

em que $I\{x_{ij} \leq x_0\} = 1$ se $x_{ij} \leq x_0$ ou $I\{x_{ij} \leq x_0\} = 0$ em caso contrário, ρ é um parâmetro com interpretação semelhante àquela de δ ou β, x_0 corresponde ao instante em que dif(ARPL) se estabiliza (atinge a assíntota) e h_{ij} são erros aleatórios com distribuições gaussianas independentes de média 0 e variância ν^2.

4. RESULTADOS

Inicialmente ajustamos variações dos três modelos (incluindo ou não efeitos aleatórios associados a cada um dos parâmetros) por meio da programoteca `nlme` [Pinheiro et al. (2014)] do pacote computacional `R`.
Essencialmente, os três modelos produziram estimativas da assíntota comparáveis. Com base nas estatísticas AIC e BIC (para a mesma classe de modelos) e porque a análise descritiva (ver Figura 1) sugere assíntotas e velocidades de crescimento diferentes para unidades amostrais diferentes, o modelo (2) escolhido para análises subsequentes inclui efeitos aleatórios para ambos os parâmetros. Uma análise de resíduos segundo a linha preconizada por Singer, Rocha e Nobre (2016) sugeriu um modelo parcialmente heterocedástico com variâncias intraunidades amostrais diferentes para três delas. A análise de resíduos desse modelo modificado não apresenta evidências contrárias ao seu ajuste, que apesar de sub-ótimo, é satisfatório. Para o modelo adotado, as correspondentes estimativas ($\pm$ erros padrões) para a assíntota populacional (α) e velocidade de crescimento (β) são, respectivamente, $\widehat{\alpha} = 52{,}7 \pm 5{,}6$ e $\widehat{\beta} = 1{,}99 \pm 0{,}68$.
Como o objetivo do método proposto é estimar o PMI para valores fixados de dif(ARPL) de alguma unidade amostral específica, digamos y_{i0}, podemos inverter (3), obtendo

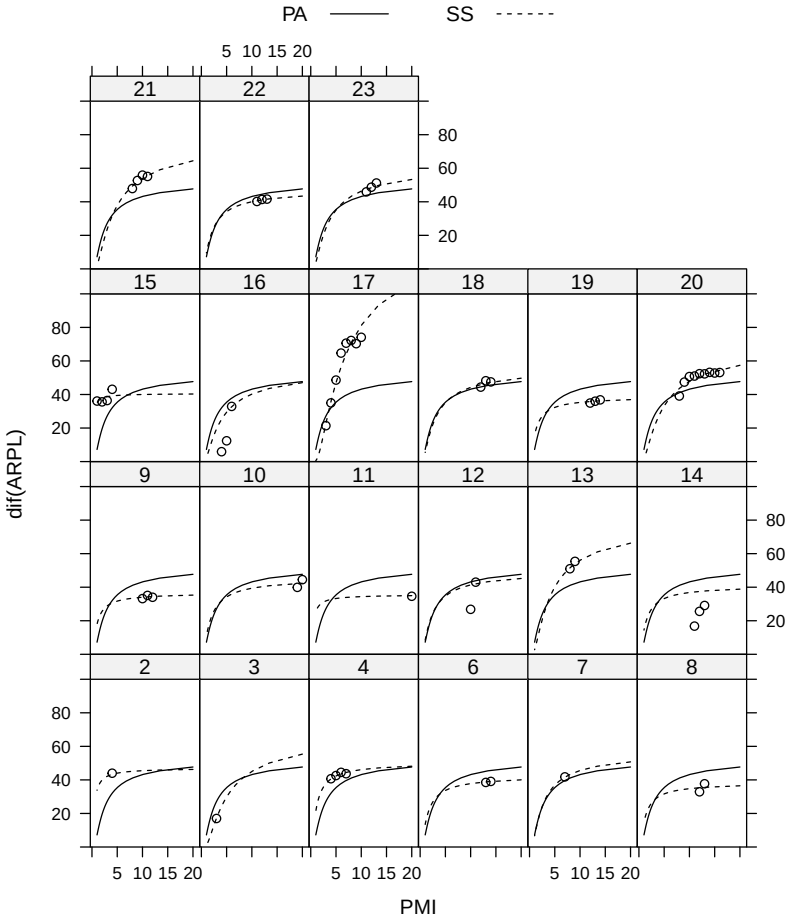
$$\widehat{PMI}_i(y_{i0}) = \frac{\widehat{\beta}^* + \widehat{b}_i}{\log(y_{i0}) - (\widehat{\alpha}^* + \widehat{a}_i^*)} \tag{5}$$

em que $\widehat{\alpha}^*$ e $\widehat{\beta}^*$ são as estimativas de máxima verossimilhança de α e β e $\widehat{a}_i^*$ e $\widehat{b}_i$ são os preditores lineares ótimos (BLUP) de a_i^* e b_i no modelo (3). Os erros padrões correspondentes são obtidos por meio do Método Delta. Curvas correspondentes às 21 unidades amostrais estão dispostas na Figura 2. Códigos escrito em R para o ajuste dos modelos (1), (3) e (4), para obtenção dos valores de PMI preditos e erros padrões correspondentes bem como para a construção dos perfis individuais estão disponíveis em http://www.ime.usp.br/∼jmsinger/hypostasis.zip.

5. CONCLUSÕES

Com base no exemplo, mostramos que os resultados obtidos a partir dos três modelos são bastante similares relativamente às estimativas dos parâmetros de interesse. O modelo de regressão segmentada misto é atraente não só pela sua simplicidade mas também porque pode ser ajustado por meio de metodologia amplamente disseminada e dispõe de técnicas de diagnóstico bem estabelecidas. A desvantagem é que a assíntota precisa ser determinada por um processo de tentativa e erro. Embora sejam interessantes por razões de interpretação [ver Lindsey (2004)], os modelos não lineares mistos como (1) estão associados a problemas de convergência dos algoritmos computacionais necessários para o seu ajuste. Modelos mistos linearizáveis, por outro lado, podem ser ajustados por meio de algoritmos bem estáveis e não requerem a especificação de valores iniciais. Além disso, curvas de respostas individuais podem ser facilmente estimadas.
O método proposto para a estimação do PMI apresenta resultados melhores que os métodos usuais [Kaliszan et al. (2009)] nas primeiras 12 horas após a morte desde que pelo menos 4 medidas tenham

Figura 2: Perfis individuais de dif(ARPL) estimados pelo modelo exponencial; PA = populacional, SS = específico para a unidade amostral



sido realizadas. Apesar de reconhecermos as limitações do estudo motivador, que foi conduzido sob condições não ideais, lembramos que todos os métodos atualmente empregados têm limitações, de forma que o uso concomitante de dois ou mais deles pode gerar resultados mais confiáveis.

AGRADECIMENTOS

Este projeto recebeu apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, processo 3304126/2015-2) e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo 2013/21728-2), Brasil.

Referencias

[1] Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley, New York.

[2] Kaliszan, M., Hauser, R. and Kernbach-Wighton, G. (2009). Estimation of the time of death based on the assessment of post mortem processes with emphasis on body cooling. *Legal Medicine (Tokyo)*, 11, 111–117.

[3] Lindsey, J.K. (2004). *Statistical analysis of stochastic processes in time*. Cambridge University Press, Cambridge.

[4] Pinheiro, J., Bates, D.M., DebRoy, S., Sarkar, D. and R Core Team (2014). `nlme`: *Linear and Nonlinear Mixed Effects Models.* `R` package version 3.1-117. url = http://CRAN.R-project.org/package=nlme

[5] Singer, J.M., Rocha, F.M.M. and Nobre, J.S. (2016). Graphical tools for detecting departures from linear mixed models and some remedial measures. (Submetido para publicação).

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

MODELO MULTINÍVEL NA AVALIAÇÃO DO NÍVEL ÓSSEO MARGINAL PERI-IMPLANTAR EM FUNÇÃO DA HIGIENE BUCAL

Omar Pereira[1], Mariana Tormena[2], Flávia Matarazzo[2],
Maurício Guimarães Araújo[2], Isolde Previdelli[1]

[1]Programa de Pós-Graduação em Bioestatística
[2]Programa de Pós-Graduação em Odontologia

RESUMO

O objetivo deste estudo foi avaliar a influência da falta de acesso adequado à higiene bucal no nível ósseo marginal peri-implantar por meio de um modelo multinível. A amostra com 41 indivíduos tiveram seus implantes classificados em dois grupos de acordo com a higiene bucal: i) Grupo Controle: indivíduos que apresentaram todas as superfícies dos implantes livres de placa; ii) Grupo Sem Acesso: indivíduos que apresentavam implantes com pelo menos uma face com biofilme dental. O nível ósseo marginal foi obtido através de radiografias periapicais padronizadas e avaliadas por examinador previamente calibrado em 5, 7 e 9 anos após a realização do implante. O modelo evidenciou que, em média, um implante pertencente ao grupo Sem Acesso 9 anos após a instalação do implante, pode ter sua perda óssea aumentada quando comparado com o grupo Controle. Conclui-se que, a falta de acesso adequado à higiene bucal de prótese sobre-implantes acarretou em acúmulo de placa e redução significativa do nível ósseo marginal peri-implantar.

Palavras e frases chave: Dados longitudinais, modelo hierárquico, acesso à higiene bucal, nível ósseo marginal, implantes dentários, placa dental.

1. INTRODUÇÃO

O uso dos implantes osseointegráveis tem apresentado altos índices de sucesso, porém os mesmos não estão imunes a complicações técnicas e biológicas. Dentre as complicações biológicas pode-se incluir as doenças peri-implantares que são classificadas em: i) mucosite peri-implantar, caracterizada pela presença de inflamação nos tecidos ao redor do implante mas sem presença de perda óssea; ii) peri-implantite, caracterizada pela presença de perda óssea progressiva somada aos sinais clínicos de inflamação. A maior parte dos estudos sobre a etiologia das doenças peri-implantares concordam que tal condição está relacionada com a resposta inflamatória frente ao biofilme bacteriano (Zitzmann et al., 2001). Entretanto, avaliando-se o nível de higiene bucal, a maioria das pessoas não são capazes de realizá-la satisfatoriamente, com total remoção da placa bacteriana (Cumming and Löe, 1973). Em um estudo transversal, Serino and Ström (2009) observaram que cerca de 74 % dos implantes não tinham acesso adequado à higienização. Até o presente momento e nos limites da nossa busca bibliográfica, não encontrou-se estudos longitudinais que avaliassem o efeito do acesso inadequado à higiene bucal no nível ósseo marginal peri-implantar. Este estudo testou a hipótese que o acesso inadequado a higiene bucal na região peri-implantar pode estar associado com a perda óssea marginal.

2. PROCEDIMENTOS EXPERIMENTAIS

Este estudo foi composto por 41 indivíduos que apresentavam reabilitações parciais com implantes em função há 5 anos. Todos os indivíduos apresentavam implantes com profundidade de sondagem $\leq 5mm$ associados ou não ao sangramento. Os indivíduos foram considerados como

tendo ou não acesso adequado à higienização após higiene oral supervisionada. Os implantes que apresentavam pelo menos uma face com biofilme dental remanescente foram considerados como sem acesso adequado à higiene bucal (Grupo Sem Acesso) e aqueles que apresentaram todas as superfícies livres de placa, foram considerados como tendo acesso adequado à higiene bucal (Grupo Controle). A variável resposta, nível ósseo marginal peri-implantar, foi obtida por meio de radiografias periapicais padronizadas realizadas em três momentos diferentes, cinco, sete e nove anos após o implante ser realizado.

3. MODELO MULTINÍVEL

Os dados de nível ósseo marginal peri-implantar observados no t-ésimo momento, com $t = 1, \ldots, 3$ correspondendo a 5, 7 e 9 anos após o implante dental, no i-ésimo dente pertencente ao k-ésimo grupo, $k = 1, 2$, aninhado dentro do j-ésimo paciente e denotado por MBL_{ijkt}, foram ajustados por meio do modelo de efeitos mistos

$$MBL_{ijkt} = (\beta_0 + b_{0j} + b_{0i|j}) + \gamma_k + \tau_t + (\alpha\tau)_{kt} + \epsilon_{ijkt} \quad (1)$$

em que β_0 é uma média global, γ é o efeito dos grupos, τ é o efeito do tempo, b_{0j} é o efeito aleatório associado ao $j$ésimo paciente, $\mathbf{b}_{0j} \sim N(\mathbf{0}, \mathbf{\Phi}_2)$, $b_{0i|j}$ é o efeito aleatório associado ao dente i aninhado dentro de um paciente, $(\mathbf{b}_{0i|j} \sim N(0, \mathbf{\Phi}_1))$, e ϵ_{ijkt} é o resíduo $(\boldsymbol{\epsilon}_{ij} \sim N(0, \mathbf{R}_{ij}))$. Neste modelo, os grupos e o tempo foram tratados como fatores, com dois e três níveis respectivamente. O grupo com acesso e o tempo inicial (5 anos) foram considerados os grupos controles.

4. RESULTADOS E DISCUSSÃO

No início do estudo a média e erro padrão dos dados de perda óssea foram 1.88(0.05) e 2.02(0.07) mm, aos 7 anos de 1.98(0.05) e 2.16(0.07) e aos 9 anos de 2.13(0.06) e 2.41(0.07) para os grupos controle e sem acesso respectivamente. A Figura 1 apresenta a distribuição dos dados de perda óssea nos dois grupos experimentais em cada um dos tempos estudados. É possível observar que, em todos os momentos a variabilidade dos dados de perda óssea foi maior no grupo Sem Acesso, assim como indicam os erros padrão. Observa-se que o grupo Controle apresentou alguns pontos candidatos a outliers. Ajustou-se um modelo de regressão, conforme equação (1), com efeitos

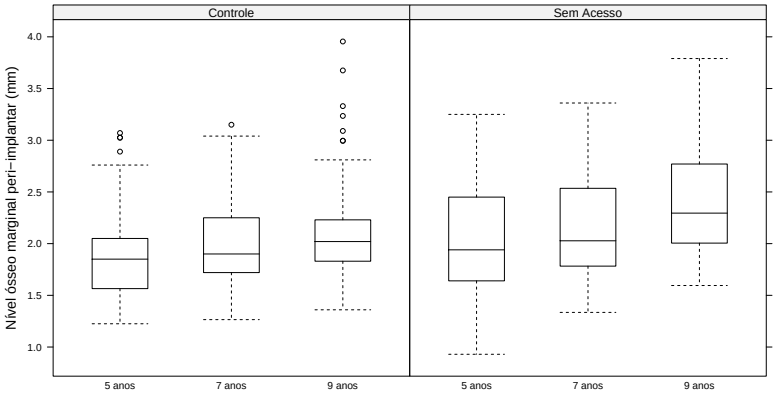


Figura 1: Boxplot para os dados do nível ósseo marginal peri-implantar em função dos momentos de observação e dos grupos.

aleatórios nos pacientes e nos implantes aninhados dentro dos pacientes. Testou-se a possibilidade de omitir algum efeito fixo, e o modelo completo com estrutura de covariância auto-regressiva de primeira ordem para os resíduos, $\mathbf{R}_{ij}$, foi o escolhido.

A Figura 2 apresenta três gráficos para o diagnóstico do modelo ajustado. Os dois primeiros gráficos apresentam os resíduos padronizados e os valores observados, ambos em função dos valores ajustados pelo modelo. Estes gráficos não indicam grandes desvios do modelo, exceto por alguns pontos estimados para seis valores do nível ósseo marginal peri-implantar. O outro gráfico, quantil-quantil, mostra que os resíduos atendem a suposição de normalidade.

A Tabela 1 mostra os valores estimados com seus intervalos de confiança, os erros padrão (SE) e os p-valores para os efeitos fixos do modelo ajustado. Observa-se que a interação do grupo Sem

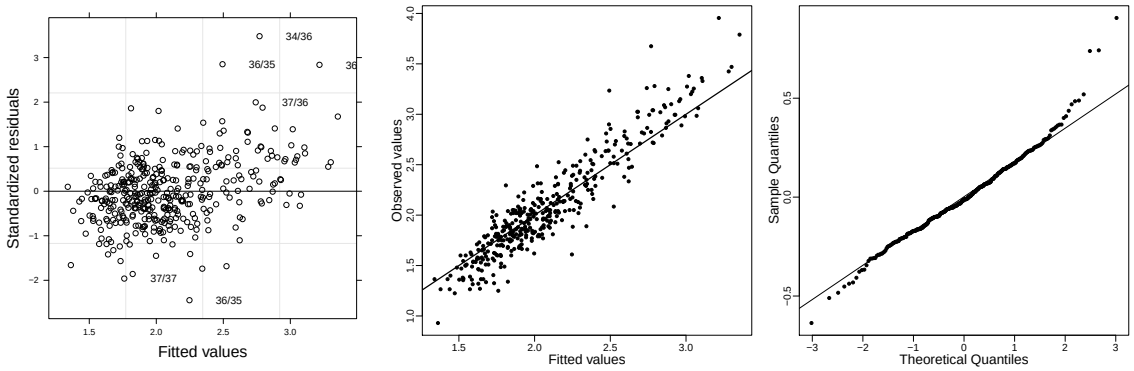


Figura 2: Gráficos de diagnósticos do modelo multinível para o nível ósseo marginal peri-implantar. O primeiro e segundo gráficos mostram os resíduos padronizados e os valores observados em função dos valores ajustados. O último é o gráfico quantil-quantil.

Acesso e o momento 9 anos após a realização do implante, $(\gamma\tau)_{23}$, foi significativa ($p < 0{,}01$). Isto indica que as diferenças nos valores da perda óssea marginal entre os dois grupos estudados não é a mesma em todos os tempos. Mais especificamente, pode-se dizer que existe uma grande evidência de que, em média, um implante pertencente ao grupo Sem Acesso 9 anos após colocar o implante, pode ter sua perda óssea aumentada, quando comparado com o grupo Controle. Este fato também está evidente na Figura 1. Os boxplots dos grupos Controle e Sem Acesso são muito semelhantes quando compara-se a perda óssea marginal entre os momentos 5 e 7 anos. No entanto, estes boxplots para os grupos são diferentes aos 9 anos após a aplicação do implante. Neste momento, o grupo Sem Acesso apresenta valores de mediana e quantis maiores que o grupo Controle.

Parâmetro	Estimativa	LI	LS	SE	p-value
β_0	1.890	1.760	2.020	0.066	0.0000
γ_2	0.067	-0.120	0.255	0.094	0.4768
τ_2	0.100	0.037	0.163	0.032	0.0018
τ_3	0.247	0.172	0.322	0.038	0.0000
$(\gamma\tau)_{22}$	0.049	-0.047	0.144	0.048	0.3145
$(\gamma\tau)_{23}$	0.152	0.037	0.267	0.058	0.0096

Tabela 1: Estimativas, limite inferior e superior (LI and LS), erro padrão (SE) e p-valores para o modelo multinível.

5. CONSIDERAÇÕES FINAIS

No presente estudo, a falta de acesso para uma higienização adequada foi observada em cerca de 50 % dos pacientes e 42 % dos implantes avaliados. Estes dados corroboram com os achados de Serino and Ström (2009) que avaliaram 23 pacientes que apresentavam peri-implantite e observaram que 74 % dos implantes não tinham acesso adequado para higienização. Logo, sugere-se que a falta de acesso para uma correta higienização pode ser um achado frequente nas reabilitações com implantes osseointegráveis. Observou-se que a perda óssea marginal foi maior para o grupo Sem Acesso em comparação com o grupo Controle 9 anos após a realização do implante. Kourtis et al. (2004) realizaram um estudo que avaliou o nível ósseo ao redor dos implantes e o grau de higiene bucal de uma amostra expressiva com 12 anos de acompanhamento. Os autores observaram uma relação clara entre essas duas variáveis (presença de placa e perda óssea marginal peri-implantar), assim como relatado por Hultin et al. (2002) e Renvert et al. (2007). Outros estudos clínicos não encontraram a mesma associação entre placa bacteriana e nível ósseo marginal (Karoussis et al., 2004; Koldsland et al., 2011). A razão para esta contrariedade ainda é obscura, mas sugere-se que possa estar relacionada com a apresentação dos valores de perda óssea marginal como média por paciente, e tal abordagem pode diluir ou mascarar valores clínicos significativos de perda óssea marginal. Este não é o caso dos modelos multiníveis em que um mesmo paciente pode ter implantes em diferentes grupos. Mesmo assim, todos eles terão sua perda óssea modelada e correlacionada ao longo do tempo (West et al., 2014). Considerando que todos os pacientes

apresentavam uma condição inicial semelhante, como produto dos critérios de exclusão e inclusão instituídos, os resultados mostraram uma grande evidência que, em média, um implante que não possua acesso adequado para remoção completa da placa bacteriana através dos procedimentos básicos diários de higiene bucal, com o passar do tempo pode ter uma perda óssea aumentada, quando comparado a um sítio peri-implantar com acesso adequado a higienização. A ocorrência dessa perda óssea progressiva pode culminar com a instalação de uma peri-implantite devido a exposição das roscas do implante e maior acúmulo de microorganismos, facilitando a quebra da homeostase entre hospedeiro e agente agressor e progressão da doença peri-implantar (Serino and Turri, 2011).

Referencias

Cumming, B. R. and Löe, H. (1973). Consistency of plaque distribution in individuals without special home care instruction. *Journal of periodontal research*, 8(1):94–100.

Hultin, M., Gustafsson, A., Hallström, H., Johansson, L.-Å., Ekfeldt, A., and Klinge, B. (2002). Microbiological findings and host response in patients with peri-implantitis. *Clinical Oral Implants Research*, 13(4):349–358.

Karoussis, I. K., Müller, S., Salvi, G. E., Heitz-mayfield, L. J., Brägger, U., and Lang, N. P. (2004). Association between periodontal and peri-implant conditions: a 10-year prospective study. *Clinical Oral Implants Research*, 15(1):1–7.

Koldsland, O. C., Scheie, A. A., and Aass, A. M. (2011). The association between selected risk indicators and severity of peri-implantitis using mixed model analyses. *Journal of clinical periodontology*, 38(3):285–292.

Kourtis, S. G., Sotiriadou, S., Voliotis, S., and Challas, A. (2004). Private practice results of dental implants. part i: survival and evaluation of risk factors—part ii: surgical and prosthetic complications. *Implant dentistry*, 13(4):373–385.

Renvert, S., Roos-Jansåker, A.-M., Lindahl, C., Renvert, H., and Rutger Persson, G. (2007). Infection at titanium implants with or without a clinical diagnosis of inflammation. *Clinical oral implants research*, 18(4):509–516.

Serino, G. and Ström, C. (2009). Peri-implantitis in partially edentulous patients: association with inadequate plaque control. *Clinical Oral Implants Research*, 20(2):169–174.

Serino, G. and Turri, A. (2011). Outcome of surgical treatment of peri-implantitis: results from a 2-year prospective clinical study in humans. *Clinical oral implants research*, 22(11):1214–1220.

West, B. T., Welch, K. B., and Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. CRC Press.

Zitzmann, N., Berglundh, T., Marinello, C., and Lindhe, J. (2001). Experimental peri-implant mucositis in man. *Journal of Clinical Periodontology*, 28(6):517–523.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de junho a 2 de Julho de 2016

A SUPERIORIDADE DO MODELO MULTINÍVEL PARA AVALIAR DADOS ODONTOLÓGICOS

Tiago Peres da Silva Suguiura[1], Omar Cléo Neves Pereira[2],
Waenya Fernandez de Carvalho[3], Isolde Previdelli[4]

[1]Programa de Pós Graduação em Bioestatística - Universidade Estadual de Maringá
[2]Pós-doc em Bioestatística - Universidade Estadual de Maringá
[3]Programa de Pós Graduação em Odontologia - Universidade Estadual de Maringá
[4]Professora Associada do Departamento de Estatística - Universidade Estadual de Maringá

RESUMO

A utilização de modelos mistos multiníveis/hierárquicos na análise de dados provenientes da odontologia têm crescido recentemente devido às características do experimento, por exemplo, os 'lados' dos dentes estarem aninhados em um dente, que por sua vez está aninhado em cada paciente e dividem características comuns de observações correlacionadas. Por esse motivo, análises assumindo independência das observações é inapropriado. Qualquer método de análise que tente agrupar os dados em nível de paciente resulta em perda de informação valiosa e pode não refletir a associação específica. Neste trabalho é feita uma comparação entre um modelo linear misto e um modelo linear misto multinível utilizando, como estimação dos parâmetros, as equações dos modelos mistos de Henderson. Como aplicação, é avaliado a utilização do Cloridrato de Nafazolina (Colírio Legrand) como agente de retração gengival em 24 pacientes que receberam, cada um, três tratamentos em dentes diferentes. Os resultados indicaram que o modelo multinível apresentou melhores ajustes na análise de resíduos, bem como corroborou com a hipótese de pesquisa biológica.
Palavras chave: Cloridrato de Nafazolina, equações de Henderson, modelos hierárquicos, modelos lineares mistos, retração gengival.

1. INTRODUÇÃO

A maior parte da análise estatística tradicional assume observações independentes. Tal suposição de independência indica que as respostas sobre o sujeito não estão correlacionadas entre si. Quando se utilizam amostras aleatórias de uma grande população, é esperado que tal independência ocorra. Porém, quando se trata de grupos com uma natural hierarquia (escolas, hospitais, empresas), respostas de pessoas do mesmo grupo tendem a exibir certo grau de relação. Ou então filhos de uma mesma família tendem a serem baixos, talvez por seus pais serem pequenos ou devido à ações socioeconômicas. Muitos experimentos desenhados, tais como testes clínicos realizados em centros aleatoriamente escolhidos ou grupos de indivíduos, também criam dados hierárquicos.

Segundo Hancock (2010), as vantagens de modelos lineares hierárquicos não são somente estatísticos. Análise multinível nos permite explorar a informação contida em amostras dos grupos para explicar a variabilidade de uma variável de interesse tanto 'entre os grupos' como 'dentro do grupo'. Tais modelos nos permitem usar preditores em ambos os níveis para explicar a variância na variável dependente.

É comum em dados clínicos de doenças periodontais (conjunto de condições inflamatórias e de caráter crônico e origem bacteriana que afetam a gengiva) entre outras, os 'lados' dos dentes estarem aninhados em um dente, que por sua vez estão aninhados em cada paciente e dividem características comuns de observações correlacionadas. Devido a sua estrutura hierárquica, análises assumindo independência das observações são inapropriadas. Qualquer método de análise que tente agrupar os dados em nível de paciente resulta em perda de informação valiosa e pode não refletir a associação específica. O uso de efeitos aleatórios em modelagem

multinível é uma maneira comum e conveniente de modelar tal estrutura de grupos. É, portanto, além de apropriado, necessário considerar 'lado do dente' e 'dente dentro do paciente' como fatores aleatórios (Mdala et al., 2012).

E para isso, uma parceria com pesquisadores do Programa de Pós Graduação em Odontologia da Universidade Estadual de Maringá tem sido desenvolvida. No estudo em questão foi realizado um ensaio clínico com o objetivo de avaliar o efeito do uso do cloridrato de nafazolina em comparação com o cloreto de alumínio (elemento de referência) no afastamento gengival vertical.

Como enfoque estatístico, tem-se interesse em detalhar as características do modelo misto multinível, isto é, seu modelo marginal, e suas matrizes de covariâncias. Portanto, para o caso em estudo, utilizou-se as equações dos modelos mistos de Henderson, a qual se baseia no método dos mínimos quadrados, para encontrar o BLUE para β e o BLUP para $\mathbf{b}$.

2. METODOLOGIA

Henderson (1950) nos diz que uma maneira de encontrarmos as estimativas dos parâmetros, tanto dos efeitos fixos quanto dos efeitos aleatórios, é utilizando as Equações do Modelo Misto (EMM), dada na forma matricial por:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}) \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}. \tag{1}$$

Singer e Andrade (1986) apresentam dados provenientes de estudos com medidas repetidas por meio de modelos mistos da forma:

$$\mathrm{y}_i = g(\mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\beta}, \mathbf{b}_i) + \boldsymbol{\varepsilon}_i \quad i = 1, \cdots, n \tag{2}$$

onde $\mathrm{y}_i = (y_{i1}, \cdots, y_{ip_i})^T$ com dimensão $(p_i \times 1)$ é o perfil de respostas da i-ésima unidade experimental, β um vetor com dimensão $(t \times 1)$ de parâmetros (efeitos fixos) desconhecidos, $\mathbf{X}_i = (x_{i1}, \cdots x_{im})$ uma matriz de especificação dos efeitos fixos com dimensão $(p_i \times t)$, conhecida e de posto completo, $\mathbf{b}_i$ um vetor com dimensão $(q \times 1)$ de parâmetros para os efeitos aleatórios, $\mathbf{Z}_i$ uma matriz de especificação dos efeitos aleatórios com dimensão $(p_i \times q)$ conhecida e de posto completo e $\boldsymbol{\varepsilon}_i$ é um vetor de erros aleatórios com dimensão $(p_i \times 1)$. Em muitos casos é razoável assumir que $\mathbf{b}_i \sim \mathbf{N}_q(\mathbf{0}, \mathbf{G})$ e $\boldsymbol{\varepsilon}_i \sim \mathbf{N}_{p_i}(\mathbf{0}, \mathbf{R}_i)$, em que $\mathbf{G}$, com dimensão $(q \times q)$ e $\mathbf{R}_i$ com dimensão $(p_i \times p_i)$ são matrizes simétricas definidas positivas e, além disso, que $\mathbf{b}_i$ e $\boldsymbol{\varepsilon}_i$ são variáveis aleatórias independentes. Quando $\mathbf{R}_i = \sigma^2\mathbf{I}_{p_i}$, o modelo é chamado de **modelo de independência condicional homocedástico**.

Para utilizar as equações de Henderson, são necessárias as matrizes $\mathbf{X}_i$ das especificações dos efeitos fixos, o vetor $\mathbf{Y}$ das observações, a matriz $\mathbf{Z}_i$ das especificações dos efeitos aleatórios, a matriz $\mathbf{G}$ da variância dos efeitos aleatórios e a matriz $\mathbf{R}$ de covariância dos resíduos. As matrizes $\mathbf{Z}_i$ e $\mathbf{G}$ serão geradas de acordo com a disposição dos dados.

O modelo linear misto será dado por

$$Afastamento_{ij} = \underbrace{\beta_0 + \beta_j \times TRAT_j}_{\text{Efeitos Fixos}} + \overbrace{b_{0i} + \varepsilon_{ij}}^{\text{Efeitos Aleatórios}}.$$

Já o modelo linear misto multinível será

$$Afastamento_{ij} = \underbrace{\beta_0 + \beta_j \times TRAT_j}_{\text{Efeitos Fixos}} + \overbrace{b_{0i} + \mathbf{b}_{0i|j} + \varepsilon_{ij}}^{\text{Efeitos Aleatórios}},$$

em que β_0 representa o efeito fixo associado à média, β_j representa o efeito fixo associado ao tratamento em cada dente $(j = 1, 2, 3)$, b_{0i} representa o efeito aleatório do paciente $(i = 1, \ldots, 24)$, $b_{0i|j}$ representa o efeito

aleatório do dente dentro de cada paciente e $\varepsilon_{i,j}$ representa os erros.

O efeito multinível adotado no estudo representa a variabilidade dos dentes dentro de cada paciente. Para tal, foi acrescentado o fator $\mathbf{b}_{0i|j}$ no modelo linear misto. O acréscimo do efeito multinível não afeta a estimativa dos efeitos fixos, somente a parte dos efeitos aleatórios.

3. Resultados

Pode-se observar a análise dos resíduos dos modelos misto e multinível na Figura1 a seguir.

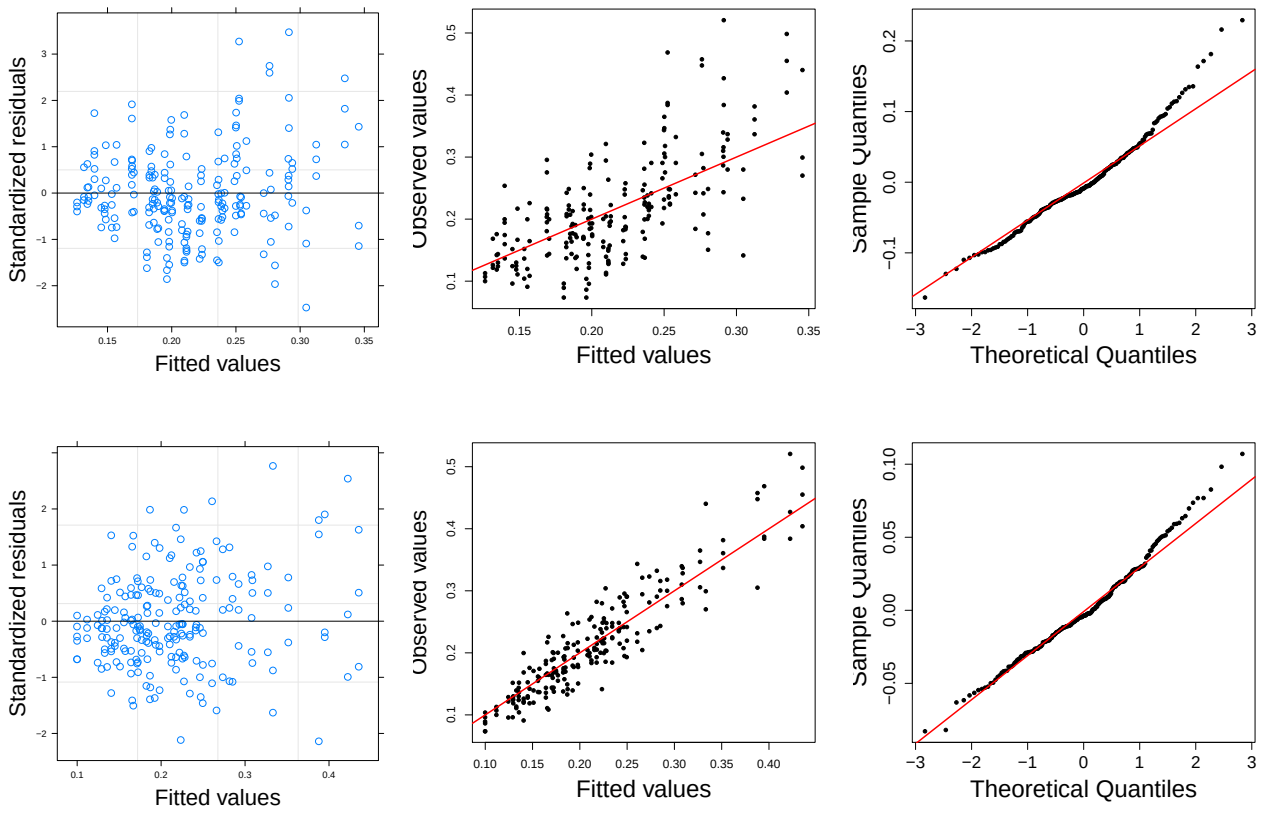


Figura 1: Gráficos de diagnósticos dos modelos misto e multinível para o afastamento gengival. O primeiro/segundo e quarto/quinto gráficos mostram os resíduos padronizados e os valores observados em função dos valores ajustados para os modelos misto e multinível respectivamente. O terceiro e sexto são os gráficos quantil-quantil para os modelos misto e multinível respectivos.

Em todas as comparações, o gráfico de resíduos para o modelo com efeito multinível se apresenta melhor ajustado com os dados do estudo.

Além da comparação com os gráficos, podemos utilizar os critérios de informação AIC, BIC e o teste da razão de verossimilhança. Conforme o software `R`, temos o seguinte resultado na comparação entre os dois modelos. O `fit1` é o modelo com hierarquia e `fit1.2` é o modelo sem hierarquia.

```
> anova(fit1,fit1.2)
Model         df    AIC       BIC   logLik   Test  L.Ratio p-value
fit1       1  6 -588.7879 -568.6202 300.3940
fit1.2     2  5 -488.1282 -471.3217 249.0641 1 vs 2 102.6598  <.0001
```

De acordo com tais resultados e mais a análise dos gráficos dos resíduos, há evidências de que o modelo que mais se ajustou aos dados foi o modelo multinível.

Utilizando o placebo como *baseline*, o tratamento com o Colírio não apresentou um p-valor significativo. A hipótese nula de que eles não são diferentes não foi rejeitada. Já o tratamento com o Hemostop, obteve um p-valor significativo na comparação com o placebo, indicando que há diferença entre os tratamentos. Portanto, de acordo com os dados, há evidências de que o Cloridrato de Nafazolina (Colírio Legrand), quando utilizado como agente de retração gengival não é capaz de aumentar o afastamento gengival em comparação com a substancia padrão.

4. CONSIDERAÇÕES FINAIS

A escolha de um modelo estatístico que se adeque melhor ao conjunto de dados a ser estudado representa uma das principais etapas da modelagem estatística. A cada dia, novas teorias e modelos surgem para tanto facilitar quanto aprimorar a análise dos dados. Teorias mais completas ajudam a evitar transtornos na compreensão de tais informações e interpretação dos dados. Para que tais teorias sejam aproveitadas ao máximo, devem ser comprovadas que são úteis no processo estatístico e que, de fato, façam com que as análises sejam mais informativas para permitir uma superioridade nas conclusões.

Dados provenientes da odontologia geralmente requisitam de uma análise que envolvam dados agrupados em mais de um nível. Como sempre existe a variação para cada paciente, também existe a variação nas unidades experimentais que, normalmente, são os dentes. A utilização de modelos multiníveis para dados com hierarquia se mostra superior a um modelo que não considera tal estrutura. Além de um melhor ajuste aos dados (comprovados pela análise de resíduos), conseguimos caracterizar o modelo com mais informações e parcimoniosidade

Referências

Aitkin, M., Anderson, D., and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, pages 419–461.

Bennett, N., Jordan, J., Long, G., and Wade, B. (1976). Teaching styles and pupil progress.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Frost, J. (2012). Why you need to check your residual plots for regression analysis: Or, to err is human, to err randomly is statistically divine.

Goldstein, H. (2011). *Multilevel statistical models*, volume 922. John Wiley & Sons.

Hancock, Gregory R Mueller, R. O. (2010). *The reviewer's guide to quantitative methods in the social sciences*. Routledge.

Henderson, C. R. (1950). Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210.

Mdala, I., Haffajee, A. D., Socransky, S. S., De Blasio, B. F., Thoresen, M., Olsen, I., and Goodson, J. M. (2012). Multilevel analysis of clinical parameters in chronic periodontitis after root planing/scaling, surgery, and systemic and local antibiotics: 2-year results. *Journal of oral microbiology*, 4.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Singer, Julio M. , Andrade, D. d. (1986). Análise de dados longitudinais. *Simpósio Nacional de Probabilidade e Estatística*, 7.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Equações de Estimação Generalizadas para modelagem de dados longitudinais comportamentais farmacológicos.

Matheus Henrique Dal Molin Ribeiro[1], Omar Pereira[2], Humberto Milani[3] e Isolde Previdelli[2]

[1]Universidade Tecnológica Federal do Paraná (UTFPR). Departamento de Matemática (DAMAT).
[2]Universidade Estadual de Maringá (UEM). Departamento de Estatística (DES).
[3]Universidade Estadual de Maringá (UEM). Departamento de Farmácia (DEF).

RESUMO

Modelos animais associados a estudos longitudinais têm sido utilizados constantemente na área da farmacologia, com objetivo de entender o comportamento cognitivo de animais submetidos, por meio de diferentes métodos, a isquemia cerebral na presença de diferentes tratamentos. Dessa forma, o objetivo deste estudo foi aplicar a metodologia das Equações de Estimação Generalizadas (EEG) para modelar dados de um experimento cujo escopo é avaliar o efeito de diferentes tratamentos com atorvastatina em animais cuja isquemia cerebral foi induzida por meio de hipoperfusão cerebral crônica. Um modelo longitudinal log-binomial negativo, com matriz de trabalho autorregressiva de ordem 1 [AR(1)], é utilizado para modelar a variável resposta erro de referência (Er). Com base nos resultados obtidos, verificou-se que, animais tratados com atorvastatina quando comparados com animais tratados com placebo, independente da duração do tratamento, tiveram sua capacidade cognitiva restruturada.

Palavras-chave: Atorvastatina; correlação; matriz de trabalho.

1. INTRODUÇÃO

A expectativa de vida do ser humano cresceu consideravelmente nos últimos anos, e com ela, a necessidade de desenvolvimento de estratégias farmacológicas para o tratamento e/ou prevenção de doenças neurodegenerativas agudas[1] ou crônicas[2]. Nesse sentido, o interesse de pesquisadores da área farmacêutica em descobrir tratamentos que minimizem os efeitos de uma série de doenças, cresce constantemente com objetivo de melhorar a saúde da população como um todo [9], [2], [3]. O foco central das análises estatísticas nestas pesquisas associam-se a modelagem da correlação oriunda das diversas medidas obtidas ao longo do tempo para um mesmo indivíduo. Entre as metodologias presentes na literatura, que permitem considerar a estrutura de dependência das unidades amostrais, é possível citar as equações de estimação generalizadas (EEG). Por meio desse método, modela-se a estrutura de correlação das observações por intermédio das EEG propostas por [11], as quais possibilitam acomodar as variações intra-indivíduos por meio de uma matriz de trabalho. Nessa conjectura, diversos autores tem utilizado as EEG em estudos farmacológicos para avaliar dados relacionados a isquemia cerebral, como observa-se nos trabalhos de [5] e [10].

[1]Lesão cerebral de origem isquêmica, infecciosa, traumática.
[2]Doença de Parkinson, Doença de Alzheimer entre outras.

O objetivo deste trabalho é apresentar o método das EGG aplicado a dados de um experimento farmacológico, em que, a variável resposta erro de referência (Er - número de erros cometidos por um animal ao entrar em um esconderijo falso pela primeira vez) é avaliada após isquemia cerebral, para verificar se o deficit cognitivo é amenizado pelos diferentes tratamentos utilizados.

2. MATERIAL E MÉTODOS

2.1. MATERIAL

Os dados utilizados neste estudo são resultados de um experimento realizado no delineamento inteiramente casualizado com cinco repetições longitudinais (testes de memória[3]), e 6 tratamentos arranjados num esquema fatorial 3x2 (3 grupos de tratamentos: Sham[4] (falso-isquêmico), Veículo (tratado com placebo) e Atorvastatina[5] (tratado com atorvastatina); 2 tempos de tratamento: 15 e 42 dias). Ratos com idade entre 12 e 15 meses foram treinados na tarefa do labirinto radial aversivo (LRA) [6] por 15 dias (fase pré-isquêmica). Iniciados os períodos de tratamentos, os animais foram avaliados quanto sua capacidade de se lembrar da tarefa aprendida durante o treino, ou seja, memória retrógrada, expressa por meio do número de erros de referência (Er - variável resposta) e outras variáveis não discutidas aqui.

2.2. MÉTODOS

Considerando a variável resposta em estudo, e que $\boldsymbol{Y_{ijkl}}$ representa o Er para o i-ésimo rato (i=1,...,67), de acordo com o j-ésimo tempo de tratamento (j=1,2), para o k-ésimo grupo de avaliação (k = 1,2,3) na l-ésima avaliação longitudinal (Teste de memória) (l=1,...,5) assume-se que

$$\boldsymbol{Y_{ijkl}} \sim BN(\boldsymbol{\mu_{ijkl}}, \boldsymbol{\phi}). \tag{1}$$

Nessa conjectura, adota-se um modelo longitudinal log-binomial negativo descrito na equação (2) para modelagem da variável resposta Er, em que por meio das EEG, é possível modelar a variação entre as observações intra-individuais incluindo na análise uma matriz de trabalho [1]. Assim, o modelo proposto é dado por

$$\eta = \log(\boldsymbol{\mu_{ijkl}}) = \alpha + \boldsymbol{\beta_j} + \boldsymbol{\gamma_k} + \boldsymbol{\tau} t_l + (\boldsymbol{\beta\gamma})_{jk}, \tag{2}$$

sendo α o efeito relacionado a média geral, $\boldsymbol{\beta_j}$ o efeito associado ao j-ésimo tempo de tratamento, $\boldsymbol{\gamma_k}$ o efeito relacionado ao k-ésimo grupo de avaliação e $\boldsymbol{\tau}$ é o parâmetro associado a l-ésima avaliação longitudinal. O último termo representa o efeito para a interação entre o j-ésimo tempo de tratamento e k-ésimo grupo de avaliação. Serão considerados efeitos de referência o grupo Sham, bem como o tratamento realizado durante 42 dias.
Inicialmente estima-se os parâmetros relacionados aos efeitos discriminados anteriormente, considerando um modelo linear generalizado padrão. Posteriormente, devido a dependência das observações, a matriz de variância e covariância assume a forma a seguir

$$V_i = A_i^{\frac{1}{2}} R(\theta) A_i^{\frac{1}{2}}, \tag{3}$$

com $A_i^{\frac{1}{2}}$ uma matriz diagonal, cujos elementos compõe a função de variância V_i [4], sendo $R(\theta)$ a matriz de trabalho. Ressalta-se que R(θ) pode assumir diferentes estruturas de acordo com as necessidades e característica dos dados avaliados. Nesse estudo, será utilizada a estrutura autorregressiva de ordem 1 (AR(1)), em que a correlação entre duas observações longitudinais adjacentes obedece a expressão

[3]Em cada teste de memória os animais foram submetidos a três tentativas. O resultado final da variável resposta foi registrado como soma dos resultados de cada tentativa.
[4]Grupo submetido apenas as manipulações experimentais.
[5]Fármaco utilizado no cotrole dos níveis de colesterol no sangue.

$$cor(\mu_{ijkl}, \mu_{ijkl'}) = \theta^{|l-l'|}. \quad (4)$$

sendo $l \neq l'$.

Comparações entre os tratamentos serão feitas por meio de contrastes e as diferenças serão consideradas significativas se $p < 0,05$. Os gráficos serão construídos com auxílio do *software* R, versão 3.2.4 [7] e o ajuste do modelo proposto será realizado no *software* SAS versão 9.4 [8] com uso da *procedure* genmod.

3. RESULTADOS E DISCUSSÕES

A Figura 1 representa os gráficos de perfis médios, com intervalo de confiança, para média da variável resposta na l-ésima avaliação longitudinal, para os tratamentos ao longo de 15 (Figura a esquerda) e 42 dias (Figura a direita) para animais dos grupos sham, veículo e atorvastatina.

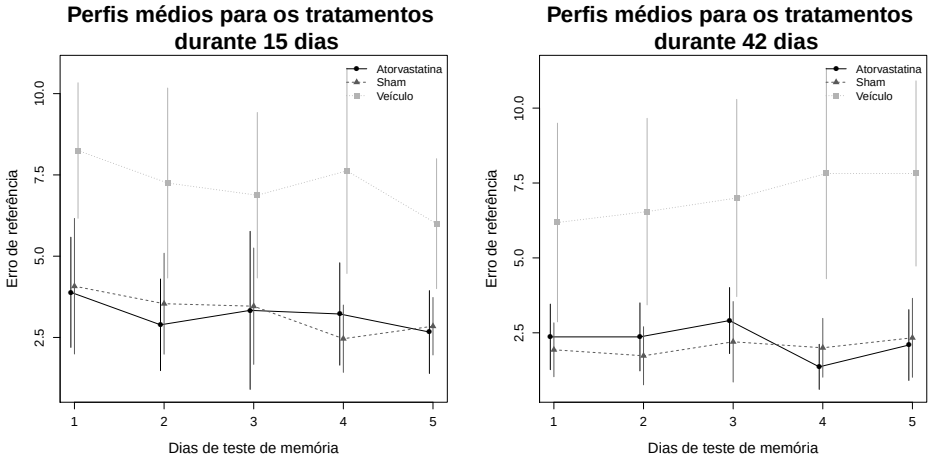


Figura 1: Gráfico de perfis médios para os tratamentos realizados durante 15 e 42 dias.

Por meio dos gráficos apresentados na Figura 1, verifica-se que, independente do tempo de duração do tratamento, animais do grupo veículo cometeram um número maior de erros de referência, esquecendo-se da tarefa realizada antes de serem submetidos a isquemia cerebral.

Na Tabela 1 são apresentadas as estimativas (Estimativa), erros padrão (EP) e intervalos de confiança, com limite inferior (LI) e limite superior (LS), para os parâmetros do modelo ajustado dado pela equação (2).

	LI	Estimativa	LS	EP
α	0,42	0,84	1,26	0,21
β_1	-0,41	0,06	0,54	0,24
β_2	0,68	1,19	1,69	0,26
γ_1	-0,02	0,44	0,91	0,24
$(\beta\gamma)_{11}$	-0,69	-0,99	0,50	0,31
$(\beta\gamma)_{12}$	-0,97	-0,39	0,19	0,29
τ	-0,07	-0,03	0,01	0,02
θ	0,43			

Tabela 1: Estimativas, limite inferior e superior (LI e LS), erro padrão (EP) para os parâmetros do modelo ajustado.

A partir das estimativas dos parâmetros do modelo, dado pela equação (2), independente do tempo de tratamento, evidencia-se que há grupos que diferem entre si quanto aos Er ($\chi^2_2 = 18{,}76$,

$p < 0,05$). O mesmo ocorre para as combinações de tempo de tratamento (15 e 42 dias) com os grupos de tratamento ($\chi^2_2 = 11{,}79$, $p < 0,05$). Isso sugere que de acordo com o grupo que o animal foi alocado (sham, atorvastatina, veículo) o número de Er se altera.

Quando comparados os Er para os animais do grupo veículo com grupo sham e atorvastatina houve diferenças estatísticas significativas, independente do tempo de tratamento ($\chi^2_1 = 8{,}88$ - 11,71, $p < 0,05$; $\chi^2_1 = 8{,}87$ - 11,99, $p < 0,05$), indicando que esses animais esqueceram da tarefa aprendida durante os treinamentos do LRA. Contudo, não houve diferenças significativas quando comparados os animais do grupo atorvastatina com sham ($p > 0,05$), sugerindo que animais desse grupo tiveram capacidade cognitiva parcialmente restabelecida, ou seja, lembram das tarefas aprendidas antes da isquemia.

4. CONCLUSÕES

Neste estudo foi aplicada a metodologia das EEG para análise de dados provenientes de um estudo longitudinal comportamental, onde verificou-se que o tratamento com atorvastatina foi capaz de aliviar o déficit cognitivo causado pela isquemia cerebral.

Referencias

[1] COSTA, S. C.; DEMETRIO, C. G. B.; SUGUINO, E.; GIOLO, S. R. **Modelando dados longitudinais de proporções de pegamento de camu-camu**. Revista de Matemática e Estatística, v. 22, n. 3, p. 113–127, 2004.

[2] FERREIRA, E. D.; ROMANINI, C. V.; OLIVEIRA, J. N.; PREVIDELLI, I. T.; MELO, S. R.; OLIVEIRA, R. M.; MILANI, H. **Fish oil provides a sustained antiamnesic effect after acute, transient forebrain ischemia but not after chronic cerebral hypoperfusion in middle-aged rats**. Behavioural Brain Research, v. 26, p. 101 – 110, 2014.

[3] GODINHO, J.; OLIVEIRA, J. N. de; FERREIRA, E. D. F.; ZAGHI, G. G. D.; BACARIN, C. C.; OLIVEIRA, R. M. W. de; MILANI", H. **Cilostazol but not sildenafil prevents memory impairment after chronic cerebral hypoperfusion in middle-aged rats**. Behavioural Brain Research, p. 61–68, 2015.

[4] HILBE, J. M. **Negative Binomial Regression**. 2. ed. [S.l.]: Cambridge University Press, 2011.

[5] LIU, K.; WANG, J.; YU, Z.; QIN, X.; WU, Y.; LI, N.; KUI, Y.; FANG, K.; WANG, X.; WU, T. et al. **Association study between PDE4D gene polymorphism and ischemic stroke**. Journal of Peking University. Health sciences, v. 45, n. 3, p. 359–363, 2013.

[6] PAGANELLI, R. A.; BENETOLLI, A.; LIMA, K. C. M.; CESTARI-JUNIOR, L. A.; FILHO, L. A. F.; MILANI, H. **A novel version of the 8-arm radial maze: effects of cerebral ischemia on learning and memory**. Journal of Neuroscience Methods, v. 132, n. 1, p. 9 – 18, 2004.

[7] R Core Team (2015). **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[8] SAS Institute Inc. SAS/STAT Software, Version 9.4. Cary, NC, 2015.

[9] SCHIAVON, A. P.; SOARES, L. M.; BONATO, J. M.; MILANI, H.; GUIMARÃES, F. S.; OLIVEIRA, R. M. Weffort de. **Protective Effects of Cannabidiol Against Hippocampal Cell Death and Cognitive Impairment Induced by Bilateral Common Carotid Artery Occlusion in Mice**. Neurotoxicity Research, v. 26, p. 301–316, 2014.

[10] WU, Y.; LI, S.; LIU, K.; TANG, X.; FANG, K.; LI, N.; LI, J.; WU, T.; CHEN, D.; HU, Y. **Study on the association between leukotriene A4 hydrolase gene polymorphism and ischemic stroke**. Zhonghua liuxingbingxue zazhi, v. 36, n. 8, p. 786–789, 2015.

[11] ZEGER, S. L.; LIANG, K. Y. **Longitudinal Data Analysis for Discrete and Continuous Outcomes**. Biometrics, v. 42, n. 1, p. 121–130, 1986.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

UMA APLICAÇÃO DOS POLINÔMIOS FRACIONÁRIOS EM MODELOS MISTOS

Edijane Paredes Garcia[1,2] e Luzia Aparecida Trinca[3]

[1]Programa de Pós-Graduação em Biometria, IBB, Universidade Estadual Paulista "Júlio de Mesquita Filho" - Unesp, SP/Brasil.
[2]Departamento de Estatística, ICE, Universidade Federal do Amazonas - UFAM, AM/Brasil.
[3]Departamento de Bioestatística, IBB, Universidade Estadual Paulista "Júlio de Mesquita Filho" - Unesp, SP/Brasil.

RESUMO

Os modelos de regressão utilizando polinômios fracionários fornecem uma abordagem alternativa e flexibilidade considerável no ajuste de modelos em que não há linearidade entre a resposta e as regressoras. A metodologia pode ser aplicada tanto em modelos de efeitos fixos quanto mistos. Neste trabalho, um conjunto de dados, em que a natureza da relação entre a resposta e a regressora é não linear, é analisado. Comparamos os ajustes de um modelo linear misto via polinômios convencionais com os polinômios fracionários.

Palavras-chave: Não Linearidade. Regressão. Suavização.

1. INTRODUÇÃO

A família dos polinômios fracionários (PFs) que foi introduzida por [8] é uma generalização da classe de funções dos polinômios usuais. Neste trabalho a flexibilidade dos modelos PFs é explorada como uma alternativa quando o ajuste do modelo por polinômios usuais apresenta potências de alta ordem, complicando o modelo e dificultando a interpretação dos parâmetros. Tem sido demonstrado que os PFs estão frequentemente entre os métodos de suavização menos enviesados para ajustes não lineares [3]. Vários pesquisadores aplicaram PFs a dados longitudinais não-lineares [5] mostrando sua utilidade na modelagem da parte fixa dos modelos mistos. Neste trabalho faremos uma comparação entre modelos mistos através dos polinômios usuais e fracionários utilizando o estudo sobre a taxa de ultrafiltração em função da pressão transmembrana.

2. MATERIAIS E MÉTODOS

Os dados utilizados neste trabalho encontram-se disponíveis na biblioteca *nlme* do software R com a denominação *Dialyzer*. Consistem de 140 medições realizadas em dialisadores de alto fluxo a fim de avaliar suas características de ultrafiltração *in vivo*. As taxas de ultrafiltração (em mL/hr) de 20 dialisadores de alto fluxo foram medidas em 7 diferentes pressões da transmembrana (tmp, em dmHg). Os dialisadores são avaliados *in vitro* usando sangue bovino, sendo: 10 dialisadores com taxas de fluxo sanguíneo (QB) de 200 dL/min e os outros 10 com QB de 300 dL/min. A natureza não linear da relação entre a taxa de ultrafiltração e pressão transmembrana é apresentada na

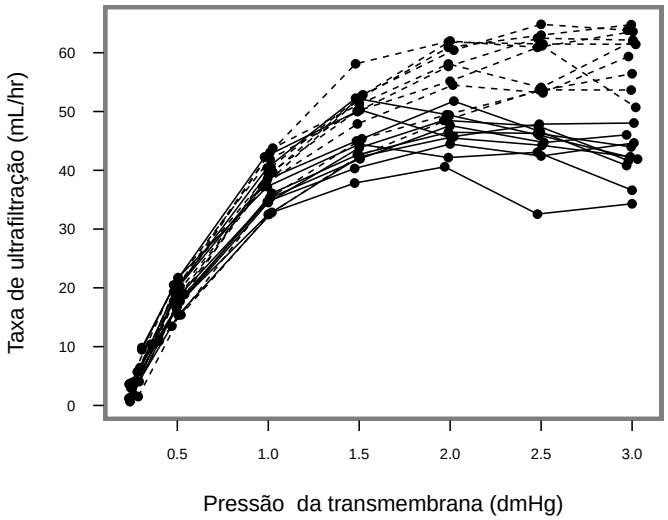


Figura 1: Gráfico de perfis individuais para as taxas de ultrafiltração (— $QB = 200$dL/min; --- $QB = 300$dL/min).

Figura 1. O modelo proposto na literatura para dados com tais características é originalmente não linear [11]. [4] e [6] apresentaram uma análise alternativa utilizando os modelos lineares de efeitos mistos. O modelo linear misto clássico considerado por [2] é expresso por:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \ldots, N, \tag{1}$$

em que $\mathbf{y}_i$ é o vetor de respostas $(n_i \times 1)$ para a i-ésima unidade, $\mathbf{X}_i$ é a matriz de especificação dos efeitos fixos $(n_i \times p)$, $\boldsymbol{\beta}$ é o vetor de parâmetros $(p \times 1)$, $\mathbf{b}_i$ é um vetor de efeitos aleatórios $(q \times 1)$ com $E(\mathbf{b}_i) = \mathbf{0}$ e $Var(\mathbf{b}_i) = \mathbf{D}$, $\mathbf{Z}_i$ é uma matriz de especificação para os efeitos aleatórios $(n_i \times q)$ e $\mathbf{e}_i$ é o vetor de erros aleatórios $(n_i \times 1)$ independente de $\mathbf{b}_i$, com $E(\mathbf{e}_i) = \mathbf{0}$ e $Var(\mathbf{e}_i) = \mathbf{R}_i$. As matrizes $\mathbf{D}$ e $\mathbf{R}_i$ são simétricas e positiva definidas de tamanho $(q \times q)$ e $(n_i \times n_i)$, respectivamente. Assim temos,

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad \text{e} \quad Var(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^{'} + \mathbf{R}_i,$$

em que o primeiro e o segundo termo da $Var(\mathbf{Y}_i)$ representam, respectivamente, a variabilidade entre as unidades amostrais e a covariância das observações intra unidades amostrais. Quando $\mathbf{R}_i = \sigma^2\mathbf{I}_{n_i}$, temos o modelo homocedástico, mas são possíveis outras estruturas de covariâncias. A análise de modelos mistos envolve duas partes: a análise da parte fixa e a análise da parte aleatória. Tanto a estimação dos efeitos fixos como a predição dos efeitos aleatórios depende da estimação das componentes de variância. A estimação dessas componentes depende da estrutura da matriz $\mathbf{D}$ e do método de estimação utilizado. Várias estruturas de covariâncias podem ser especificadas para a matriz $\mathbf{D}$. Mais detalhes sobre modelos mistos podem ser encontrados em [10], [2] ou [1], entre outros.

O melhor modelo linear de efeitos mistos selecionado por [6] para representar a taxa de ultrafiltração, y_{ij}, da j-ésima pressão transmembrana, x_{ij}, para o i-ésimo dialisador foi:

$$y_{ij} = (\beta_0 + \gamma_0 Q_i + b_{0i}) + (\beta_1 + \gamma_1 Q_i + b_{1i})\, x_{ij} + (\beta_2 + b_{2i})\, x_{ij}^2 + \beta_3 x_{ij}^3 + \beta_4 x_{ij}^4 + e_{ij}, \tag{2}$$

$$\mathbf{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \qquad e_{ij} \sim \mathcal{N}\left(0, \sigma^2 x_{ij}^{2\delta}\right), \delta \in \mathbb{R}$$

com

$$Q_i = \begin{cases} 0, & \text{se} \quad QB = 200\ dl/min \\ 1, & \text{se} \quad QB = 300\ dl/min, \end{cases}$$

os parâmetros $\beta_0, \beta_1, \beta_2, \beta_3$ e β_4 são o intercepto, os efeitos linear, quadrático, cúbico e quártico da pressão transmembrana, respectivamente; γ_0 e γ_1 representam o efeito do fluxo sanguíneo associado

com os efeitos fixos de β_0 e β_1; $\mathbf{b}_i$ o vetor de efeitos aleatórios, cuja estrutura de covariância especificada para a matriz $\mathbf{D}$ corresponde a matriz positiva definida simétrica geral e e_{ij} o erro aleatório cuja heterocedasticidade foi especificada como uma função da covariável *pressão transmembrana* por meio de um modelo de variância.

O método investiga se a hipótese de um efeito linear é aceitável ou se uma função não linear, a partir da classe dos PFs, melhora o ajuste dos dados sendo então preferível para descrever a influência funcional de uma covariável. A forma funcional de um polinômio fracionário de grau 1, denotado por PF1, é definida por $\beta_1 x^{p_1}$ com p_1 proveniente do conjunto de potências $S = \{-2; -1; -0,5; 0; 0,5; 1; 2; 3\}$, em que $x^0 = log(x)$. Se um PF1 não for suficiente para encontrar a melhor transformação em x, eles definem os polinômios fracionários de grau 2, denotado por PF2, com potências $\mathbf{p} = (p_1, p_2)$, cuja forma funcional é definida por $\beta_1 x^{p_1} + \beta_2 x^{p_2}$. Se $p_1 = p_2$, temos a forma funcional com potências repetidas que é definida por $\beta_1 x^{p_1} + \beta_2 x^{p_2} log(x)$. Para estimar $\mathbf{p}$ o método da máxima verossimilhança é aplicado ao conjunto S discreto.

O melhor modelo ajustado é aquele cujo valor de $\mathbf{p}$ apresenta maior verossimilhança. Para PF1, oito modelos devem ser ajustados, enquanto 36 modelos são examinados para PF2 (28 se $p_1 \neq p_2$ mais 8 com $p_1 = p_2$). Os autores em [8] sugerem um procedimento para selecionar a curva que melhor ajusta os dados. Antes de aplicar o procedimento deve-se decidir sobre o nível de significância (α) e sobre o grau do modelo PF mais complexo permitido. Escolhas típicas são $\alpha = 0.05$ e PF2. O procedimento para seleção de função (PSF), discutido por [9], foi implementado para seleção entre os modelos PF1, PF2 e o modelo linear misto.

Uma extensão do modelo PF univariado para o contexto dos polinômios fracionários multivariado nos possibilita incluir outras covariáveis consideradas relevantes. Para avaliar o efeito de preditores contínuos no contexto multivariado a combinação da *eliminação backward* com o PSF definem a construção do modelo multivariado por meio dos PFs. [9] fornecem um relato detalhado do método, juntamente com suas extensões, contendo numerosos exemplos numéricos no contexto de modelos de efeitos fixos.

Todas as análises foram feitas utilizando o pacote *nlme* do software R [7].

3. RESULTADOS E DISCUSSÃO

Usando as estruturas de covariâncias para as matrizes $\mathbf{D}$ e $\mathbf{R}_i$ definidas no modelo 2, os melhores modelos PF1 e PF2, apresentando menor AIC, tem potências $p = 0$ (AIC=746.60) e $p : (1; 1)$ (AIC=685.85), respectivamente. Aplicamos o procedimento PSF para selecionar o melhor modelo dentre os modelos: Linear Misto, PF1 e PF2. Os resultados são apresentados na Tabela 1. A comparação do PF2 versus modelo nulo mostra que a covariável pressão transmembrana com potências $(1; 1)$ é altamente significativa e o teste de PF2 versus modelo linear misto nos dá forte evidência que a função é não linear. A comparação final de PF2 versus PF1 nos mostra que é necessário um modelo PF2, uma vez que o modelo PF1 apresenta falta de ajuste significativa quando contrastado com PF2.

Modelo	Deviance	Potência	Comparação	Dif. Dev.	valor p
PF2	661.8547	1; 1	PF2 *vs* nulo	358.6426	< 0.001
PF1	724.5999	0	PF2 *vs* linear	199.5441	< 0.001
Linear	861.3988	1	PF2 *vs* PF1	62.7452	< 0.001
Nulo	1020.4970	–			

Tabela 1: Aplicação do procedimento de seleção funcional para análise das características de ultrafiltração de dialisadores de alto fluxo.

Como o melhor modelo é o PF2$(1; 1)$, duas covariáveis são requeridas: X e $Xlog(X)$. Considerando variação aleatória aos efeitos destas covariáveis, o modelo misto selecionado é dado por:

$$y_{ij} = (\beta_0 + \gamma_0 Q_i + b_{0i}) + (\beta_1 + \gamma_1 Q_i + b_{1i})\, x_{ij} + (\beta_2 + b_{2i})\, x_{ij} log(x_{ij}) + e_{ij}. \quad (3)$$

Embora o modelo misto com polinômios usuais tenha apresentado menor AIC (648,96), o modelo misto ajustado através dos polinômios fracionários possui menor número de parâmetros na parte fixa do modelo e melhor interpretação do fenômeno.

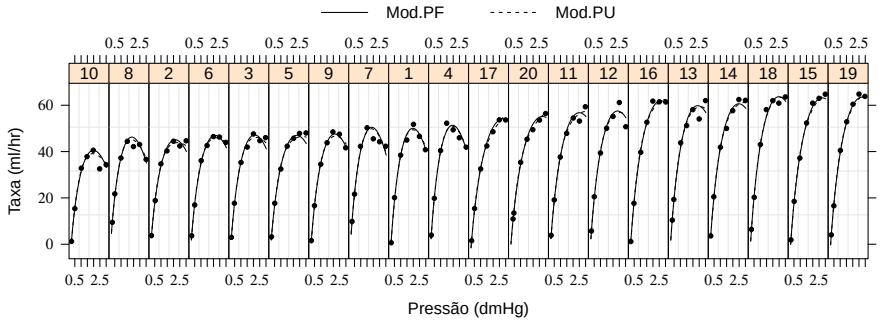


Figura 2: Gráfico da taxa de ultrafiltração predita versus pressão transmembrana por dialisador para os modelos 2 e 3.

Uma avaliação dos modelos 2 e 3 é fornecida pelo gráfico das predições por indivíduo, Figura 2. As diferenças entre as curvas geradas pelos dois modelos são muito pequenas. O resíduo padronizado do modelo 3 apresentou um padrão razoavelmente homogêneo de variabilidade com o aumento da pressão indicando que a modelagem da variância captou a heterogeneidade dos dados.

3. CONCLUSÃO

A taxa de ultrafiltração de dialisadores de alto fluxo, avaliada em diferentes pressões da transmembrana, pode ser modelada utilizando-se procedimentos regressão linear através de polinômios fracionários, com termos aleatórios. Embora a modelagem por polinômios usuais tenha apresentado um menor AIC, ela precisa de um número maior de parâmetros que a do modelo PF. Este último, por sua vez, mostrou-se mais atraente, pois as estimativas dos seus parâmetros tem uma melhor interpretação, uma vez que não apresenta potências de alta ordem.

AGRADECIMENTOS

Os autores agradecem o apoio financeiro da Pró-Reitoria de Pós-Graduação da Universidade Estadual Paulista "Júlio de Mesquita Filho" - Unesp.

Referências

[1] Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*, 2 edn. Wiley, New York.

[2] Fitzmaurice, G.; Laird, A. M.; Ware, J. H. (2011). *Applied Longitudinal Analysis*. Wiley, New York.

[3] Govindarajulu, U.S., Malloy, E.J., Ganguli, B., Spiegelman, D., Eisen, E.A. (2009). The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *International Journal of Biostatistics* 5, 1–19.

[4] Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. (1996). *SAS System for Mixed Models*. SAS Institute Inc., Cary, NC.

[5] Long, J. and Ryoo, J. (2010). Using fractional polynomials to model non-linear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology* 63, 177–203.

[6] Pinheiro, J. C., Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*. Springer-Verlag, New York.

[7] Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D., R Core Team. (2016). _nlme: Linear and Nonlinear Mixed Effects Models_. R package version 3.1-126, <URL: http://CRAN.R-project.org/package=nlme>.

[8] Royston, P., Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics* 43, 429–467.

[9] Royston, P., Sauerbrei, W. (2008). *Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables*. Wiley, New York.

[10] Verbeke, G., Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

[11] Vonesh, E. F., Carter, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* 48, 1–17.

Comunicacións orais

Sesión: Premio Ecoloxía-Medio Ambiente

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Multivariate Neighborhood Penalties for Analyzing Mussel Abundance at the Galician Coast

Hauke Thaden[1], Nadja Klein[1], Thomas Kneib[1], María P. Pata[2] and Carmen Cadarso-Suárez[2]

[1]Georg-August University Goettingen
[2]University of Santiago de Compostela

ABSTRACT

Mussel production plays a key role in the Galician economy. Moreover, Galicia is the major producer of mussels in Europe. In addition to the economic importance of the *Mytilus galloprovincials*, understanding the relationship between species behavior and environmental factors is of vital ecological interest. In this contribution, we develop a multivariate neighborhood penalty and embed it into the framework of simultaneous equation models in order to analyze the interrelationships between adult mussels, mussel seeds and their environment at the Galician coast.

Keywords: geoadditive regression, neighborhood penalties, correlated spatial effects, Mediterranean mussels

1. INTRODUCTION

The Mediterranean mussel, *Mytilus galloprovincialis*, plays a key role in Galician aquaculture: about 95% of the Spanish mussel production takes place in Galicia. Furthermore, Spain is the most important European provider of mussels. Therefore, understanding the abundance of mussels with respect to their environment is not only of ecological, but also of great economic interest ([1]).
We simultaneously investigate the influence of environmental factors on the presence of adult mussels and mussel seeds in order to understand the interrelationships between the former and the latter. We use discrete spatial effects on the scale of administrative areas at the Galician coast to represent environmental factors (like tidal height, magnetic course, ...) which influence the abundances of adult mussels and mussel seeds. To this end, we develop a spatial penalty which accounts for the idea that the spatial structures within the abundances of adults and seeds might be correlated. For the analysis, we embed this penalty into the framework of simultaneous equation models to additionally quantify the interrelationships between young and adult mussels assuming the ecological pathways displayed in Figure 1.

2. MULTIVARIATE NEIGHBORHOOD PENALTIES

In regression analysis with response y and predictor η, spatial effects are in general included as *smooth* functions of the locations of observations, i.e.

$$y_i = \eta_i + f_{spat}(location_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

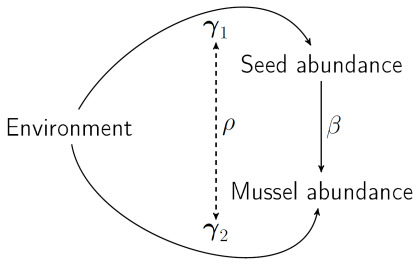


Figure 1: Assumed underlying ecological relationships between adult mussels, mussel seeds and environmental factors. The environmental effects on mussel and seed abundances are represented by $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, respectively. The correlation between these effects is denoted by ρ, whereas β is the direct effect of mussel seed abundance on adult mussel abundance.

Spatial smoothness is motivated by the common assumption that observations that are located close to each other are more similar than observations from sites far apart. If the spatial information is available on a discrete scale, e.g. as d administrative regions or grid cells, this proximity is generally modeled in terms of adjacency of the regions. In this case, a region specific effect γ_s is assigned to each region $s \in \{1, \ldots, d\}$ and (1) can be written as

$$y_i = \eta_i + \sum_{j=1}^{d} z_{ji}\gamma_{s_i} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where

$$z_{ji} = \begin{cases} 1, & \text{if observation } i \text{ is located in region } j, \\ 0, & \text{otherwise} \end{cases}$$

is an indicator for the regional affiliation of the ith observation. The spatial smoothness is then enforced by penalizing deviations between neighboring regional effects by adding the term[1]

$$\text{pen}_{spat} = \lambda \boldsymbol{\gamma}' \boldsymbol{K} \boldsymbol{\gamma}, \quad \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)', \tag{3}$$

to the optimization criterion (e.g. negative log-likelihood, least squares) during the estimation. Here, the penalty matrix $\boldsymbol{K}$ represents the neighborhood structure of the regions and the smoothing parameter λ controls the degree of spatial interdependencies between the regions. As a consequence, strong deviations between the effects of neighboring regions are penalized resulting in the desired spatial smoothness. In this univariate[2] scenario, the penalty term (3) can analogously be motivated as the precision $\tau^{-2}\boldsymbol{K}$ of correlated random regional effects (see [2, Chapter 8.2] for details). Note that there is a one-to-one correspondence between the smoothing parameter λ and the random effects variance τ^2.

In the spirit of this analogy, we extend the neighborhood penalization approach to regression models with multiple responses and more than one spatial effect. Assume a setup with responses y_1 (amount of mussel seeds in the introduced example) and y_2 (amount of adult mussels in the introduced example) observed on the same spatial scale (administrative areas in Galicia in the introduced example) with spatial effects

$$\boldsymbol{\gamma}_1 = (\gamma_{1,1}, \ldots, \gamma_{1,d})' \quad \text{and} \quad \boldsymbol{\gamma}_2 = (\gamma_{2,1}, \ldots, \gamma_{2,d})',$$

respectively. In case of independent spatial effects, penalization is straightforward and can be achieved by using

$$\text{pen}_{spat} = \tau_1^{-2}\boldsymbol{\gamma}_1'\boldsymbol{K}\boldsymbol{\gamma}_1 + \tau_2^{-2}\boldsymbol{\gamma}_2'\boldsymbol{K}\boldsymbol{\gamma}_2. \tag{4}$$

In terms of random effects, the parametrization of the precision matrix now is

$$\boldsymbol{A} \otimes \boldsymbol{K} = \begin{pmatrix} \tau_1^{-2}\boldsymbol{K} & \boldsymbol{0} \\ \boldsymbol{0} & \tau_2^{-2}\boldsymbol{K} \end{pmatrix} \in \mathbb{R}^{2d \times 2d},$$

[1] See [2, Chapter 8.2.4] for details about the penalty matrix $\boldsymbol{K}$.

[2] In the sense, that one spatial effect for one response is estimated.

with

$$\boldsymbol{A} = \begin{pmatrix} \tau_1^{-2} & 0 \\ 0 & \tau_2^{-2} \end{pmatrix}.$$

Setting $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')$, it is easily verified that (4) can equivalently be written as

$$\text{pen}_{spat} = \boldsymbol{\gamma}' (\boldsymbol{A} \otimes \boldsymbol{K}) \boldsymbol{\gamma}. \tag{5}$$

Finally, we achieve a general correlation structure of the spatial effects by setting

$$\boldsymbol{A} = \begin{pmatrix} \tau_1^2 & \rho\tau_1\tau_2 \\ \rho\tau_1\tau_2 & \tau_2^2 \end{pmatrix}^{-1} \tag{6}$$

in (5). As visualized in Figure 2, the resulting parametrization yields a straightforward decomposition of the overall spatial structure: as in the univariate case, τ_j describes the variability between the regional effects on $y_j, j = 1, 2$ (red arrows), whereas ρ quantifies the correlation between the two spatial effects (blue arrow).

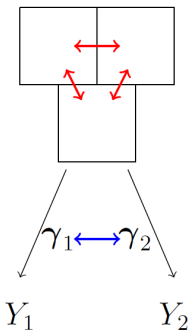


Figure 2: Decomposition of spatial correlation into interdependencies (red arrows) between regions (squares) and correlation between the effects $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ (blue arrow).

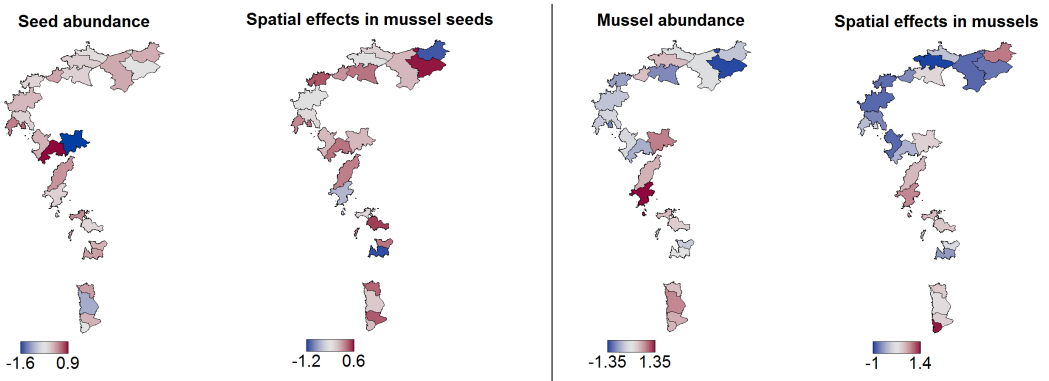
3. EMPIRICAL RESULTS

The data were collected at the Galician coast between 2005 and 2009. Abundances of adult mussels and mussel seeds are represented as percentages of coverage on predefined transects. In total, $n = 89$ transects have been investigated in the intertidal zone along the coast of $d = 25$ administrative regions. Figure 3 (a) and (c), respectively, show the spatial structures of adult mussel and mussel seed abundances on the log-scale.[3]
We translate the assumed underlying ecological pathway from Figure 1 into the simultaneous equations

$$\log(seed_i) = f_{spat,1}(s_i) + \varepsilon_{1i}, \tag{7}$$
$$\log(mussel_i) = \beta \log(seed_i) + f_{spat,2}(s_i) + \varepsilon_{2i}. \tag{8}$$

As explained in Section 2, the spatial effects $f_{spat,j}, j = 1, 2$, are included as regional dummy variables and the according effects are penalized via (5). Note that in this case, estimation is based on the assumption of joint normality and independence of ε_1 and ε_2. Hence, the presented estimation strategy is a *bivariate* penalized maximum likelihood approach. The estimated spatial effects are displayed in Figure 3 (b) and (d). The patterns of the effects within mussel seeds and adult mussels are - to a certain degree - conflictive. The regions at the northern coast of Galicia seem to favor the abundance of mussel seeds, whereas the environmental influences on mussel abundance in the same area are predominantly negative. This inverse pattern is also captured by the negative estimate $\hat{\rho} = -0.35$ of the correlation of the two spatial effect. In the southern regions, the differences are not that pronounced. Along with the estimated spatial effects, we find a slightly positive effect of

[3] See [1] for details on the data collection procedure.



(a) Seed abundances on a log-scale (centered)

(b) Estimated spatial effects for seed abundance

(c) Mussel abundances on a log-scale (centered)

(d) Estimated spatial effects for mussel abundance

Figure 3: Spatial structures and estimated regional effects of the response variables *seed abundance* ((a) and (b)) and *mussel abundance* ((c) and (d)).

mussel seed abundance on the presence of adult mussels on the log-scale, namely $\hat{\beta} = 0.05$.
Note that the assessment of uncertainty is not straightforward in this case, since the uncertainty resulting from the additional estimation of the penalty parameters τ_1, τ_2 and ρ needs to be accounted for. It is therefore at this point still under development.

4. CONCLUSIONS

This contribution extends the concept of neighborhood penalties to regression setups with bivariate responses. Additionally, the penalty is further generalized by including an additional correlation parameter for the occurring spatial effects. For the application, this penalty is embedded in a bivariate simultaneous equation model alongside with a maximum likelihood estimation approach. It is - however - conceptually straightforward to include the idea into structured additive regression models with alternative distributional assumption. Also, an extension in cases of more than two spatial effects is obvious, but caution is advised in terms of identifiability and interpretability.
In the data example, we find that - based on the present sample - mussel seeds and adult mussels react to some degree inversely to their environment. The direct effect of mussel seed abundance on the presence of adult mussels is, if existing at all, slightly positive.
The assessment of uncertainty of the resulting estimates is still under development and might be facilitated in a Bayesian approach.

ACKNOWLEDGMENT

The work of Hauke Thaden, Nadja Klein and Thomas Kneib is supported by the German Research Foundation via the research training group 1644 on “Scaling problems in statistics”. This collaboration was set up within the DAAD project “Extended Structured Additive Regression and its Applications”.

References

[1] Pata, M. P., Kneib, T., Cardaso-Suárez, C., Lustres-Pérez, V. and Fernández-Pulpeiro, E., (2012). Categorical structured additive regression for assessing habitat suitability in the spatial distribution of mussel seed abundance. *Environmetrics* 23, 75–84.

[2] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression*. Springer Berlin Heidelberg, Berlin, Heidelberg.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Modelling intra- and inter-day variability of NO_2 concentrations in Portugal

Andreia Monteiro[1], Raquel Menezes[2] and Maria Eduarda Silva[3]

[1]Department of Mathematics and Applications & Centre of Mathematics, University of Minho & CIDMA; andreiaforte50@gmail.com
[2]Department of Mathematics and Applications & Centre of Mathematics, University of Minho; rmenezes@math.uminho.pt
[3]CIDMA & Faculty of Economics, University of Porto; mesilva@fep.up.pt

ABSTRACT

Nitrogen dioxide (NO_2), is a pollutant that is toxic by inhalation and there is evidence that long-term exposure to this, at high concentrations, has adverse health effects, namely in respiratory and cardiovascular systems. The goal of this study is to characterize the spatial and temporal evolution of NO_2 concentration levels, taking into account that environmental data often incorporate distinct recurring patterns in time, imposed by social habits. We aim at capturing the cyclic nature of these environmental indicators, identifying the intra and inter-day variability. Simultaneously, we aim at modelling the temporal and spatial correlation inherent to this type of data. This study focus on NO_2 hourly data collected in Portugal from October 1 to December 31, 2014. An initial exploratory study suggests that there are two main seasonal effects in the data and identifies variables such as the type of site, environment, and the day of the week as possible explanatory variables. Furthermore, the analysis of the correlation between meteorological parameters, as air temperature, wind speed and relative humidity and NO_2 levels identifies significant negative associations among them. After describing the trend function, geostatistical approaches are applied to the resulting residuals with the aim of characterizing the space-time variability and deriving the predicted values through the kriging tools. This methodology can be used to complement the current design sampling, where there are districts without monitoring stations or with many missing values. Moreover, as meteorological data are available earlier than NO_2 levels, we draw scenarios for NO_2 levels for 2015.

Keywords: Geostatistics; Spatio-temporal modelling; Time Series; Environment; NO_2.

1. MOTIVATING EXAMPLE

This study analyzes hourly measurements of NO_2 from October 1 to December 31, 2014 corresponding to highest mean NO_2 concentrations along the year. The data set is available in QualAr site. The available data include information about the type of the site where the station is placed (background, industrial or traffic) and the environment of the zone (urban, suburban or rural). From the 49 stations analyzed, 33 are background, 10 traffic and 6 industrial, 29 are located in urban areas, 11 in rural areas and 9 in suburban areas. The mean of the hourly NO_2 concentrations is 20.6 $\mu g/m^3$ and standard deviation 21.9. Hourly mean values for both weekdays and weekends, represented in Figure 1 indicate that NO_2 levels show two daily peaks, one in the morning (8:00 A.M.) and one in the afternoon (6:00 P.M.) which coincide with rushhour traffic, with the second peak being more pronounced than the first. Moreover, the mean NO_2 concentrations

are much lower on weekends than on weekdays, displaying, also, smaller variations on weekends, which reflect reduced levels of vehicular emissions on non-working days. Thus, the two main seasonal effects in the data: intra-day as well as intra-week periodicities, may be, at least partially, explained by characteristics of the station. In fact, the stations located in traffic areas and urban zones present higher values for their NO_2 quartiles as well higher variability. This analysis indicates that the type of site and environment must be considered as explanatory variables.

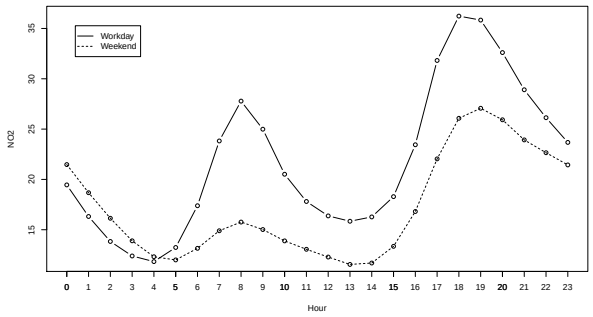


Figure 1: Mean hourly NO_2 concentrations, for workdays and weekends.

Furthermore, it is acknowledge that meteorological conditions influence NO_2 levels. Hourly data from the following meteorological variables were obtained from Weather Underground site: wind speed (km/h); air temperature (^oC) and relative humidity (%). The analysis of the correlation between these meteorological parameters and NO_2 levels identified significant negative associations among them with the strongest correlations occurring at a 9-hour lag with air temperature, a 1-hour lag with wind-speed and a 5-hour lag with relative humidity. Therefore, these meteorological variables are also be considered as explanatory variables.

2. MODELLING THE LARGE AND SMALL SCALE VARIATION

Consider a random function, $\{Z(\mathbf{s},t) : (\mathbf{s},t) \in \mathbb{R}^d \times \mathbb{R}\}$ index in space by $\mathbf{s} \in \mathbb{R}^d$ and in time by $t \in \mathbb{R}$. The random process $Z(\mathbf{s},t)$ can be decomposed into, $Z(\mathbf{s},t) = \mu(\mathbf{s},t) + \delta(\mathbf{s},t)$, where $\mu(\mathbf{s},t)$ denotes the trend and $\delta(\mathbf{s},t)$ is a zero-mean stationary residual. For NO_2 concentrations, exploratory analysis revealed that it is a continuous variable with an asymmetric distribution. To estimate $\mu(\mathbf{s},t)$ we use a Generalized Linear Model (GLM) and assume that the response variable follows a Gamma distribution with log link. The seasonal effects in the data can be modeled by adding some components of a mixed model [1], leading to:

$$\eta(\mathbf{s},t) = \alpha + \beta_1 X_1(\mathbf{s},t) + \beta_2 X_2(\mathbf{s},t) + ... + \beta_k X_k(\mathbf{s},t) + \sum_{l=1}^{L} [\phi_{l,1}\cos(l\,2\pi t f_l) + \phi_{l,2}\sin(l\,2\pi t f_l)] \quad (1)$$

where X_i are the regressors, function of the explanatory variables, α, β_i, $\phi_{l,1}$, $\phi_{l,2} \in \mathbb{R}$ the regression parameters estimated by the model and f_l is the frequency of the lth harmonic.
After estimation of the large-scale variation (trend), the next step is to analyze the dependence structure of the small-scale variation (residuals) with the aim of characterizing the space-time variability, following a similar approach to that described in [2].

3. CHARCTERIZATION OF $\mu(\mathbf{s},t)$ and $\delta(\mathbf{s},t)$

As described in Section 2, we first model the trend of NO_2 data using a GLM. We consider six explanatory variables: type of site, type of environment, if weekend, air temperature (9-hour lag effect), wind speed (1-hour lag effect) and relative humidity (5-hour lag effect). For modeling the seasonal effects in the data set we use the model given in (1).
To estimate the frequency of the data, the periodicity for stations without missing values was calculated and we obtain periodicities equal to 12 and 24 hours. The results of the gamma regression of the hourly NO_2

concentrations show that the values of NO_2 concentrations are greater during the week and in monitoring stations where the environment is urban or suburban and the type of site is traffic. Regarding the meteorological variables, these variables have significant negative associations with NO_2 levels, wind speed has a stronger influence on NO_2 concentrations than humidity and air temperature. The acquired coefficient of determination shows that 41% of the large-scale variation of NO_2 concentrations is explained under this trend model.
After estimation of the large-scale variation, we fit a valid variogram. We start by analyzing the marginal spatial and marginal temporal correlation structures. To decide which non-separable covariance structure to adopt we use a cross-validation study. We select the sum-metric model with a Exponential function for the temporal variogram and Gaussian functions for the spatial and the spatio-temporal components. Under this selection, the fitted final model is represented in Figure 2 (right). According to the results, we conclude that the majority of the total variation is explained by the spatial component. The temporal and spatio-temporal components have a smaller contribution. Furthermore, NO_2 concentrations have a significative spatial correlation up to 40 km and a temporal correlation up to 100 hours (4 days).

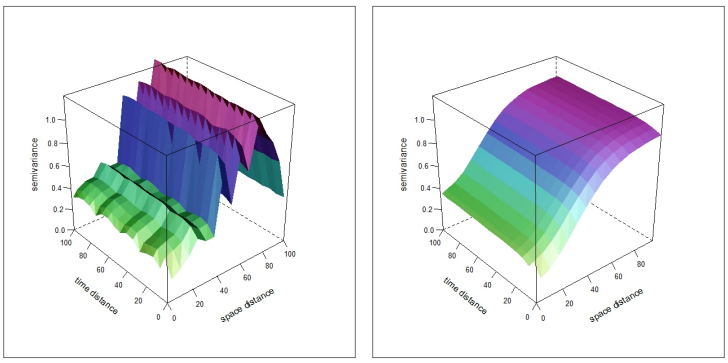


Figure 2: Plots of the empirical estimator (left) and the fitted model (right) for the space-time variogram.

4. PREDICTION AND FORECASTING IN THE PORTUGUESE CASE

This methodology is important because it can be used to complement the current design sampling, where there are districts without monitoring stations or with many missing values. To illustrate this idea, we obtain the predicted maps for two days, 21 and 23 of November, Friday and Sunday which correspond, respectively, to the days of the week with higher and lower levels of NO_2 and select three times of the day, 8:00 A.M. and 6:00 P.M. which are the daily peaks and 13:00 P.M. a daily minimum. The results are displayed in Figure 3, and confirm that the general spatial pattern of the NO_2 concentrations are not persistent during the day and over the days, achieving higher values on the coast, namely in zones of Braga, Porto, Aveiro, Lisbon and Sines.

Knowledge of the space-time dependence allows to predict missing values in a specific station. These missing values may occur occasionally at some time points or when the station becomes inactive. The estimation of small-scale and large-scale variation from 2014-10-06 00:00 to 2014-10-11 00:00 for Vila Nova da Telha, a suburban and background station from Maia county are displayed in Figure 4.

ACKNOWLEDGMENTS

The authors acknowledge Foundation FCT (Fundação para a Ciência e Tecnologia) for funding through Individual Scholarship PhD PD/BD/ 105743/2014, Centre of Mathematics of Minho University and Center for Research & Development in Mathematics and Applications of Aveiro University within project UID/MAT/04106/2013.

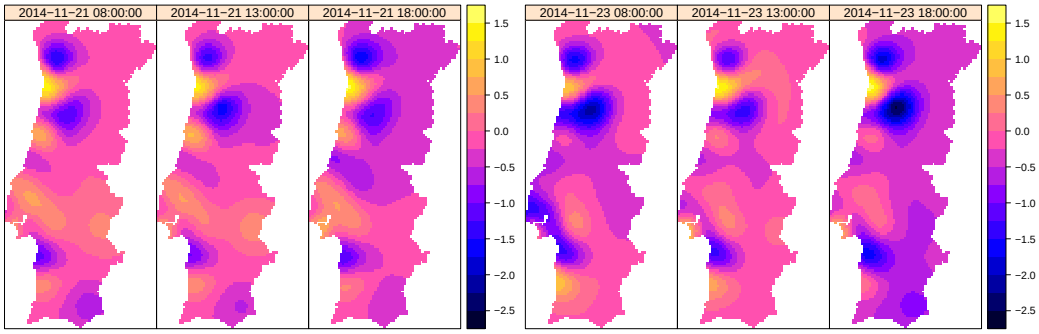


Figure 3: Space kriging maps for 2014-11-21 (Friday) and 2014-11-23 (Sunday).

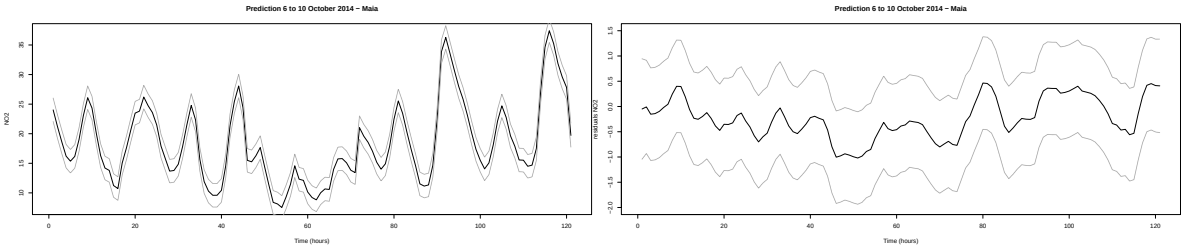


Figure 4: Estimation of the large-scale (left) and small-scale (right) variation of NO_2 concentrations in Maia Station. The dashed lines identify the 95% confidence intervals.

References

[1] Kyriakidis, P. and Journel, A. (1999). Geostatistical space-time models: A review. *Mathematical Geology* **31**, 651–684.

[2] Menezes, R., Piairo, H., García-Sóidan, P. and Sousa, I. (2015). Spatial–temporal modellization of the NO_2 concentration data through geostatistical tools. *Statistical Methods Applications* **25**, 107–124.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

ESTIMATING BLAINVILLE'S BEAKED WHALES ABUNDANCE OVER TIME: IMPROVING DIVE COUNTING METHODS

Patrícia A. Jorge[1], Jessica W. Shaffer[2], Tiago A. Marques[3,4], Helena Mouriño[1], Len Thomas[3], David J. Moretti[2], Nancy DiMarzio[2]

[1]Departamento de Estatística e Investigação Operacional, Faculdade de Ciências da Universidade de Lisboa, Bloco C6, Piso 4, 1749-016, Lisboa, Portugal
[2]Naval Undersea Warfare Center, 1176 Howell Street, Newport, RI 02841, USA
[3]Centre for Research into Ecological and Environmental Modelling, University of St Andrews, The Observatory, Buchanan Gardens, Fife, KY16 9LZ, UK
[4]Centro de Estatística e Aplicações da Universidade de Lisboa, Faculdade de Ciências, Universidade de Lisboa, Bloco C6, Piso 4, 1749-016, Lisboa, Portugal

RESUMO

Efficient and precise methods for estimating the abundance of natural populations is required for their effective management and conservation. Consequently, it is important to optimize existing models. Methods allowing *in situ* monitoring with the minimum amount of human intervention are becoming more popular to study natural populations.

Blainville's beaked whales (*Mesoplodon densirostris*, *Md*) are the widest ranging mesoplodont whales. Nonetheless, they are rarely seen, as the majority of their time is spent deep in the ocean. This species is known to associate in groups, exhibiting metronomic dive behaviour. They tend to vocalize only during deep foraging dives using broadband, ultrasonic echolocation signals known as clicks, with a bandwidth from 26 to 51 kHz [1]. Where these clicks occur indicates where whales search and encounter prey [2].

The AUTEC (Atlantic Undersea Test and Evaluation Center) is a U.S. Navy testing and training range located in the Tongue of the Ocean (TOTO) in the Bahamas. The range includes a large network of hydrophones cabled to shore that can be used to detect *Md* echolocation clicks. Given the hydrophone spacing and sensitivity, combined with the animals' source level, the *Md* dives occurring on the AUTEC range are certain to be detected [3].

Using *Md* click data collected on AUTEC hydrophones, a method for estimating *Md* abundance is presented. As the individuals spend little time on the surface, other traditional abundance estimation methods like line transect distance sampling may lead to inconclusive results. Passive acoustic methods are used to detect and classify *Md* echolocation clicks. Click detections are then assigned to an *Md* group foraging dive. The start of a foraging dive is assumed to be the time of the first detected echolocation click.

The proposed approach extends previous methods developed by Moretti *et al.* (2010) [3] and DiMarzio et al. (2008) [4]. Instead of considering an average group size value based on literature, the size of each group will be estimated considering variables derived from the acoustic data, via a generalized linear model.

We consider two different data sets. The first is used to build the model of group size as a function of the groups acoustic footprint, and the second is used to estimate density leveraging on the group size model.

The first data set consists of 43 deep dives identified between 2005 and 2008, for which the group size was visually confirmed. Potential explanatory variables include, for each detected dive, the number of the hydrophones which detected the echolocation clicks, the number of clicks detected in each hydrophone, the corresponding start and end of the echolocation period, and a binary variable which indicates whether or not the particular hydrophone was located on the edge. Further, a number of derived variables based on these were constructed. The group size in this modelling data ranged between 1 and 6 whales, and some groups would remain within the site for several hours.

The second data set is a long time series of AUTEC data from which density will be estimated. It includes 3 separate periods of time in 2011: (1) 61 days from the 27th of April to the 27th of June, (2) 19 days from the 19th of October to the 7th of November, and (3) 30 days from the 1st to the 31st of December. These data were processed using the same procedure that generated the data for the group size model, and the group size will be estimated for all deep dives detected. This method allows to quantify how many dives occurred on the AUTEC range during that period, and to estimate the total number of animals involved. This in turn allows the estimation of density over time using this improved version of the dive counting method proposed by Moretti *et al.*

Preliminary analysis indicates that the group size can be predicted from acoustic footprint of the group via the available covariates. The most significant variables appear to be the click duration, the number of hydrophones involved and the click rate. In total 15493 potential deep dives were detected in the second data set. A preprocessing of the data to exclude false positives was implemented, based on a set of biologically infeasible characteristics: (1) detections occurring on a single hydrophone and (2) a minimum threshold of 400 clicks detected. This resulted in a much more biologically plausible distribution of observed vocal lengths, matching what would be expected given described values in the literature [5]. We further removed groups detected only on edge hydrophones, considering these would correspond to groups outside the area of inference (as in [3]). This led to 8275 detections considered to correspond to true beaked whale deep dives. The first period of time recorded 4551 dives, with an average of 74 dives per day; the second period showed 1450 dives with an average of 76 dives per day; and the third one registered 2294 dives with an average of 76 dives per day.

The temporal patterns in density and group sizes will be described, allowing a better understanding how these animals use the AUTEC range over time for feeding.

Palavras e frases chave: Blainville's beaked whales, Echolocation, Dive counting, Density Estimation, Group size.

AGRADECIMENTOS

We would like to acknowledge the effort of the NUWC M3R team involved in the data processing. Also, Diane Claridge and Charlotte Dunn of the Bahamas Marine Mammal Research Organisation for all the effort involved in the visual confirmation of group sizes. This research is an output of GROUPAM, a project funded by the Office of Naval Research (USA). TAM is partially funded by FCT, Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

REFERÊNCIAS

[1] Johnson, M., Madsen, P. T., Zimmer, W. M. X., Aguilar de Soto, N., Tyack, P. L. (2006). Foraging Blainville's beaked whales (*Mesoplodon densirostris*) produce distinct click types matched to different phases of echolocation. *The Journal of Experimental Biology* 209, 5038-5050

[2] Arranz, P., Aguilar de Soto, N., Madsen, P., Johnson, M. (2011). Following a foraging fish-finder: diel foraging habits of Blainville's beaked whale revealed by echolocation. *Plos One* 6(12): e28353.

[3] Moretti, D., Marques, T., Thomas, L., DiMarzio, N., Dilley, A., Morrissey, R., McCarthy, E., Ward, J., Jarvis, S. (2010). A dive counting density estimation method for Blainville's beaked whale (*Mesoplodon densirostris*) using a bottom-mounted hydrophone field as applied to a Mid-Frequency Active (MFA) sonar operation. *Applied Acoustics* 71, 1036-1042

[4] DiMarzio, N., Moretti, D., Ward ,D., Morrissey, R., Jarvis, S., Izzi, A., Johnson, M., Tyack,P., Hansen, A., (2008). Passive acoustic measurement of dive vocal behavior and group size of Blainville's beaked whale (*Mesoplodon densirostris*) in the tongue of the ocean (TOTO). *Canadian Acoustics* 36(1), 166-173

[5] Tyack, P. L., Johnson M., Soto, N. A., Sturlese, A., Madsen P. T. (2006). Extreme diving of beaked whales. *The Journal of Experimental Biology* 209, 4238-4253

[6] Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., Tyack, P. L. (2013). Estimating animal population density using passive acoustics. *Biological Reviews* 88, 287-309

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Captando a orixe dos patróns de distribución de abundancia de especies do plancto microbiano: unha análise da dispersión e a ocupación de nichos

Miguel Branco [1] , Salvador Pueyo [2], Isabel G. Teixeira[3], Francisco G. Figueiras 3[3] e Pedro Cermeño [4]

[1] Universidade de Vigo [2] Institut Català de Ciències del Clima (IC3) [3] Instituto de Investigacións Mariñas de Vigo (IIM-CSIC) [3] Institut de Ciències del Mar (ICM-CSIC)

RESUMO

A alta riqueza de especies raras nas comunidades de plancto microbiano fai cuestionar cales son os mecanismos que as manteñen ao longo do tempo. Mentres que algúns modelos ecolóxicos xustifican que eses patróns son froito de altas taxas de dispersión, outros describen como comunidades con especies ocupando cadanseu nicho poden seguir ese mesmo tipo de distribucións. Unha avaliación dos patróns de β-diversidade do plancto microbiano da Ría de Vigo mostrou que as taxas de inmigración son un fenómeno secundario. A diversidade funcional, como medida de nichos ocupados pola comunidade, revelou que os cambios na distribución de abundancia de especies implican re-estructuracións funcionais e explican a persistencia desas especies. Nos momentos de maior produción primaria do plancto microbiano as especies dominantes ocupan o mesmo nicho e compiten entre si polos mesmos recursos. Iso vese reflectido como unha maior desviación no patrón de distribución de abundancia de especies.

Palabras e frases clave: patróns de diversidade, distribución de abundancia de especies, log-series, teoría idiosincrática, β-diversidade, diversidade funcional

1. INTRODUCIÓN

As comunidades do plancto microbiano mariño fórmanas milleiros de individuos e centos de especies. Namentres que un pequeno grupo de especies son abundantes, unha "gran fracción" aparecen en baixas cantidades: son *especies raras*. Para explicar estas distribucións en forestas temperadas, os modelos neutrais (NM) definiron a distribución de abundancia de especies (SAD) *Zero-Sum Multinomial (ZSM)* argumentando que a estocasticidade demográfica e a dispersión son os únicos factores que as xeran e manteñen. Iso supón considerar que todas as especies cumpren a mesma función no ecosistema e omitir que haxa diferenzas de capacidade de competición por nichos ecolóxicos (espazo, recursos...) entre elas, e que por exclusión competitiva deriven en extinción de especies. Por iso, como resposta aos NM, a Teoría Idiosincrática demostrou que a ZSM equivale á clásica *log-series*. Argüíu como é que cada especie pode seguir cadansúa distribución de abundancias e interaccionar de diferente xeito coas outras especies sen que iso repercuta na SAD. As diferenzas por distinta capacidade de ocupación de nichos veríanse canceladas a nivel de comunidade. Argumentado isto, cambios pequenos pero significativos dunha log-series fan que esta mude, e siga unha lei potencial, reflectindo a macroescala os mecanismos ecolóxicos que actúan dentro da comunidade (*Fig. 1*).

Mentres que os NM consideran a dispersión como fundamental, a Teoría Idiosincrática sostén que a existencia de competición non é inconsistente con estes patróns observados, nos que coexisten ducias de especies. Aquí faremos unha análise de comunidades de microplancto mariño para elucidar que mecanismos xeran as súas distribucións de abundancia e manteñen nelas sempre presente o grupo de *especies raras*.

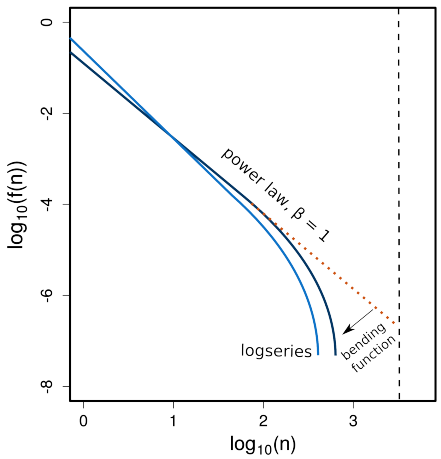


Figura 1: **A razón da log-series**. Función probabilística log-series, $f(n) = \kappa n^{-\beta} e^{\theta n}$, dunha comunidade de n individuos e *f(n)* especies en escala logarítmica [azul escuro] (Pueyo 2006, Pueyo *et al.* 2007). Segundo os NM, un poboación cunha medra por deriva estocástica seguirá unha log-series cunha $\kappa n^{-\beta}$ dunha $\beta = 1$, ata chegar ao máximo de individuos que poida acoller a comunidade . Segundo a Teoría Idiosincrática, as desviacións de β da log-series reflicten os procesos de ocupación de nichos dominantes [azul claro].

2. MÉTODOS E RESULTADOS

Empregando unha serie anual de mostras de fitoplancto mariño recollidos na Ría de Vigo no 2012 observamos que sempre seguen unha SAD log-series. Seguindo a Pueyo 2006, a pendente da distribución en escala log-log, ou o expoñente β da *pdf*, mostrou desviacións con compoñentes estacionais. Para medir cal foi a fracción de especies raras que entraron por inmigración na comunidade, elaboramos os gráficos ternarios de Koleff 2003 e medimos a β-diversidade seguindo a Baselga 2010, 2013 (*Fig. 2*). As entrada de especies ao longo do ano debida a inmigración foi menor do 11 % e os cambios na estrutura da comunidade debéronse a aumentos e diminucións nas abundancias relativas de cada especies. Con iso, a inmigración non xustificou a persistencia constante de especies a baixas abundancias. Para cuantificar os nichos de Hutchinson que foron ocupados en cada estación, empregamos como *proxy* os índices de diversidade funcional (FD) (*Fig. 3*). A medida da FD fíxose considerando as características taxonómicas de cada especie ligadas ao seu papel como principais produtores primarios dos ecosistemas mariños. O FD de Dispersión Funcional, que mide a distancia funcional entre os grupos de especies máis abundantes, e a Redundancia Funcional, ou duplicidade de especies coas mesmas características funcionais, mostraron unha relación directa coas variacións estacionais do expoñente β. Nos momentos de alta produción primaria, a Dispersión Funcional chegou aos mínimos e β desviouse máis de 1. Observouse o patrón inverso para a relación entre a produción primaria neta da comunidade e os cambios da *pdf* (*Fig. 4*).

3. CONCLUSIÓNS

A dispersión, que é o factor que os NM consideran clave, non deu explicado por si soa nin a persistencia nin os cambios na log-series do microplancto mariño. Si o foron as variacións na FD, xa que cuantifican as re-estruturacións da comunidade por intercambios no papeis que xogan as especies ao longo da sucesión ecolóxica. A Dispersión Funcional e a Redundancia Funcional, ao chegar aos seus maiores valores, mostraron cando é que as especies dominantes se solaparon e realizaron a mesma actividade no ecosistema. Neses períodos relegan ás outras especies a baixas abundancias. Porén, cando estes índices de FD acadaron valores baixos, foi cando as especies ocuparon diferentes funcións no ecosistema e equipararon as súas abundancias. Así é que a maior distancia funcional das especies reflicte a ocupación de novos nichos ou do aproveitamento dos baleiros.

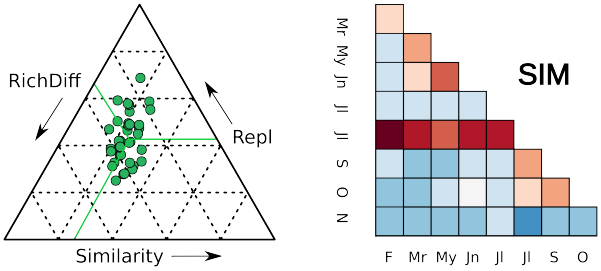


Figura 2: **Análise da β-diversidade. a)** Gráfico ternario comparando por pares as comunidades da serie temporal (Koleff *et al. 2003*, Legrende 2014). A nube de puntos centrada, alonxada do eixo de diferenza por riqueza de especies cuantifica a inmigración e diferenzas por cambios en abundancia de especies. **b)** Baseados en Sørensen, diferenzas da β-diversidade en escala temporal. Compáranse os pares de mostreos e as similitudes representase coa cor azul (Baselga 2010, 2013)

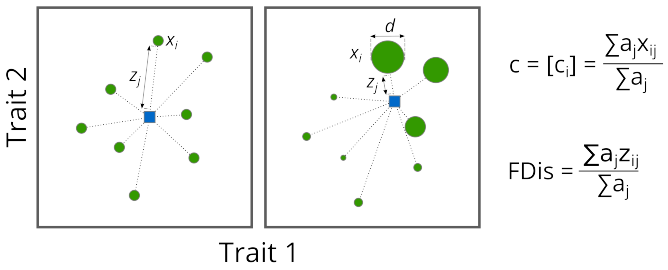


Figura 3: **Os extremos da dispersión funcional**. Considerando dúas características funcionais (como o volume celular ou capacidade de flotación), a distancia media entre todas as especies dunha comunidade, ponderada polas súas abundancias, é o índice de Dispersión Funcional. (Maior valor á esquerda, menor á dereita)

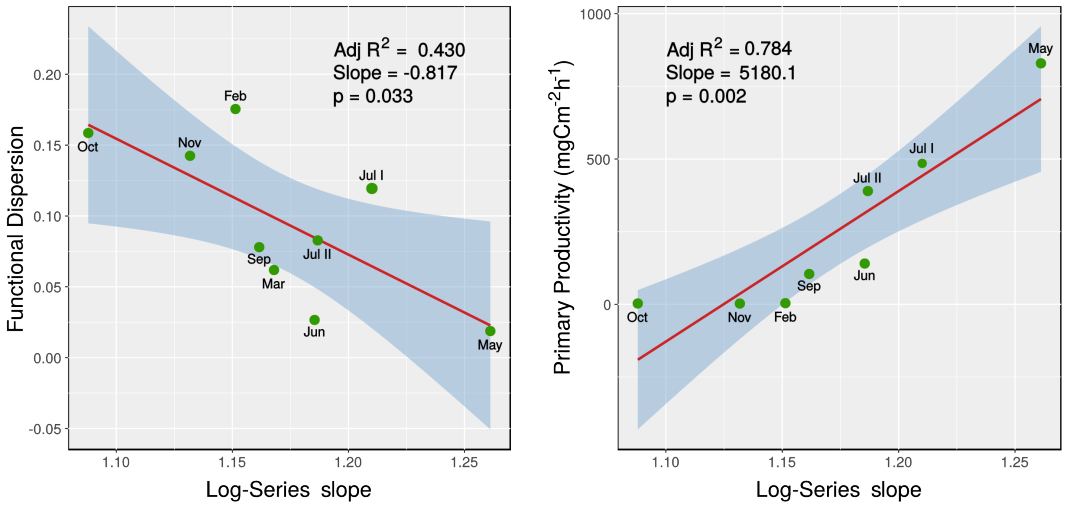


Figura 4: **As desviacións na compoñente β da log-series.** Os momentos de outubro e maio representan casos críticos na sucesión ecolóxica do microplancto mariño da ría de Vigo. En maio é o momento de maior afloramento de primavera, cando son as condicións máis favorables para o crecemento das poboacións. En outubro a mestura vertical da auga é o paso aos momentos desfavorables de crecemento. Ambos son momentos opostos da produción primaria e de espazos funcionais, ou nichos, ocupados. Os cambios na compoñente β da log-series responden a estes dous factores.

Do mesmo xeito, obsevouse que durante os momentos favorables e de alta produtividade primaria a desviación de β foi maior e a Dispersión Funcional menor. Isto revela un proceso de exclusión competitiva dun reducido grupo de especies, similares entre si, que teñen unha mellor capacidade de asimilación de nutrientes e maior posibilidade de crecemento poboacional. Pola contra, nos momentos desfavorables, cando a Dispersión Funcional foi a máxima, a desviación de β estivo no mínimo, preto de $\beta = 1$. Isto suxírenos que o aproveitamento de novos nichos é a resposta que ten a comunidade para conseguir persistir. Como entre os múltiples procesos de interacción entre especies non prevalece ningún, non hai ningún mecanismo que se vexa reflectido en desviacións de β.
Así é que, como si acepta a Teoría Idiosincrática para a SAD log-series, os continuos cambios temporais na interacción entre especies e cambios nos papeis de explotación de nichos explican a función e a persistencia dun amplo grupo de especies raras nas comunidades de microplancto mariño.

Referencias

[1] Baselga A. Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*. 2010;19: 134–143.

[2] Baselga A. Separating the two components of abundance-based dissimilarity: balanced changes in abundance vs. abundance gradients. *Methods in Ecology and Evolution*. 2013;4: 552–557.

[3] Botta-Dukát Z, Wilson JB. Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of Vegetation Science*. 2005;16: 533–540.

[4] Hubbell SP. *Neutral Theory*. In: Editors-in-Chief: Sven Erik Jorgensen, Brian Fath, editors. Encyclopedia of Ecology. Oxford: Academic Press; 2008. pp. 2504–2511.

[5] Hubbell SP. Neutral Theory and the Evolution of Ecological Equivalence. *Ecology*. 2006;87: 1387–1398.

[6] Koleff P, Gaston KJ, Lennon JJ. Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*. 2003;72: 367–382.

[7] Laliberte E, Legendre P. A distance-based framework for measuring functional diversity from multiple traits. *Ecology*. 2010;91: 299–305.

[8] Pueyo S. Diversity: between neutrality and structure. *Oikos*. 2006;112: 392–405.

[9] Pueyo S, He F, Zillio T. The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol Lett*. 2007;10: 1017–1028.

[10] Villéger S, Mouillot D. Additive partitioning of diversity including species differences: a comment on Hardy & Senterre (2007). *Journal of Ecology*. 2008;96: 845–848.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

UN TEST NO PARAMÉTRICO PARA COMPARAR EL PATRÓN ESPACIAL DE LOS INCENDIOS FORESTALES REGISTRADOS EN GALICIA

Isabel Fuentes-Santos [1], Wenceslao González-Manteiga [2], Jorge Mateu [3] y Manuel F. Marey-Pérez [4]

[1]Instituto de Investigaciones Marinas - Consejo Superior de Investigaciones Científicas. *isafusa@gmail.com*
[2]Departamento de Estatística e Investigación Operativa - Universidade de Santiago de Compostela. *wenceslao-gonzalez@usc.es*
[3]Departamento de Matemáticas - Universitat Jaume I. *mateu@mat.uji.es*
[4]Departamento de Enxeñería Agroforestal - Universidade de Santiago de Compostela. *manuel.marey@usc.es*

RESUMEN

En los último años ha aumentado el uso de técnicas de análisis e inferencia de procesos puntuales espaciales en el estudio del comportamiento de los incendios forestales. Este trabajo propone un test no parámetrico que compara las densidades de localización de eventos de dos procesos puntuales para determinar si tienen la misma estructura de primer orden. Este test permitirá determinar, por ejemplo, si la distribución espacial de los incendios forestales registrados en Galicia depende de su causa o varia a medida que pasa el tiempo.

Palabras e frases clave: bootstrap suavizado, estimador núcleo, función de intensidad, incendios intencionados, proceso puntual espacial.

1. INTRODUCCIÓN

Los incendios forestales suponen un gran problema ambiental en todo el mundo. En particular, en Galicia los incendios intencionados representan más del 80 % del total y son la principal causa de destrucción de los bosques. Este situación ha motivado el uso de distintas técnicas estadísticas para estudiar el comportamiento de los incendios forestales. Entre ellas, los procesos puntuales espaciales permiten describir la distribución espacial de los incendios en una región y caracterizar tanto la interacción ente incendios como su dependencia de variables ambientales o socioeconómicas.

Un proceso puntual espacial es un proceso estocástico que genera un número finito de eventos $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ en una región $W \subset \mathbb{R}^2$. Si los eventos tienen asociadas medidas o marcas tenemos un proceso puntual con marcas. Marcas discretas que dividen el proceso en varios grupos dan lugar a un proceso puntual multitipo. Por ejemplo, si conocemos la localización espacial de los incendios registrados en una región junto con otras característica de interés, como causa del incendio o superficie quemada, tenemos una realización de un proceso puntual espacial con marcas.

Cuando analizamos un procesos puntual multitipo es natural preguntarse si dos tipos de eventos tienen la misma distribución espacial. En nuestro caso podemos preguntarnos si la distribución espacial de incendios intencionados y con causa natural es la misma. Sin embargo, hasta el momento el análisis de procesos puntuales muttitipo se ha centrado en sus características de segundo orden, es decir en contrastar si existe interacción espacial entre dos tipos de eventos.

Si dos procesos puntuales, $\mathbf{X}_1$ y $\mathbf{X}_2$, tienen la misma distribución espacial, sus intensidades de primer orden son proporcionales y sus densidades de localización de eventos son iguales. Por tanto, podemos comparar la distribución espacial de dos procesos puntuales mediante un test que contraste la igualdad de sus densidades de localización de eventos. La consistencia del estimador núcleo de la densidad de localización de eventos y la equivalencia entre esta función y la función de densidad bivariante [1] sugieren extender al ámbito de los procesos puntuales el test desarrollado por [2] para comparar distribuciones multivariantes.

2. EL ESTADÍSTICO DE CONTRASTE

Condicionados al número de eventos, $N_1 = n_1$ y $N_2 = n_2$, los patrones puntuales espaciales pueden verse como muestras aleatorias de las distribuciones bivariantes con densidades $\lambda_{01}(\cdot)$ y $\lambda_{02}(\cdot)$. Entonces, podemos contrastar la hipótesis nula, $\mathcal{H}_0 : \lambda_{01}(x) = \lambda_{02}(x)$, mediante el siguiente estadístico

$$\hat{T} = \int_W \left(\hat{\lambda}_{01}(x) - \hat{\lambda}_{02}(x)\right)^2 dx = \hat{\psi}_{1,G_1} + \hat{\psi}_{2,G_2} - \left(\hat{\psi}_{12,G_1} + \hat{\psi}_{21,G_2}\right) \tag{1}$$

donde $\hat{\psi}_i$ y $\hat{\psi}_{ij}$ son los estimadores núcleo de $\psi_i = \int_W \lambda_{0i}(x)^2 dx$ y $\psi_{ij} = \int_W \lambda_{0i}(x)\lambda_{0j}(x) dx$, para $j = 1, 2$. G_1, G_2 son las matrices de suavizado de $\hat{\psi}_1$ y $\hat{\psi}_2$, que pueden obtenerse mediante el algoritmo plug-in propuesto por [3].

La mormalidad asintótica de $\hat{T}$ bajo la hipótesis nula, probada por [2] para distribuciones multivariantes, se puede extender de forma natural al caso de procesos inhomogéneos de Poisson. Sin embargo la calibración asintotica del test puede no ser adecuada para patrones espaciales con pocos eventos y, por otra parte, no podemos garantizar la normalidad de $\hat{T}$ si los procesos no son de Poisson. Por tanto, proponemos un algoritmo de bootstrap suavizado que genera réplicas de la hipótesis nula para calibrar el estadístico de contraste.

Hemos realizado un estudio de simulación, que contempla distintas formas y grados de desviación de la hipótesis nula, para analizar la validez del test. Los resultados de este estudio muestran que algoritmo bootstrap proporciona una mejor calibración y es más potente que la calibración asintotica, que puede lugar a un test conservador. El test detecta distintos tipos de desviación de la hipótesis nula y, como es deseable, su potencia aumenta a medida que nos alejamos de $\mathcal{H}_0$. Por último, la potencia del test se reduce cuando los procesos puntuales tienen pocos eventos y con diseños no equilibrados, es decir cuando el número de eventos de los dos procesos puntuales que se están comparando es muy distinto.

3. COMPARACIÓN DEL PATRÓN ESPACIAL DE DISTINTOS TIPOS DE INCENDIOS

Hemos aplicado el test (1), calibrado mediante $B = 1000$ realizaciones del algoritmo bootstrap, para contrastar si la distribución espacial de los 6904 incendios forestales registrados en Galicia en el año 2006 depende de su causa o varía a medida que pasa el tiempo.

La Figura 1 muestra el estimador núcleo de la intensidad para los incendios clasificados por causa. La comparación 2 a 2 entre causas ha confirmado que sus distribuciones espaciales son distintas (p-valor $< 0{,}001$). Para analizar la variabilidad temporal de la distribución espacial de los incendios forestales, hemos clasificado los incendios registrados entre enero y septiembre por meses (ver estimadores de intensidad en la Figura 2). Las comparaciones 2 a 2 solo identifica como iguales las distribuciones espaciales de abril y mayo (p-valor = 0.09).

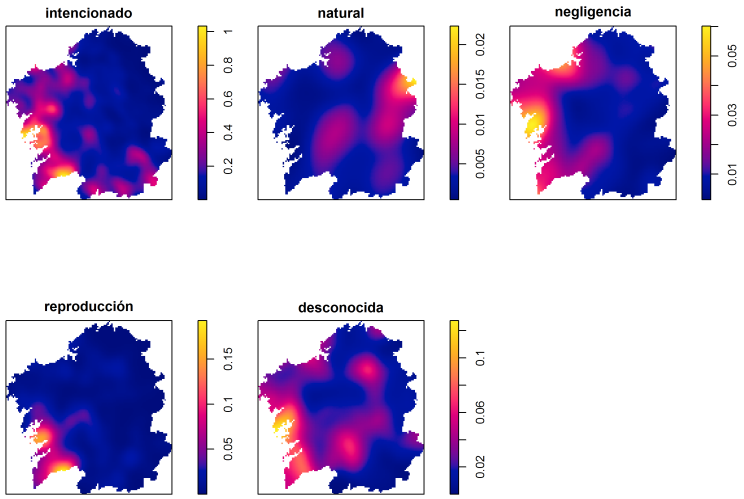


Figura 1: Estimador núcleo de la intensidad de primer orden de los incendios registrados en Galicia por causa.

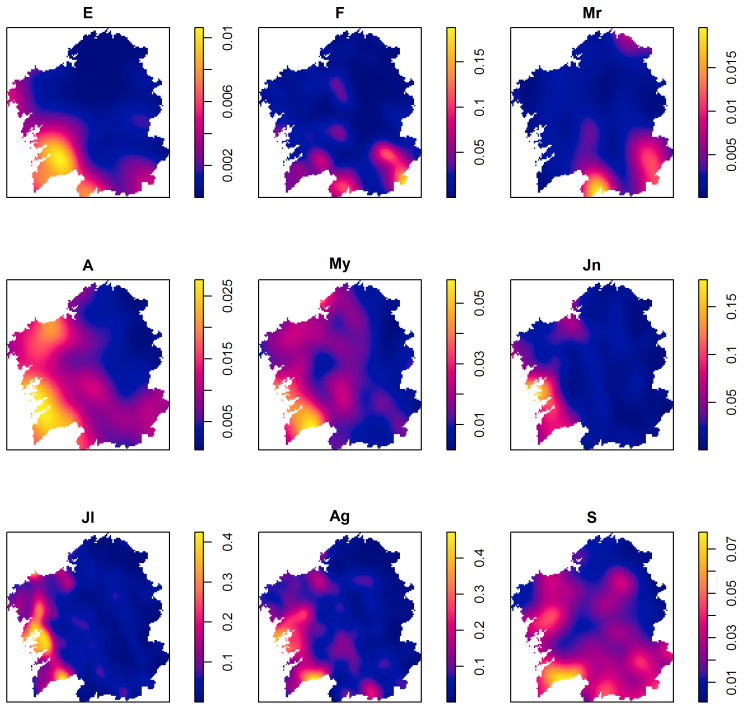


Figura 2: Estimador núcleo de la intensidad de primer orden de los incendios registrados en Galicia entre enero y septiembre de 2006.

3. CONCLUSIONES

En este trabajo se propone una medida de discrepancia cuadrática para contrastar la igualdad de las densidades de localización de eventos y, por tanto, establecer cuando dos procesos puntuales espaciales tienen la misma estructura de primer orden. La aplicación de este test al análisis de incendios forestales registrados en Galicia ha confirmado que su distribución espacial de los incendios depende de su causa. Este resultado sugiere la necesidad de utilizar distintos modelos y covariables para tratar de predecir el riesgo de incendios intencionados y naturales. Por último hemos visto que este test puede usarse para contrastar si un proceso espacio-temporal es separable, es decir si su distribución espacial se mantiene constante a medida que pasa el tiempo, cuando estratificamos la componente temporal dando lugar a un proceso multitipo.

AGRADECIMIENTOS

Este trabajo ha sido financiado por los proyectos MTM2008-0310 del Ministerio de Ciencia e Innovación, MTM2010-14961, MTM2013-41383-P y MTM2013-43917-P del Ministerio de Economía y Competitividad, P1-1B2012-52 de la Fundación Bancaja y la red IAP-StUDyS de la Belgian Science Policy.

Referencias

[1] Fuentes-Santos I., González-Manteiga W., and Mateu J. (2015). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial Poisson point processes. *Scandinavian Journal of Statistics*, DOI: 10.1111/sjos.12183.

[2] Duong T., Goud S., and Schauer K. (2012). Closed-form density-based framework for au- tomatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences of the United States of America* 109(22), 8382-8387.

[3] Chacón, J. E., and Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *Test*, 19(2), 375:398.

II Encontro Galaico-Portugues de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Políticas de pesca sustentáveis ótimas com esforço constante *versus* políticas de pesca ótimas com esforço variável: comparação e aplicação em ambiente aleatório com o modelo de Gompertz

Nuno M. Brites[1] e Carlos A. Braumann[1,2]

[1]Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora
[2]Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora

RESUMO

Para descrever o crescimento de uma população sujeita a pesca e a flutuações aleatórias do ambiente podemos usar modelos (ver, por exemplo, [1], [2] e [3]) baseados em equações diferenciais estocásticas (EDE). Neste trabalho consideramos que o crescimento da população evolui de acordo com o modelo de Gompertz ao qual subtraímos uma taxa de capturas da forma $h(t) = qE(t)X(t)$, onde $q > 0$ representa o coeficiente de capturabilidade, $E(t) \geq 0$ é o esforço de pesca e $X(t)$ representa o tamanho da população no instante de tempo t. Obtemos a EDE

$$dX(t) = rX(t) \ln\left(\frac{K}{X(t)}\right) dt - qE(t)X(t)dt + \sigma X(t)dW(t), \quad X(0) = x_0,$$

onde $r > 0$ é a taxa intrínseca de crescimento da população, $K > 0$ é a capacidade de sustento do meio, $\sigma > 0$ mede a intensidade das flutuações aleatórias do ambiente, $W(t)$ é um processo de Wiener e $x_0 > 0$ é a população no instante inicial, cujo valor é conhecido.

Em [3] otimizou-se a taxa de captura para os modelos logístico e de Gompertz; [5] fez um tratamento semelhante para o modelo de Gompertz. Aqui vamos, em vez disso, otimizar o lucro por unidade de tempo, que supomos ser da forma $\Pi(t) = P(t) - C(t)$, onde $P(t) = ph(t)$ é o preço de venda (sendo $p > 0$ o preço unitário) e $C(t) = c(E)E(t)$ são os custos de pesca. O custo por unidade de esforço $c(E)$ pode variar com o esforço aplicado; aqui vamos considerar o caso particular de $c(E) = c_1 + c_2E(t)$, com $c_1 > 0$ e $c_2 > 0$.

Existem trabalhos anteriores relacionados com a escolha da política ótima de pesca com o objetivo de maximizar o lucro acumulado (descontado por uma taxa de desvalorização) ao longo de um horizonte de tempo finito (ver, por exemplo, [4]). As políticas ótimas conduzem a um esforço de pesca variável $E^*(t)$ que, sob certas condições, podem ser do tipo 'arranca-pára', isto é, alternam entre períodos de pesca e períodos de ausência de pesca. O esforço ótimo varia de acordo com os valores (que variam aleatoriamente) do tamanho da população. Este tipo de políticas pode ser aplicada, por exemplo, a ativos financeiros, que dada a sua natureza podem ser quantificados quase continuamente, mas não são aplicáveis a populações sujeitas a pesca. Com efeito, a estimação do tamanho de uma população é difícil, dispendiosa e demorada, e a logística

da atividade pesqueira não é compatível (quer do ponto vista prático quer do ponto de vista social) com alterações muito frequentes determinadas por variações aleatórias do esforço de pesca.

Uma abordagem alternativa foi proposta em [3] com base em políticas de pesca sustentáveis e aplicáveis com esforço constante ($E(t) \equiv E$), que também conduzem à sustentabilidade da população e à existência de distribuição estacionária para o tamanho da população. Em [3] obteve-se o esforço E que otimiza a taxa de captura para os modelos logístico e de Gompertz em regime estacionário, conduzindo a políticas sustentáveis ótimas de pesca.

Neste trabalho, determinamos a política de pesca de esforço constante que, em vez da taxa de captura, otimiza o lucro sustentável (também constante) esperado por unidade de tempo. Sejam E^{**} o esforço ótimo de pesca e Π^{**} a respetiva taxa de lucro ótima. Com recurso a simulações de Monte Carlo comparamos as duas metodologias, ou seja, comparamos os valores de E^{**} com $E^*(t)$ e Π^{**} com o lucro ótimo variável esperado por unidade de tempo $\Pi^*(t)$ correspondente a $E^*(t)$. Podemos verificar que, em situações típicas, há uma ligeira redução do lucro quando se escolhe esta nova política em vez da política inaplicável com esforço variável.

Terminamos com a comparação dos resultados obtidos pelas duas abordagens quando são utilizado os cálculos de Itô e de Stratonovich e mostramos que, desde que sejam tomadas as devidas precauções, não existem diferenças entre os dois cálculos.

Palavras e frases chave: Equações diferenciais estocásticas, Esforço de pesca, Modelo de Gompertz, Políticas ótimas, Políticas sustentáveis.

AGRADECIMENTOS

Nuno M. Brites e Carlos A. Braumann são membros do Centro de Investigação em Matemática e Aplicações (UID/MAT/04674/2013), financiado pela Fundação para a Ciência e Tecnologia (FCT). O primeiro autor é financiado pela FCT através de uma bolsa de doutoramento com a referência SFRH/BD/85096/2012.

Referências

[1] Beddington, J. R. and May, R. M. (1977). Harvesting natural populations in a randomly fluctuating environment. *Science*. 197 (4302), pp. 463–465.

[2] Braumann, C. A. (2002). Variable effort harvesting in random environments: generalization to density-dependent noise intensities. *Math.Biosciences*. 177 & 178, pp. 229–245.

[3] Braumann, C. A. (1985). Stochastic differential equation models of fisheries in an uncertain world: extinction probabilities, optimal fishing effort, and parameter estimation. *In Mathematics in Biology and Medicine (V. Capasso, E. Grosso, and S. L. Paveri-Fontana, editors)*. Springer, Berlin, pp. 201–206.

[4] Suri, R. (2008). Optimal harvesting strategies for fisheries: A differential equations approach. PhD thesis, Massey University, Albany, New Zealand.

[5] Zou, Z., Li, W., and Wang, K. (2013). Ergodic method on optimal harvesting for a stochastic Gompertz-type diffusion process. *Applied Mathematics Letters*. 26, pp. 170–174.

Comunicacións orais

Sesión III: Modelos de regresión distribucional aditivos estructurados e GAM

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

UTILIDAD DE LOS MODELOS DE REGRESIÓN DISTRIBUCIONAL ADITIVOS ESTRUCTURADOS EN LA TOMA DE DECISIONES CLÍNICAS. A PROPÓSITO DEL POTASIO

Jenifer Espasandín-Domínguez[1],Carmen Cadarso-Suárez[1], Thomas Kneib[2], Balbina Casas-Méndez[3], Alfonso Javier Benitez-Estévez[4], Tegra Barreiro-Martínez[4], Francisco Gude[5].

[1] Grupo de Bioestadística, Centro de Investigación en Medicina Molecular y Enfermedades Crónicas (CiMUS). Universidad de Santiago de Compostela.
[2] Chair of Statistics, Georg-August-Universität Göttingen, Germany.
[3] Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela.
[4] Servicio de Análisis Clínicos. Hospital Clínico Universitario. Santiago de Compostela.
[5]Unidad de Epidemiología Clínica. Hospital Clínico Universitario. Santiago de Compostela.

RESUMEN

Una de las cuestiones que más frecuentemente se plantean los especialistas clínicos ante la interpretación de resultados analíticos se refiere al hecho de si los valores de potasio obtenidos, reflejan los verdaderos niveles que presentan los pacientes o pueden verse alterados por variables como la edad y el sexo o algún otro factor preanalítico como el procedimiento de extracción, el manejo o el transporte de las muestras. Para identificar todas las variables que pueden estar causando alteraciones sobre la media y la varianza de los niveles de potasio se utilizarán modelos de regresión distribucional aditivos estructurados. Estos modelos nos permitirán realizar una modelización no paramétrica de los parámetros de la distribución de la variable respuesta e incorporar efectos espaciales.

Palabras y frases clave: potasio, regresión distribucional, efectos espaciales, p-splines.

1. INTRODUCCIÓN

El potasio es un catión intracelular cuya concentración se determina con frecuencia en los laboratorios debido a que es ampliamente monitorizado en pacientes diabéticos, cardiópatas o con insuficiencia renal. Entre las causas más frecuentes de aumento de la concentración de potasio, figura la hemólisis, la cual se asocia a diversos factores preanalíticos, tales como el procedimiento de extracción, transporte, conservación y/o pretratamiento de las muestras.
En los últimos años, estamos asistiendo a una descentralización de los servicios sanitarios para lograr un acercamiento de la atención sanitaria a los ciudadanos. En el ámbito de los laboratorios clínicos se ha procedido a la descentralización en la recolección de las muestras biológicas a través de los centros de salud. El alejamiento de los puntos de extracción ha supuesto un cambio en los factores preanalíticos anteriormente mencionados que pueden cuestionar la veracidad de los resultados analíticos obtenidos.
Recientemente, tanto los profesionales médicos como los responsables de laboratorio de un hospital terciario, han mostrado su preocupación por el elevado porcentaje de pacientes cuyos niveles de potasio se sitúan fuera de los intervalos de referencia, [1]. El objetivo de este trabajo es analizar los factores que pueden influir sobre las concentraciones séricas de potasio teniendo en cuenta que el Área Sanitaria sujeta al estudio cuenta con numerosos puntos de extracción periféricos. Pero además

de determinar estas concentraciones medias de potasio, los clínicos están interesados en conocer su variabilidad.
Para abordar este problema clínico se emplearán modelos de regresión distribucional aditivos estructurados (en adelante modelos de regresión distribucional) de localización y escala para la media y varianza de los niveles de potasio, [2].
La mayoría de los modelos de regresión clásica se centran únicamente en la estimación de la media, sin embargo, en muchas ocasiones, no es suficiente para capturar la verdadera naturaleza de los datos. La ventaja que nos ofrecen este tipo de modelos es la consideración de prácticamente cualquier tipo de distribución para la variable respuesta además de la modelización conjunta de todos sus parámetros. Por otro lado, nos permitirán estimar de forma suave covariables continuas, efectos aleatorios, posibles tendencias espaciales ó variables categóricas entre otros efectos.

2. MATERIAL Y MÉTODO

2.1. Descripción de la base de datos

A partir del sistema de información del Laboratorio de Análisis Clínicos del Hospital Clínico Universitario de Santiago de Compostela, se recolectaron todas las analíticas realizadas en el período comprendido entre el 1 de junio y el 31 de diciembre de 2015 correspondiente a las determinaciones séricas de potasio, sodio y creatinina. Después de excluir todas aquellas muestras que presentaran criterios de hemólisis o valores de creatinina y/o sodio fuera de los rango de referencia, el número de muestras incluidas fue de 95096. Mencionar que el transporte de muestras está subcontratado a una empresa externa, y que ésta realiza diferentes rutas en el traslado de las muestras biológicas desde los puntos de extracción hasta su llegada al Hospital. Se determinó el *tiempo* hasta el procesamiento como el transcurrido desde las 8:00h hasta su registro en el sistema robótico Aptio (Siemens, Erlangen, Alemania) del laboratorio.

2.2 Modelos de regresión distribucional

En los últimos años los modelos aditivos estructurados, conocidos como modelos STAR (Structured Additive Regression) están mereciendo gran interés pues permiten generalizar a los modelos GLM [3] y GAM [4], además de incorporar efectos suaves de las covariables (empleando splines penalizados), efectos aleatorios, datos clúster, o efectos espacio-temporales, entre otros [5]. Sin embargo, en muchas aplicaciones médicas, como en este caso, no solo es interesante explicar el efecto de las covariables en función de la media de la respuesta, sino conocer su distribución completa. Por este motivo recurriremos a la regresión distribucional considerando modelos STAR como predictores.
En la regresión distribucional se supone para cada uno de los parámetros un predictor semiparamétrico, que en este caso particular toman la forma, [2]:

$$\begin{cases} \eta^{\mu} = \beta_0^{\mu} + v'\beta^{\mu} + f_1^{\mu}(v_1) + \cdots + f_q^{\mu}(v_q) + f_{spat}^{\mu}(s), \\ \eta^{\sigma^2} = \beta_0^{\sigma^2} + v'\beta^{\sigma^2} + f_1^{\sigma^2}(v_1) + \cdots + f_q^{\sigma^2}(v_q) + f_{spat}^{\sigma^2}(s), \end{cases} \tag{1}$$

donde η es una variable transformada del parámetro de la respuesta original (el potasio); $v'\beta$ denota los efectos paramétricos β de covariables, v; $f_i, i = 1, \ldots, q$ son funciones desconocidas suaves que nos permiten modelar efectos no lineales de las covariables continuas. $f_{spat}(s)$, representa los efectos espaciales. En este caso, incluiremos en el estudio únicamente efectos espaciales no estructurados, para captar posibles tendencias locales en los municipios del Área Sanitaria de Santiago de Compostela, debido a la existencia de diferentes rutas para municipios próximos. Tal y como reflejan los superíndices, la expresión (1) incluye una ecuación correspondiente a la media y una correspondiente a la varianza, en las que además las covariables consideradas para cada uno de estos parámetros pueden ser diferentes.
La inferencia de los modelos regresión distribucional se realiza mediante métodos puramente Bayesianos, (Full Bayes, FB) empleando técnicas de simulación MCMC (*Markov Chain Monte Carlo*),

[2]. En este contexto, los efectos no lineales de las covariables continuas, se modelarán mediante versiones bayesianas de los splines penalizados (P-splines; [5]), mientras que para los efectos espaciales no estructurados supondremos que siguen una distribución a priori gaussiana.

2.3 Resultados

En primer lugar, se ha comprobado que el potasio tiene una densidad aproximadamente normal. Como posibles covariables a introducir en la modelización (1) se han considerado la *edad* y el *sexo* del paciente, el tiempo transcurrido hasta el procesamiento de las muestras (*tiempo*) y el efecto espacial de los municipios (*muni*) pertenecientes al Área Sanitaria de Santiago de Compostela, donde se obtuvieron las muestras.

Se ajustaron varios modelos y los resultados se presentan en la Tabla 1 (*M1-M6)*, en términos del criterio DIC (*Deviance Information Criterion*) [6]. Según este criterio, el mejor resultado se obtiene con el "*Modelo M5*". En la Tabla 2, y las Figuras 1 y 2 se muestran los resultados para este modelo.

Modelo	***Especificación del modelo***	***DIC***
M1	$\eta_i^\mu = \beta_0^\mu + sexo_i\beta_1^\mu + f_2^\mu(edad)_i,$ $\eta_i^{\sigma^2} = \beta_0^{\sigma^2} + sexo_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(edad)_i.$	90247
M2	$\eta_i^\mu = \beta_0^\mu + sexo_i\beta_1^\mu + f_2^\mu(edad)_i + f_{spat}^\mu(muni)_i,$ $\eta_i^{\sigma^2} = \beta_0^{\sigma^2} + sexo_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(edad)_i + f_{spat}^{\sigma^2}(muni)_i.$	87936.3
M3	$\eta_i^\mu = \beta_0^\mu + sexo_i\beta_1^\mu + f_2^\mu(edad)_i + f_3^\mu(tiempo)_i,$ $\eta_i^{\sigma^2} = \beta_0^{\sigma^2} + sexo_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(edad)_i + f_3^{\sigma^2}(tiempo)_i.$	89407.4
M4	$\eta_i^\mu = \beta_0^\mu + sexo_i\beta_1^\mu + f_2^\mu(edad)_i + tiempo_i\beta_3^\mu,$ $\eta_i^{\sigma^2} = \beta_0^{\sigma^2} + sexo_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(edad)_i + tiempo_i\beta_3^{\sigma^2}.$	89535.9
M5	$\eta_i^\mu = \beta_0^\mu + sexo_i\beta_1^\mu + f_2^\mu(edad)_i + f_3^\mu(tiempo)_i + f_{spat}^\mu(muni)_i,$ $\eta_i^{\sigma^2} = \beta_0^{\sigma^2} + sexo_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(edad)_i + f_3^{\sigma^2}(tiempo)_i + f_{spat}^{\sigma^2}(muni)_i.$	**87823.9**
M6	$\eta_i^\mu = \beta_0^\mu + sexo_i\beta_1^\mu + f_2^\mu(edad)_i + tiempo_i\beta_3^\mu + f_{spat}^\mu(muni)_i,$ $\eta_i^{\sigma^2} = \beta_0^{\sigma^2} + sexo_i\beta_1^{\sigma^2} + f_2^{\sigma^2}(edad)_i + tiempo_i\beta_3\sigma^2 + f_{spat}^{\sigma^2}(muni)_i.$	87872.9

Tabla 1: DIC para cada uno de los modelos de regresión distribucional estimados.

Se ha encontrado que la edad, el tiempo hasta el procesamiento de las muestras y su procedencia son variables asociadas de forma estadísticamente significativa con los niveles de potasio (Ver Tabla 2 y Figura 1). Sin embargo, a pesar de que en media el nivel de potasio y su varianza es menor en las mujeres que en los hombres, estas diferencias no tienen relevancia clínica. En contraposición, las diferencias obtenidas por edades en los niveles de potasio sí se consideran relevantes y ponen de manifiesto la necesidad de establecer distintos niveles de referencia por rangos de edades. Además, la varianza en los jóvenes es pequeña, aumentando aproximadamente a partir de los 45 años (Figura 1). Por otra parte, aunque el efecto del tiempo hasta el procesamiento es relevante, parece observarse un efecto más acusado en la parte espacial. La incorporación de los efectos espaciales no estructurados nos permite comprobar que los niveles medios de potasio no se distribuyen de forma uniforme en el Área Sanitaria de Santiago de Compostela. En general, se detectaron valores mayores de potasio en los municipios más alejados del laboratorio, si bien cabe mencionar la existencia de algún ayuntamiento cercano con valores aún más elevados (como el caso de Trazo). Finalmente, se detectó gran variabilidad en algunos municipios, que no siempre coincide con aquellos dónde se registraron mayores valores de potasio, (Figura 2).

Parámetro	Media	Cuantil 2.5 % ($q_{2,5}$)	Mediana (q_{50})	Cuantil 97.5 % ($q_{97,5}$)
β_0^μ *(intercepto)*	4.70862	4.68185	4.70907	4.73399
$\beta_0^{\sigma^2}$ *(intercepto)*	-1.79722	-1.85615	-1.79543	-1.74459
β_1^μ *(sexo)*	-0.03105	-0.03618	-0.03093	-0.02621
$\beta_1^{\sigma^2}$ *(sexo)*	-0.05230	-0.06945	-0.05281	-0.03362

Tabla 2: Estimación de los efectos lineales (*Modelo M5*).

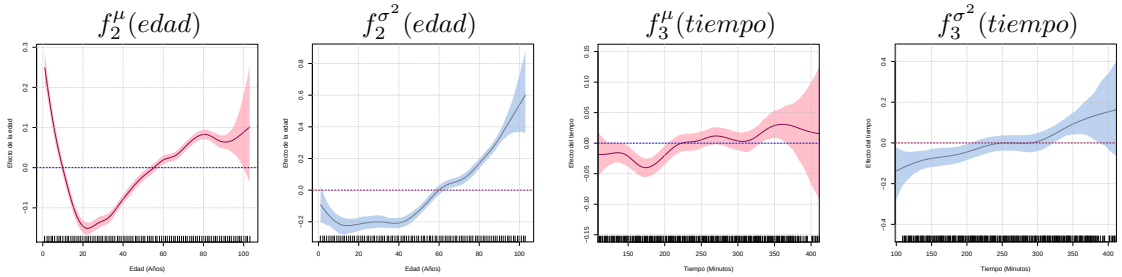


Figura 1: Efectos de las variables *edad* y *tiempo* en la media y varianza de las muestras de potasio obtenidos con el *Modelo M5*.

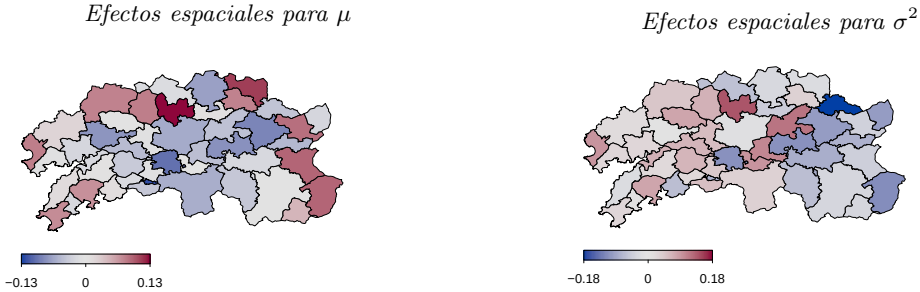


Figura 2: Efectos espaciales no estructurados.

3. CONCLUSIONES

La adecuación de los modelos de regresión distribucional a nuestro problema, nos ha permitido comprobar que la edad de los pacientes junto con el tiempo hasta el procesamiento de las muestras son variables que influyen en los niveles de potasio. Se identificaron algunos municipios donde puede haber un porcentaje de individuos con valores más altos de potasio que puede llegar a ser relevante en el ámbito clínico. Tanto en municipios con valores bajos como altos de potasio se ha observado alta variabilidad.

A la vista de los resultados, la línea futura es explorar las ineficiencias del sistema de rutas de las muestras y analizar su efecto en las concentraciones de potasio observadas. Para ello, se plantea aplicar técnicas de análisis de redes de transporte y de problemas de colas. Prestando especial atención a las muestras recogidas fuera del laboratorio y sin la supervisión directa del personal del mismo, se espera que un manejo más eficiente de los recursos permita una reducción en los errores y un diseño ajustado de los intervalos de referencia.

AGRADECIMIENTOS

Trabajo realizado en el marco del proyecto MTM2014-52975-C2-1-R, cofinanciado por el Ministerio Español de Economía y Competitividad y a través de fondos FEDER además de por la Xunta de Galicia, mediante la concesión de una beca predoctoral (Plan I2C).

Referencias

[1] CLSI. (2008). *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory Guideline- Third Edition.* CLSI document EP28-A3c. Wayne P.A. Clinical and Laboratory Institute.

[2] Klein N., Kneib T., Klasen S. and Lang S. (2014). Bayesian Structured Additive Distributional Regression for Multivariate Responses.*Journal of the Royal Statistical Society: Series C.* 64, 569–591.

[3] McCullagh P., Nelder J.A. (1989). *Generalized Linear Models.* Chapman-Hall. London.

[4] Hastie T.J., Tibshirani R.J. (1990). *Generalized Additive Models.* Chapman-Hall.Boca Raton/New York.

[5] Fahrmeir L., Kneib T., Lang S., Marx B. (2013). *Regression. Models, Methods and Applications.* Springer. Heidelberg/ Berlin.

[6] Spiegelhalter D.J., Best N.J., Carlin B.P., Van Der Linde A. (2002). *Bayesian measures of model complexity and fit.* Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(4), 583–639.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Comparing higher-dimensional tensor product and markov random fields when modeling a trivariate interaction term in a structured additive predictor using BayesX

Elisa Duarte[1], Bruno de Sousa[2], Carmen Cadarso-Suárez[1], Thomas Kneib[3] and Vítor Rodrigues[4]

[1]Unit of Biostatistics, Department of Statistics and Operations Research, School of Medicine, University of Santiago de Compostela, C/ San Francisco s/n, 15782-Santiago de Compostela, Spain
[2]Faculty of Psychology and Education Sciences, University of Coimbra, Rua do Colégio Novo, Apartado 6153, 3001-802 Coimbra, Portugal
[3]Institute of Statistics and Econometrics, Dept. of Economics, Georg-August-Universität Göttingen, Humboldtallee 3, 37073 Göttingen, Germany
[4]Faculty of Medicine, University of Coimbra, Rua Larga, 3004-504 Coimbra, Portugal

Abstract

A structured additive predictor allows the regression modeling of different type of covariates effects, such as, nonlinear effects of continuous variables, correlated spatial effects, time trends, cluster-specific heterogeneity and complex interactions between covariates of different type. BayesX software, version 3.0.2, provides regressions tools based either on Markov chain Monte Carlo (MCMC) simulation techniques, yielding fully Bayesian posterior mean or posterior mode estimates, or on a mixed model representation of penalized regression models with inference based on penalized maximum likelihood and marginal likelihood estimation (empirical Bayes). In this work, the inclusion of a trivariate interaction term in a structured additive predictor is proposed, together with several continuous covariates and spatial effects, when modeling the risk of breast cancer screening data. A comparison between higher-dimensional tensor product splines and two different approaches based on Markov random field methods is presented when modeling the trivariate interaction term using full Bayes inference.

Keywords: STAR models; Higher-order interaction term; Higher-dimensional tensor product splines; Markov random fields; BayesX software.

1. INTRODUCTION

The work presented by Duarte et al.(2016) [1] shows the importance of including a trivariate interaction term between attendance, detection and mortality rates when modeling breast cancer risk using a screening program database. The inclusion of this term allowed for a greater understanding of the complex interaction effects resulting from the diversity in municipalities' coverage and attendance rates. In this work [1], the trivariate interaction implemented based on a Markov random field, introduced to model breast cancer incidence rates, showed a significant improvement in terms of model fit when compared to a simpler geoadditive regression model. An alternative

approach in order to include a higher-order interaction term in a structured additive predictor is the higher-dimensional tensor product splines method. When the inference is carried out using the BayesX software version 3.0.2[5], this feature is only possible through a full bayesian regression tool provided by the software. In the present work, two models with the same predictor and using the MCMC simulation techniques in a bayesian perspective[2, 3] are presented. In one of them the trivariate term is estimated using a Markov random field [8] and in the other a restricted maximum likelihood (REML)[6] estimation is used. The third model in comparison will estimate the triavariate term using higher-dimensional tensor product splines [7, 9].

2. SCREENING DATA AND MODEL FORMULATION

The data set for this study is provided by the Portuguese Cancer League (LPCC/NRC – *Liga Portuguesa Contra o Cancro/Núcleo Regional do Centro*) and consists of information on all women who attended the Breast Cancer Screening Program in central Portugal between 1990 and 2009, born after 1920 and with screening ages between 44 and 70 years old. The covariates considered in our study are: year of birth, age at menarche and menopause, family history, pregnancy status, nursing status, the use of oral contraceptives, the municipality purchasing power index (PPI), and the place of residence (in terms of the municipality where a woman lives). For age at menopause and menarche, we will consider a bivariate interaction effect while a trivariate interaction will be constructed for detection, mortality and attendance rates. The outcome of this study is the breast cancer diagnosis which is a binary variable. There are a total of 134 530 women (79%) who reached menopause. A total of 1 716 women (approximately 1%) have been diagnosed with breast cancer while 132 814 women were disease-free throughout the complete screening period. Thus, a logit model with a logistic distribution $h(\eta) = \frac{\exp(\eta)}{1-\exp(\eta)}$. The predictor to model the probability of having breast cancer can be written as:

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1(famhist1_i) + \beta_2(famhist2_i) + \beta_3(famhist3_i) \\ &\quad +\beta_4(pregnancystatus_i) + \beta_5(oralcontraceptives_i) \\ &\quad +\beta_6(nursingstatus_i) + f_1(birthyear_i) + f_2(PPI_i) \\ &\quad +f_3(menopause_i) + f_4(menarche_i) + f_5(menopause, menarche) \\ &\quad +f_{str}(municipality_i) + f_{unstr}(municipality_i) \\ &\quad +f_{3Di}(attendance, detection, mortality)\end{aligned} \tag{1}$$

where f_{3Di} is the trivariate interaction term, that is modeled using Markov random field in one model (Model MRF-REML and Model MRF-MCMC), and the a 3D tensor product spline in the other model (Model TP-MCMC)

3. CONCLUSIONS

The results achieved for the fixed effects and for the non-linear effects of the continuous covariates, are similar for all models under comparison. Regardless the method used to modeled the trivariate interaction term (Markov random field or 3D tensor product spline), the full bayesian approach does not capture the spatial structured effect, that is observed in the model performed in Duarte et al. (2016) [1] (Fig.1). Worth noting that this model is also unable to capture the bivariate interaction between ages at menopause and menarche.

ACKNOWLEDGMENTS

This work was financed by POPH-QREN, the European Social Fund and national funds MCTES - Portuguese Ministry of Science, Technology and Higher Education through the research project SFRH/BD/64761/2009, the Spanish Ministry of Science and Innovation grant MTM2015-69068-REDT, and the 2014 Competitive Call for "Acciones Conjuntas Hispano-Alemans/PPP Spain"in cooperation with the Foundation for the International Promotion of Spanish Universities – Unversidad.es.

REFERENCES

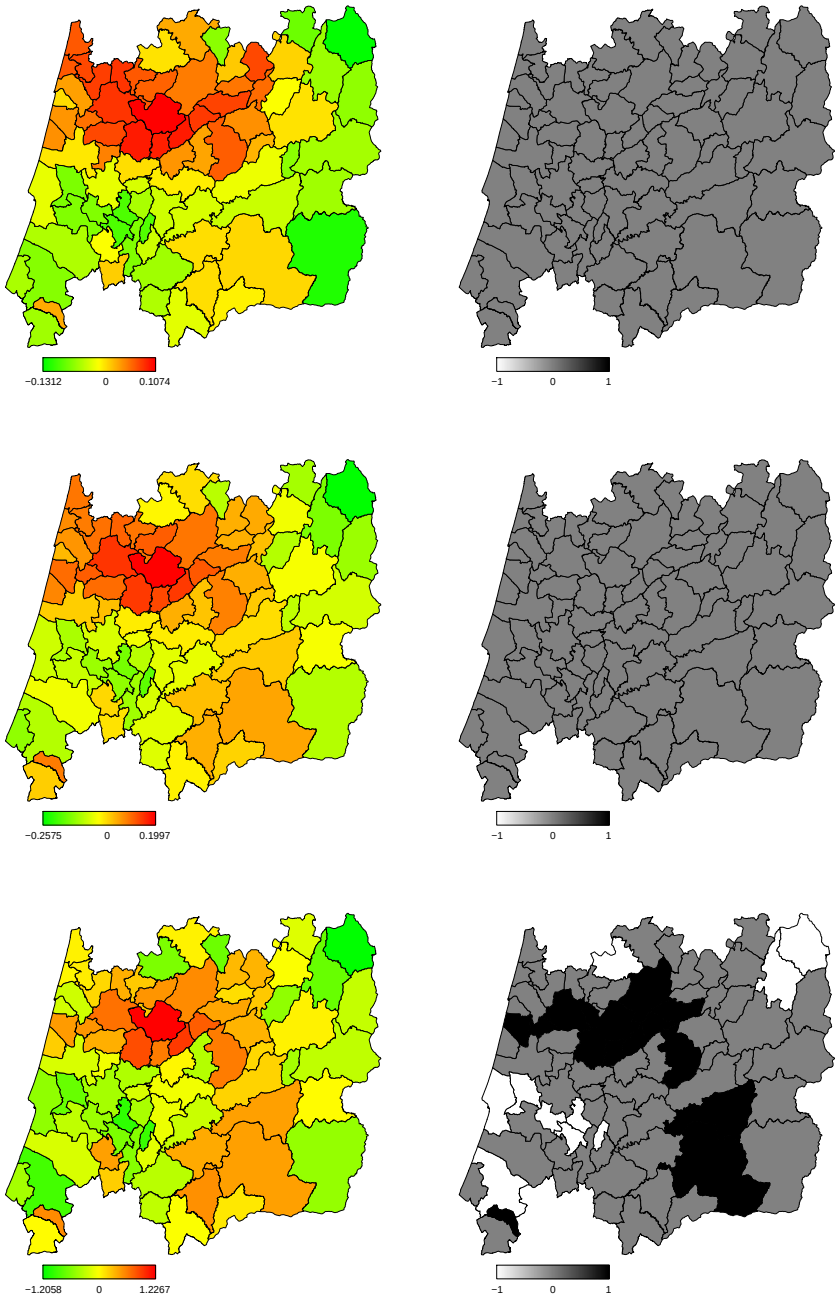


Figura 1: Structured spatial effect (left side), together with the posterior probabilities for a nominal level of 95% credible intervals (right side) - First row: Full bayesian estimation - MRF; Second row: Full bayesian estimation - 3D tensor product spline; third row: REML estimation - MRF.

Referências

[1] E. Duarte, B. de Sousa, C. Cadarso-Suárez, T. Kneib, V. Rodrigues (2016). Exploring risk factors in breast cancer screening program data using structured geoadditive models with high-order interaction. *Computational Statistics & Data Analysis* submitted work.

[2] L. Fahrmeir, S. Lang (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Fields Priors. *Journal of the Royal Statistical Society C* **50**(2), 201–220

[3] L. Fahrmeir, T. Kneib, S. Lang (2004). Penalized structured additive regression for space-time data:

a Bayesian perspective. *Statistica Sinica.* **14**, 731–761.

[4] C. Belitz, A. Brezger, T. Kneib, S. Lang (2009). BayesX - Software for Bayesian Statistics Inference in Structured Additive Regression Models. Version 2.1 *Biometrical Journal* 36, 643-660. url: `http://www.bayesx.org/`

[5] C. Belitz, A. Brezger, T. Kneib, S. Lang and N. Umlauf(2009). BayesX: Software for Bayesian Inference in Structured Additive Regression Models - Version 3.0.2 url: `http://www.bayesx.org/`

[6] T. Kneib (2006). *Mixed model based inference in structured additive regression.* PhD thesis, Fakultät für Mathematik, Informatik und Statistik der Ludwig-Maximilians-Universität München.

[7] L. Fahrmeir and T. Kneib and S. Lang and B. Marx (2013). *Regression: Models, Methods and Applications.* Springer-Verlag Berlin Heidelberg.

[8] H. Rue, L. Held (2005). *Gaussian Markov Random Fields.* Chapman & Hall/CRC - Boca Raton, FL.

[9] S. N. Wood (2006). *Generalized additive models: an introduction with R.* Chapman & Hall/CRC - Boca Raton, FL.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Studying the relationship between reproductive life span and menarche age using Bayesian multivariate structured additive distributional regression

Elisa Duarte[1], Bruno de Sousa[2], Carmen Cadarso-Suárez[1], Nadja Klein[3], Thomas Kneib[3] and Vítor Rodrigues[4]

[1]Unit of Biostatistics, Department of Statistics and Operations Research, School of Medicine, University of Santiago de Compostela, C/ San Francisco s/n, 15782-Santiago de Compostela, Spain
[2]Faculty of Psychology and Education Sciences, University of Coimbra, Rua do Colégio Novo, Apartado 6153, 3001-802 Coimbra, Portugal
[3]Institute of Statistics and Econometrics, Dept. of Economics, Georg-August-Universität Göttingen, Humboldtallee 3, 37073 Göttingen, Germany
[4]Faculty of Medicine, University of Coimbra, Rua Larga, 3004-504 Coimbra, Portugal

Abstract

Age at menarche and menopause trends have been the subject of several studies when studying breast cancer. In general, these studies point out for a downward trend of age at menarche and an upward trend for age at menopause age upwards, of an increase of a woman's reproductive lifespan cycle. In this study a Bayesian multivariate structured additive distributional regression model is used in order to analyze the correlation between a woman's age at menarche and her reproductive lifespan cycle. The data comprise 212 517 post menopause women, born between 1920 and 1965, and who attended the breast cancer screening program over a period of twenty years in the central region of Portugal.

Keywords: Correlated responses; Menarche age; Reproductive lifespan cycle, Breast Cancer Screening data.

ACKNOWLEDGMENTS

This work was financed by POPH-QREN, the European Social Fund and national funds MCTES - Portuguese Ministry of Science, Technology and Higher Education through the research project SFRH/BD/64761/2009 and the Spanish Ministry of Science and Innovation grant MTM2014-52975-C2-1-R.

REFERENCES

Referências

[1] Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **64**, 569–591. doi: 10.1111/rssc.12090

[2] de Sousa, B., Rodrigues, V., Duarte, E. e Cruz, D. (2009). A Correspondence Analysis on First Time Registries in the Breast Cancer Screening Program in Central Portugal. *Joint Statistical Meetings Proceedings, Statistics in Epidemiology Session, August 1-6. Washington, DC, VA: American Statistical Association*, 2234-2245.

[3] Nichols, H.B., Trentham-Dietz, A., Hampton, J.M., Titus-Ernstoff, L., Egan, K.M. , Willett, W.C. and Newcomb, P.A. (2006). From Menarche to Menopause: Trends among US Women Born from 1912 to 1969. *American Journal of Epidemiology* **164**(10), 1003–1011. doi: 10.1093/aje/kwj282

[4] Cabanes, A., Ascunce ,N., Vidal, E., Ederra, M., Barcos, A., Erdozain, N., Lope, V., Pollán, M.(2009). Decline in age at menarche among Spanish women born from 1925 to 1962. *BMC Public Health* **9**, 449.

[5] Padez, C. and Rocha, M. A. (2003). Age at menarche in Coimbra (Portugal) school girls: a note on the secular changes. *Annals of Human Biology* **30**(5), 622-–632.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Empleo de GAMLSS para el cálculo de percentiles de rendimiento en una prueba máxima de 2000 metros sobre ergómetro de remo

Telmo Silva Alonso[1], María del Carmen Iglesias Pérez [1] y José Luis García Soidán[1]

[1] *Facultad de Ciencias de la Educación y del Deporte. Universidad de Vigo (España)*

RESUMEN

El objetivo de este trabajo fue determinar los valores de referencia y construir las curvas de percentiles de rendimiento, para remeros hombres y mujeres entre 14 y 70 años, en una prueba máxima de 2000 metros realizada sobre ergómetro de remo. Para este propósito se realizó un estudio de diseño transversal con una muestra no probabilística consecutiva de remeros entre 14 y 70 años de edad, con peso superior a 75 kg en el caso de los hombres (n=6337) y peso superior a 61,5 kg en el caso de las mujeres (n= 4023). Con el empleo del paquete "gamlss" del software estadístico R se obtuvieron valores de referencia y curvas de percentiles de rendimiento aplicando modelos aditivos generalizados de posición, escala y forma (Rigby & Stasinopoulos, 2005) con una modelización de la distribución de datos de rendimiento por medio de una Box-Cox power exponencial y una técnica de suavización con cubic splines. Como conclusión principal de este trabajo se determinó que el rendimiento en una prueba máxima de 2000 metros en ergómetro de remo se modifica con la edad, contemplando diferentes momentos que se concretan en una fase de crecimiento desde los 14 años, una edad de rendimiento máximo entre los 18 y 25 años y una fase de declinación del rendimiento hasta los 70 años, todas ellas con patrones diferentes según sexo y percentil.

Palabras clave: ergómetro, gamlss, percentiles, referencias, remo, rendimiento

1. INTRODUCCIÓN

El remo es un conjunto de modalidades deportivas en las que un individuo o equipo desplaza un sistema embarcación-remero sobre el agua como efecto de la fuerza de propulsión generada a partir de la repetición de un ciclo de palada (Kleshnev, 2010). No sólo es un deporte de alta competición, sino también una forma de actividad física saludable que puede contribuir de forma substancial al desarrollo del fitness aeróbico y fitness muscular en otro tipo poblaciones (Shephard, 1998).

La estatura el peso y el porcentaje de grasa corporal (Secher, 1983) definen las principales características antropométricas de los remeros de alto nivel. Por otra parte, el $VO_{2\ max}$, la potencia aeróbica máxima y la producción de energía anaeróbica constituyen los factores fisiológicos clave para el rendimiento en pruebas de remo en la distancia de 2000 metros (Nevill, Allen & Ingham, 2011).

Asimismo, desde hace unos años el test de esfuerzo máximo de 2000 metros en ergómetro de remo constituye la prueba de referencia habitual para medir el estado de forma y evaluar el rendimiento en practicantes de remo convencional (Mikulić, Smoljanović, Bojanić, Hannafin, & Matković, 2009).

El objetivo principal de este trabajo fue determinar los valores de referencia y construir las curvas de percentiles de rendimiento en una prueba máxima de 2000 metros realizada sobre ergómetro de remo, para remeros hombres y mujeres entre 14 y 70 años de categoría de peso pesado, con la aplicación de Modelos Aditivos Generalizados de Posición, Escala y Forma. Los percentiles obtenidos no solo tienen interés para la evaluación del rendimiento en el remo de competición sino que, proporcionan referencias indirectas respecto al $VO_{2\ max}$ y la fuerza, muy útiles para la valoración de la condición física desde una perspectiva de la salud.

2. MÉTODOS

Para este trabajo se realizó un diseño transversal a partir de los resultados alcanzados por los sujetos de una muestra no probabilística consecutiva (n=15420), considerando los grupos de estudio según el sexo y el peso en función de las categorías deportivas existentes para el remo indoor[1].

Los datos se obtuvieron de una fuente on-line http://log.concept2.com/ranking.asp, con información disponible desde el año 2002 de resultados en competiciones nacionales e internacionales de remo indoor de participantes entre 12 y 96 años de edad.

En este estudio se analizaron los resultados correspondientes a la categoría de peso pesado, incluyendo en la muestra los hombres con peso superior a 75 kg (n=6337) y a las mujeres con peso superior a 61,5 kg (n= 4023), entre 14 y 70 años de edad y con un resultado oficial en competición de remo indoor conseguido entre el 1 de septiembre de 2010 y 30 de abril de 2014.

Se definió la edad cronológica medida en años como variable explicativa y al tiempo invertido en cubrir los 2000 metros, medido en segundos, como indicador de rendimiento y variable respuesta; además se determinó como posible variable de interacción el entrenamiento previo para la prueba.

En todas las pruebas realizadas fue empleado un ergómetro Concept 2 modelo D (Vogler, Rice & Withers, 2007), con una fiabilidad de 1,3 % como error estándar al medir la potencia media en pruebas de 2000 metros.

Para el cálculo y la obtención de las curvas de percentiles se empleó un ordenador MacBook Pro con procesador Intel Core i5 de 2,4 GHz y el paquete "gamlss" del software estadístico R para aplicar modelos aditivos generalizados de posición, escala y forma (Rigby & Stasinopoulos, 2005) modelizando la distribución de datos de rendimiento mediante una Box-Cox power exponencial, BCPE, y empleando una técnica de cubic splines como método de suavización.

Los GAMLSS[2] se vienen usando en los últimos años en diferentes áreas de investigación para la construcción y cálculo de percentiles (Rigby & Stasinopoulos, 2005). En este sentido y el ámbito de la condición física para la salud destaca el trabajo en el que construyen curvas de percentiles y valores normativos para el test shuttle-run de 20 m en población escolar (Sandercock, Voss, Cohen, Taylor & Stasinopoulos, 2012).

Los modelos generados con GAMLSS son un tipo de modelos de regresión en el que se asume que la variable respuesta, en este caso el tiempo como indicador de rendimiento, proviene de una distribución paramétrica $Y \sim f_Y(y \setminus \theta)$ siendo θ un vector de parámetros desconocidos, que pueden ser modelados como funciones de una o más variables explicativas X, en este caso la edad de los remeros.

El modelo utilizado fue el siguiente. Dado $X = x$,

$$Y \sim BCPE(\mu, \sigma, \nu, \tau)$$
$$g_1(\mu) = h_1(x)$$
$$g_2(\sigma) = h_2(x)$$
$$g_3(\nu) = h_3(x)$$
$$g_4(\tau) = h_4(x),$$

donde, para $k = 1,2,3,4$, $g_k(.)$ son funciones de enlace: $g_1 = g_3 = identidad$, $g_2 = g_4 = log$; $h_k(.)$ son funciones suaves de x, que se estimaron no paramétricamente usando splines cúbicos, y la distribución $Y \sim BCPE(\mu, \sigma, \nu, \tau)$, donde μ es un parámetro de localización, σ un parámetro de escala, ν y τ parámetros de forma, está definida en las ecuaciones (1), (2) y (3) de Rigby & Stasinopoulos (2004). Además, con el fin de mejorar el modelo la variable explicativa fue transformada: $X = edad^{\lambda}$.

Para $i = 1,2, \ldots, n$, dado $X = x_i$, se asume $Y_i \sim BCPE(\mu_i, \sigma_i, \nu_i, \tau_i)$ e independientes. La estimación del modelo que realiza el paquete "gamlss" se basa en maximizar la función de log-verosimilitud penalizada, dada por

[1] Hombre y peso superior a 75 kg, Hombre y peso igual o inferior a 75 kg, Mujer y peso inferior a 61,5 kg y Mujer con peso igual o inferior a 61,5 kg.
[2] Modelos Aditivos Generalizados de Posición, Escala y Forma

$$L_p = L_d - \frac{1}{2}\sum_{k=1}^{4} \lambda_k \int_{-\infty}^{+\infty} \{h_k''(u)\}^2 \, du,$$

donde $L_d = \sum_{i=1}^{n} L_i$, siendo L_i la log-verosimilitud de $Y_i \sim BCPE(\mu_i, \sigma_i, \nu_i, \tau_i)$ y λ_k es el parámetro de suavización del estimador spline de h_k, $k = 1,2,3,4$.

Como criterio para la selección del modelo se usó el $GAIC(3) = -2\widehat{L_p} + 3 * \text{df}$, siendo $\text{df} = \text{df}_\mu + \text{df}_\sigma + \text{df}_\upsilon + \text{df}_\tau + 1$ los grados de libertad del modelo. Cada grado de libertad (p.ej. df_μ) se define como la traza de la matriz de suavizado del spline correspondiente, y tiene una relación directa con su parámetro de suavización (p.ej. λ_1). El último sumando de df corresponde al parámetro λ usado en $X = edad^{\lambda}$.

3. RESULTADOS

El rendimiento para cada edad en las curvas de percentiles de hombres (Figura 1) y mujeres (Figura 2) se expresa en unidades de tiempo (segundos). Cuanto menor es el tiempo invertido en la prueba, mejor será el rendimiento mostrado en ella. En este sentido, los percentiles hacia P2 indican valores con mejor rendimiento frente a los valores de percentiles hacia P98 que representan un rendimiento más bajo.

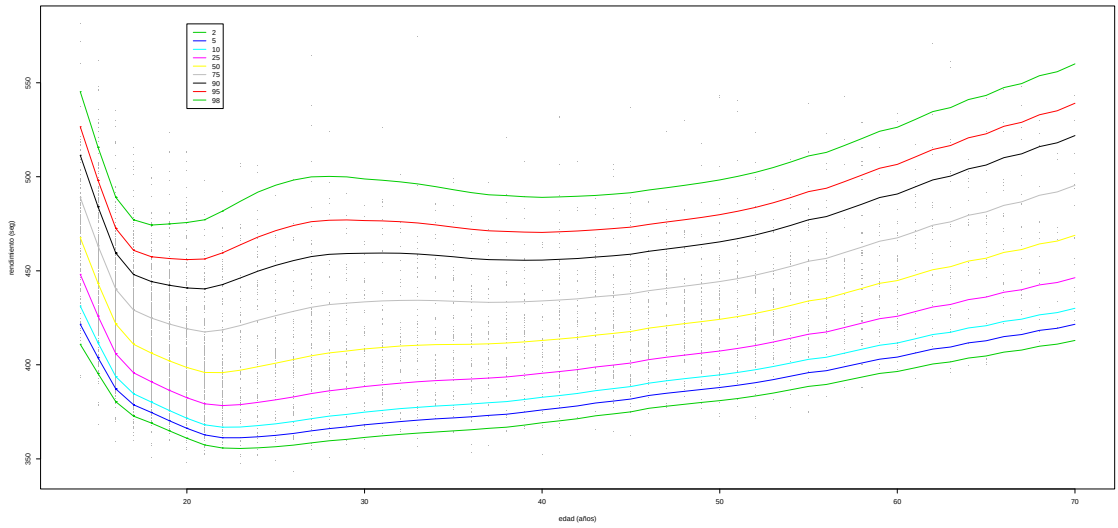


Figura 1. Curvas de percentiles P2, P5, P10, P25, P50, P75, P90, P95 y P98 para el grupo hombre y peso > 75 kg (n=6337) suavizadas con cubic splines.

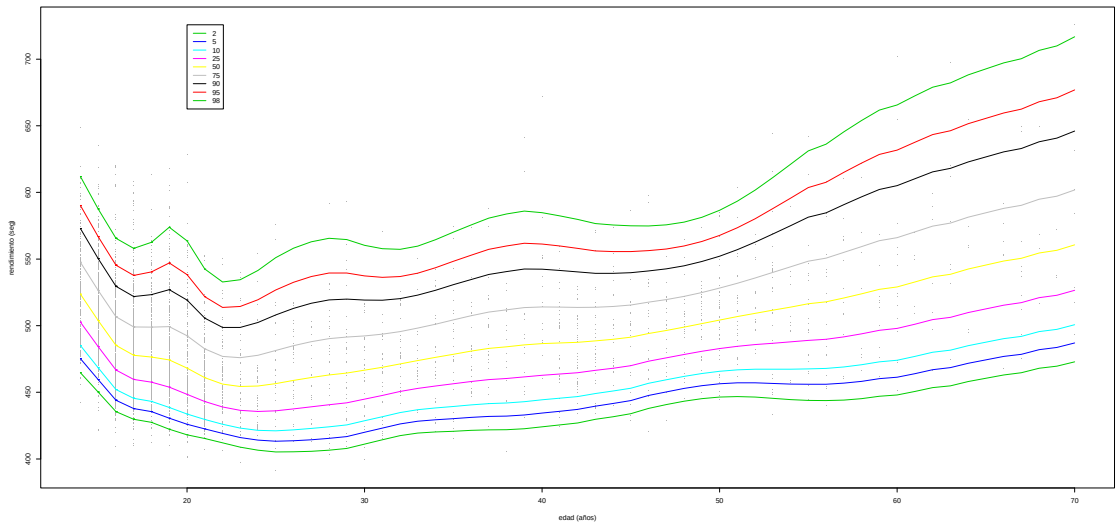


Figura 2. Curvas de percentiles P2, P5, P10, P25, P50, P75, P90, P95 y P98 para el grupo mujer y peso >61,5 kg (n=4023) suavizadas con cubic splines.

En ambas curvas (Figura 1 y Figura 2) se pueden visualizar con claridad dos fases diferenciadas en la evolución del rendimiento con la edad: una fase de crecimiento desde los 14 años, donde el rendimiento mejora hasta los 18 – 23 años en el grupo de hombres y hasta los 22 – 25 años en el grupo de mujeres y una fase de declinación, donde el rendimiento mostrado en la prueba de ergómetro para ambos grupos se reduce con la edad.

Tanto para el grupo de hombres como para el grupo de mujeres se constatan alteraciones importantes en el trazado de los percentiles iguales o superiores a P75, probablemente debido a que estos sujetos realizaron menos entrenamiento en relación al resto de los casos. Concretamente estas diferencias se manifiestan de forma algo más destacada desde los 18 hasta los 40 años en el grupo de los hombres y desde los 17 hasta los 70 años en el grupo de las mujeres.

Además de las curvas se obtuvieron tablas con los valores de los percentiles de rendimiento que, a mayores de los resultados comentados, permiten establecer de forma más exacta el rendimiento alcanzado, así como la edad de rendimiento máximo para cada percentil.

4. CONCLUSIONES

Estas son las primeras referencias y curvas de percentiles de rendimiento que se pueden utilizar, tanto en el remo de competición como en el ámbito de la salud, para comparar y valorar el rendimiento individual en la prueba de evaluación de referencia.

El trazado de las curvas y los valores de percentiles obtenidos indican que el rendimiento en esta prueba se modifica con la edad, identificándose una fase de crecimiento, un momento de rendimiento máximo y una fase de declinación del rendimiento, todas con patrones diferentes según sexo y percentil.

Concretamente, los hombres presentan un mejor rendimiento en esta prueba que las mujeres y, en ambos grupos, se constatan diferencias de trazado en los percentiles, especialmente localizadas a partir del percentil P75.

REFERENCIAS

Kleshnev, V. (2010). Boat acceleration, temporal structure of the stroke cycle, and effectiveness in rowing. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, *224*(1), 63-74.

Mikulić, P., Smoljanović, T., Bojanić, I., Hannafin, J. A., & Matković, B. R. (2009). Relationship between 2000-m rowing ergometer performance times and World Rowing Championships rankings in elite-standard rowers. *Journal of sports sciences*, *27*(9), 907-913.

Nevill, A. M., Allen, S. V. & Ingham, S. A. (2011). Modelling the determinants of 2000 m rowing ergometer performance: a proportional, curvilinear allometric approach. *Scandinavian journal of medicine & science in sports*, *21*(1), 73-78.

Rigby, R. A. & Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine,* 23, 3053–3076.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507-554.

Secher, N. H. (1983). The physiology of rowing. *Journal of Sports Sciences*, *1*(1), 23-53.

Shephard, R. J. (1998). Science and medicine of rowing: a review. *Journal of Sports Sciences*, *16*(7), 603-620.

Sandercock, G., Voss, C., Cohen, D., Taylor, M. & Stasinopoulos, D. M. (2012). Centile curves and normative values for the twenty metre shuttle-run test in English schoolchildren. *Journal of sports sciences*, *30*(7), 679-687.

Vogler, A. J., Rice, A. J. & Withers, R. T. (2007). Physiological responses to exercise on different models of the concept II rowing ergometer. *International journal of sports physiology and performance*, *2*(4), 360.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

AUC-BASED MODEL SELECTION IN GENERALIZED ADDITIVE MODELS FOR THE PREDICTION OF NEW THERAPEUTIC TARGETS

P Cacheiro[1], J Roca-Pardiñas[2], C Cadarso-Suárez[3] , A Carracedo[1,4]

[1]Grupo de Medicina Xenómica, CIBERER, CIMUS, Universidad de Santiago de Compostela, 15706 Santiago de Compostela, Spain.
[2] Departamento de Estadística e Investigación Operativa, Universidad de Vigo, 36208 Vigo, Spain.
[3] Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, 15704 Santiago de Compostela, Spain.
[4]Fundación Pública Galega de Medicina Xenómica, SERGAS, Complexo Hospitalario Universitario de Santiago, 15706 Santiago de Compostela, Spain.

ABSTRACT

Prediction of new therapeutic targets is a primary interest in the drug discovery field. Features used to build predictive models, statistical methods and criteria for selection of positive and negative instances differ among studies. Generalized Additive Models (GAM) are proposed in this work to build predictive models based on protein sequence features, topological networks parameters and functional properties. A variable and model selection method based on AUC and bootstrap confidence intervals was applied in order to find the model with the optimal number of predictors $q \leq p$ and the largest AUC. The resulting classifier might help both to discover innovative targets and to evaluate or rank previous findings.

Keywords: GAM, AUC, model selection, bootstrap, drug target

1. INTRODUCTION

One of the first steps in the drug discovery pipeline is the identification and validation of new therapeutic targets. In silico analysis, using information available in public databases might constitute a useful tool in order to predict if a gene or gene product will potentially constitute a drug target. The majority of previous classifiers are based on physicochemical properties that can be obtained directly from protein sequence data [1-4]. Network properties have also been used for prediction of novel targets since the discovery that drug targets might have specific features in a protein-protein interaction network [5,6]. Some other studies used functional or system properties [7,8]. Regarding the methodology, most predictive models are based on Support Vector Machines [1,2,3,6], but different classifiers including Random Forest [7], Logistic Regression [8] or Rotation Forest [4] were applied as well.

In this study we present an approach based on GAM models using a previously described variable selection method [9]. Different kinds of items were evaluated to test their discriminatory capacity along with different criteria to select positive and negative instances. The objective is to build a tool that will allow us not only to infer new targets but also to help to measure its potential when a new protein is proposed as a target. Our approach is novel in that introduces GAM methodology in the context of prediction of new targets, evaluates the discriminatory capacity of three sets of different properties and allows for different class selection criteria.

2. METHODOLOGY

Our initial dataset comprised all human protein-coding genes. In order to identify the therapeutic targets from this dataset, we made use of two resources, DrugBank [10] - All drug targets and approved

drug targets datasets- and Protein Atlas [11] - FDA approved drug targets and Potential drug targets. The whole set consisting of 19004 protein coding genes was annotated for three different types of attributes to use as predictors. Type A) Physicochemical properties: The R package *Peptides* (v 1.0.3) [12] was used to extract 18 protein sequence features. Type B) Network properties: Protein-protein interaction data was downloaded from BioGRID [13]. NetworkAnalyzer plug-in from Cytoscape (v3.2.1) [14] was used to compute 11 topological parameters corresponding to network nodes. Type C) Functional -System properties: A third set of 10 attributes, related to functional, expression or evolutive properties was retrieved from different resources. This set contains both numerical and categorical attributes. Two variables regarding expression data were extracted from [11] and [15]. The number of similar proteins was computed using the PSI-BLAST algorithm, establishing a cutoff evalue=0.001 using NCBI BLAST blastp suite. The number of pathways was downloaded from the Reactome database [16]. Transcript lengths were obtained with *biomaRt* R package [17]. Evolutive features such as a phylogenetic classification [18], and non-synonymous to synonymous ratios from Phase 3 1000 Genome Project variant set [19] were also accessed. Information regarding gene involvement in disease was retrieved from dbNSFP3.0 [20]. We obtained the sets of genes associated with both monogenic and complex disorders. Lastly, essentiality of the genes was retrieved from the Database of Essential Genes [21].

Generalized Additive Models [22] are flexible models which allow us to introduce non-linear effects of covariates, facilitating the interpretability and visualization of the relationship among the dependent variable and the predictor variables $X_1, \dots, X_p$, In this case, we used a logistic GAM model with binary response $Y(0/1)$ given by

$$P(Y = 1|X_1, \dots, X_p) = \frac{exp\left(\alpha + f_1(X_1) + \dots + f_p(X_p)\right)}{1 + \exp\left(\alpha + f_1(X_1) + \dots + f_p(X_p)\right)}$$

where f_i is a smooth function, estimated from the data.

We applied a modification of a previously described method for multivariate model selection [13]. K=5-fold CV was used to estimate the efficiency of each one of the classifiers, being the area under the ROC curve (AUC) [23] the criterion of choice to evaluate and compare the different models. The performance reported by 5-fold CV was calculated using the average AUC -mean AUC for the 5 folds-. The selected models were subsequently compared to find the optimal number of predictors within each category. In order to asses if differences in AUC between these models were statistically significant, bias-corrected confidence intervals were calculated for each model, by generating B=100 bootstrap samples. We then evaluated if the 95% bootstrap CI for the difference between two given models included zero. The *mgcv* R package implementation of GAM was used to build the models [24].

Since the assignment of class labels differed among previous studies, several criteria to assign positive and negative labels were assessed by introducing weights for the response variable.

3. RESULTS AND DISCUSSION

When models were built considering each type of variables separately, functional and system properties (type C) turned out to be the predictors with the highest predictive ability, followed by physicochemical properties (type A) and, lastly, network properties (type B). Additionally, by combining the different types of variables, we obtained a final combined model with an optimal number of predictors $q_{ABC} = 5$ and an AUC of 0.880 [0.865, 0.895]. This model included 2 type A variables (Hydrophobicity and Instaindex), 1 type B variable (Clustering Coefficient) and 2 type C variables (Phylostratum and RNA tissue category). Up to 20 different combinations of predictors with $q_{ABC} = 5$ were provided with no significant differences in AUC with the selected combined model. Regarding class selection criteria, the best results in terms of AUC were obtained with the most restrictive criteria both for the inclusion of protein as targets (positive instances) and non-targets (negative instances).

The proposed AUC-based method for model selection in GAM models has been successfully applied for the prediction of new drug targets, obtaining high-performance models as indicated by AUC values close to 0.90. Caution must be taken when comparing results among different studies since, the use of different criteria to exclude instances prior to building the model, might have a significant impact on the performance of the resulting classifiers.

ACKNOWLEDGMENTS

This research has been supported by INNOPHARMA (Project MINECO-FEDER, Xunta de Galicia), MTM2014-52975-C2-1-R (MINECO), CN2012/180 (Xunta de Galicia), BIA2011-26915 and MTM2011-23204 –FEDER (Spanish Ministry of Science and Innovation)

REFERENCES

[1] Li Q., Lai L. (2007). Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics* 8, 353.

[2] Bakheet T.M., Doig A.J. (2009). Properties and identification of human protein drug targets. *Bioinformatics* 25, 451–457.

[3] Wang J.T., Liu W., Tang H., Xie H. (2014). Screening drug target proteins based on sequence information. *J Biomed Inform* 49, 269–274.

[4] Kumari P., Nath A., Chaube R. (2015). Identification of human drug targets using machine-learning algorithms. *Comput Biol Med* 56, 175–181.

[5] Yildirim M.A., Goh K.I., Cusick M.E., Barabási A.L., Vidal M. (2007). Drug-target network. *Nat Biotechnol* 25, 1119–1126.

[6] Zhu M., Gao L., Li X., Liu Z., Xu C., Yan Y., Walker E., Jiang W., Su B., Chen X., Lin H. (2009). The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. *J Drug Target* 17, 524–532.

[7] Bull S.C., Doig A.J. (2015). Properties of protein drug target classes. *PLoS One* 10,e0117955.

[8] Yao L., Rzhetsky A. (2008). Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res* 18, 206–213.

[9] Roca-Pardiñas J., Cadarso-Suárez C., Tahoces P.G., Lado M.J. (2009). Selecting variables in non-parametric regression models for binary response. An application to the computerized detection of breast cancer. *Stat Med* 28, 240-259.

[10] Wishart D.S., Knox C., Guo A.C., Shrivastava S., Hassanali M., Stothard P., Chang Z., Woolsey J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 34, D668–D672.

[11] Pontén F., Jirström K., Uhlen M.(2008).The Human Protein Atlas -a tool for pathology. *J Pathol* 216, 387-393.

[12] Osorio D., Rondón-Villarreal P., Torres R. (2014). Peptides 1.0.3: Calculate indices and theoretical physicochemical properties of peptides and protein sequences.

[13] Stark C., Breitkreutz B.J., Reguly T., Boucher L., Breitkreutz A., Tyers M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34, D535-539.

[14] Doncheva N.T., Assenov Y., Domingues F.S., Albrecht M. (2012). Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc* 7,670-685.

[15] Kogenaru S., del Val S., Hotz-Wagenblatt A.,Glatting K.H.(2010). TissueDistributionDBs: a repository of organism-specific tissue-distribution profiles. *Theor Chem Acc* 125, 651–658.

[16] Croft D., Mundo A.F., Haw R., Milacic M., Weiser J., Wu G., Caudy M., Garapati P., Gillespie M., Kamdar M.R., Jassal B., Jupe S., Matthews L., May B., Palatnik S., Rothfels K.., Shamovsky V., Song H., Williams M., Birney E., Hermjakob H., Stein L., D'Eustachio P.(2014). The Reactome pathway knowledgebase. *Nucleic Acids Res* 42, D472–D477.

[17] Steffen Durinck S., Spellman P.T., Birney E., Huber W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 4, 1184-1191.

[18] Domazet-Lošo T., Tautz D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol* 25, 2699–2707.

[19] The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74.

[20] Liu X., Wu C., Li C., Boerwinkle E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Non-synonymous and Splice Site SNVs. *Hum Mutat* 37, 235-241

[21] Luo H., Lin Y., Gao F., Zhang C.T., Zhang R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res* 42, D574–80.

[22] Hastie T., Tibshirani R. (1986). Generalized Additive Models. *Statistical Science* 1, 297–310.

[23] Swets, J.A. and Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection theory*. New York: Academic

[24] Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Comunicacións orais

Sesión IV: Métodos estatísticos en Medicina I

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Propuesta de desarrollo del score global de un cuestionario con una estructura unidimensional ordenada mediante el análisis MCA.

Maider Mateo-Abad[1,2], Arantza Urkaregi[1], Jone Lázaro[4], Inmaculada Arostegui[1,3]

[1] Departamento de Matemática Aplicada, Estadística e Investigación Operativa, Universidad del País Vasco UPV/EHU
[2] Unidad de Investigación AP-OSIs Gipuzkoa. Kronikgune-Osakidetza
[3] BCAM – Basque Center for Applied Mathematics
[4] Departamento de Didáctica de la Matemática y de las Ciencias Experimentales, Universidad del País Vasco UPV/EHU

RESUMEN

El análisis de correspondencias múltiple (ACM) es uno de los métodos más usados para el análisis estadístico de variables cualitativas. El ACM también puede ser usado para el análisis de cuestionarios. Se observa en la literatura que al aplicar el ACM a un cuestionario de Q ítems con una escala ordinal como opción de respuesta, no necesariamente igual para todas las preguntas, frecuentemente aparece una estructura ordenada en las dos primeras dimensiones, con forma de arco llamada efecto Guttman. Nuestra hipótesis es que este efecto puede ser usado para definir una puntuación más eficiente que únicamente sumando los valores numéricos de cada ítem. El objetivo de este trabajo es proponer una metodología para desarrollar un score global basado en las coordenadas del ACM para cada respuesta de un cuestionario donde todas las modalidades de cada pregunta están ordenadas en la primera dimensión del ACM.
La metodología propuesta aplica el ACM a todas las preguntas del cuestionario considerándolas como variables activas. Se toma la primera dimensión del análisis y se prueba que todas las modalidades de respuesta de cada pregunta están ordenadas respecto a esta dimensión. En caso positivo, se atribuye a cada modalidad su coordenada principal. De esta manera se define la puntuación total para cada persona como la suma de las coordenadas principales de cada una de las modalidades seleccionada. Además se puede definir este score global de 0 a 100, haciendo una transformación lineal en la que se tienen en cuenta los valores mínimos y máximos teóricos, es decir, los valores extremos posibles que se podrían esperar al hacer la suma entre las coordenadas correspondientes de los valores mínimos y máximos de cada ítem. Se ha desarrollado un paquete en el software R para la implementación de la metodología propuesta.
Por último, en este trabajo se describe la aplicación a datos reales, y se comparan los resultados con una variable resumen existente en el cuestionario.

Palabras y frases clave: ACM, score global, cuestionarios, escalas, software

1. INTRODUCCIÓN

Las encuestas son una de las herramientas de evaluación más usadas. Son usadas frecuentemente como forma de recopilación de datos, porque dan la oportunidad de recoger datos directamente de un gran número de personas de forma estandarizada y sistemática. Los cuestionarios son empleados y desarrollados con el objetivo de definir un concepto y son muy útiles para recabar información sobre opiniones, conocimientos, comportamientos, experiencias, necesidades, información demográfica, etc. La psicometría es el disciplina que se encarga del desarrollo de este tipo de cuestionarios.

Se han desarrollado varios tipos de escalas de valoración para medir opiniones y actitudes, las más usadas son las escalas valoración ordinales o de tipo Likert. Likert desarrolló el principio de medir actitudes mediante las respuestas de personas a una serie de enunciados sobre un tema. De esta manera los cuestionarios están en una escala ordinal que mide el nivel de conformidad o disconformidad para cada enunciado.
El método más usado para medir la información de este tipo de escalas es sumando la respuesta de cada ítem, dándole a cada respuesta un valor numérico. Este método asume que cada ítem tiene una estructura lineal subyacente, que se representa de forma discreta. Una de las limitaciones de esta propuesta es asumir que las respuestas de cada ítem se distribuyen de forma uniforme, y además que todas las preguntas se distribuyen de la misma forma, y con los mismos pesos. Por ello planteamos una nueva metodología en la que las puntuaciones de cada respuesta las obtenemos a través de la primera dimensión del análisis de correspondencias múltiple (ACM). El ACM es frecuentemente usado en el análisis de cuestionarios en los que se tienen respuestas cualitativas, y cuando hay un único aspecto a medir, la literatura muestra que las respuestas de un cuestionario suelen presentar una forma ordenada en la primera dimensión del ACM. A partir de esta condición, la hipótesis de este trabajo es que se puede definir un score global más eficiente mediante las coordenadas del ACM, que únicamente sumando los valores numéricos de cada ítem. Y para que sea más fácil la implementación de la metodología propuesta a continuación, se ha desarrollado un paquete en software R.

2. METODOLOGÍA APLICADA A LOS DATOS REALES

El sistema de formación de especialistas médicos existente en España desde 1978, basado en la figura del Médico interno residente (MIR), ha hecho una gran contribución en la mejora y desarrollo del Sistema Nacional de Salud (SNS). Debido a esta contribución, se ha aceptado realizar periódicamente una encuesta de evaluación para detectar problemas y encontrar soluciones en la formación de los MIR.
En este contexto, queremos definir un score global basada en los resultados de las encuestas de los MIR realizadas en Osakidetza, que nos permita resumir la evaluación dada por los internos.
Se dispone de 849 encuestas MIR, de internos de diferentes años. El propósito de esta encuesta es valorar la opinión de los internos respecto a la formación recibida.
Se tienen Q=10 variables activas que se incluyen en el ACM. Todas las variables están en una escala ordinal con 4 opciones de respuesta. En este caso, la no respuesta se consideran como una categoría más de cada pregunta. De esta manera tenemos un total de J=50 modalidades diferentes.
Después de aplicar el ACM, el primer paso es comprobar que las coordenadas del ACM correspondientes a las modalidades de cada variable están ordenadas respecto a la primera dimensión. Para comprobar esta condición, no se tienen en cuenta las coordenadas de la no respuesta. En este caso, se cumple la condición. En la Figura 1 se muestran las modalidades representadas en las dos primeras dimensiones del ACM, y en ella se puede ver este efecto.
Una vez comprobada la estructura ordenada para la que hemos planteado esta metodología, podemos calcular la puntuación total. El primer paso es adjudicar la coordenada principal (C_q) de la primera dimensión del ACM a cada modalidad de cada variable q. De esta manera, la puntuación total para cada sujeto el la suma de todas las modalidades elegidas de cada variable:

$$S_{\mathrm{c}} = \sum_{q=1}^{10} C_{q,k_q}\,, \quad \text{donde } k_q \text{ es la modalidad elegida de la variable } q$$

Para obtener la puntuación desde 0 a 100, los valores mínimos y máximos teóricos se definen de la siguiente manera:

$$minS = \sum_{q=1}^{10} min_q\,, \qquad \text{donde } min_q = min\{C_{q,j_q} | \, j_q = 1, \dots, 5\}$$

$$maxS = \sum_{q=1}^{10} max_q\,, \qquad \text{donde } max_q = max\{C_{q,j_q} | \, j_q = 1, \dots, 5\}$$

Por lo tanto el score global (S_{Global}) de 0 a 100 se calcula mediante la siguiente transformación lineal:

$$S_{Global} = \frac{(S_c - minS) \text{ x } 100}{(maxS - minS)}$$

En la Tabla a se presenta el resumen del score global calculado para la muestra. Además se incluye la distribución del score respecto a la pregunta general del cuestionario que hace referencia a la evaluación general del servicio recibido, donde se han encontrado diferencias significativas ($p<0.001$) entre la distribución del score total a lo largo cada modalidad.

	N	Min	Max	Media	SD	Q1	Mediana	Q3
Score Global	849	0	100	57.2	19.6	45	55.6	69.3
Score respecto al ítem de Evaluación general del servicio recibido								
Excelente	184	48.8	100	82.2	12.6	74	84	93
Apropiada	493	23.8	90.7	55.5	11.2	48.6	54.7	61.9
Escasa	139	0	85	34.7	12.7	25.7	33.7	42.2
Muy escasa	14	5.7	39.6	20	10.9	10.2	19.7	28.6
No responde	19	25.4	91.4	52	17.3	42.6	46.8	61.6

Tabla 1: Estadísticos descriptivos del score global. Además se describe la distribución de la puntuación a través de las categorías de la variable resumen existente en el cuestionario.

En caso de querer aplicar la metodología propuesta mediante el paquete desarrollado en R (todavía no está publicado, está en proceso), en este caso el código a compilar sería:

```
> zscore(MIR, c(1:10), nr.consider=T, nr.name="No response")
```

Está función proporciona además del score de 0 a 100 para cada uno de los sujetos incluidos en el estudio, también proporciona el peso de cada modalidad de cada ítem con la que se podría calcular la puntuación total para un nuevo sujeto.

3. CONCLUSIONES

Se propone una metodología para construir un score global valida y eficaz. Teniendo la opción de incluir las no respuestas como una modalidad adicional, se puede calcular el score total para todos los sujetos independientemente de si tienen algún ítem sin contestar, lo que no ocurre cuando se aplican metodologías comunes como sumar simplemente el valor numérico dado a cada categoría.
Además este score tiene una buena capacidad de discriminación, donde las opiniones extremas pueden ser diferenciadas. Adicionalmente, esta puntuación podría usarse para determinar que factores pueden ser significativos en la valoración de un cuestionario.
Por último al desarrollar el paquete en R, se da la oportunidad de que cualquiera pueda implementar el método de forma sencilla. Pudiendo ser aplicada esta metodología a diferentes ámbitos.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT620-13).

Referencias

[1] Benzecri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à. *Les Cahiers De l'Analyse Des Données*, 4(3), 377-378.

[2] Greenacre, M., & Pardo, R. (2006). Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods & Research*, 35(2), 193-218.

[3] Pijoán J.I., Urkaregi, A. & Morán J.M. (2001). Evaluación por los médicos internos residentes de la formación recibida en los servicios hospitalarios: una herramienta de monitorización. *Gaceta Sanitaria*, 15 (5), 432-440.

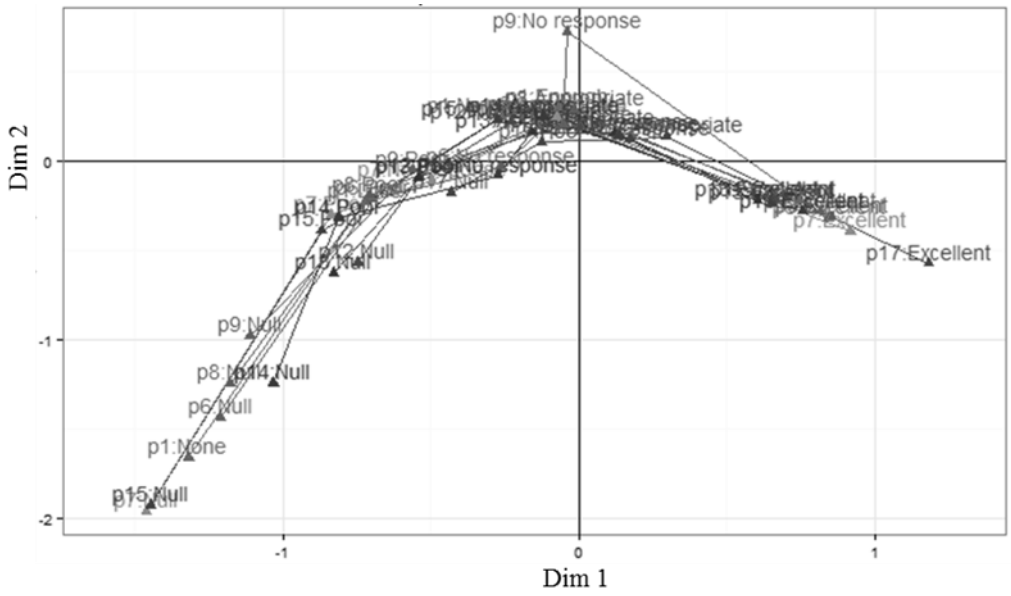


Figura 1: Representación de las modalidades de cada variable en las dos primeras dimensiones del ACM.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de julio, 1 y 2 de julio de 2016

TEORIA DA RESPOSTA AO ITEM: aplicações em enfermagem e saúde

José Wicto Pereira Borges[1], Thereza Maria Magalhães Moreira[2], Dalton Francisco de Andrade[3]

[1] Universidade Federal do Piauí, Brasil, *wictoborges@yahoo.com.br*
[2] Universidade Estadual do Ceará, Brasil, *tmmoreira@cnpq.br*
[3] Universidade Federal de Santa Catarina, Brasil, *dalton.andrade@ufsc.br*

RESUMO: A Teoria da Resposta ao Item (TRI) é uma teoria do traço latente aplicada primariamente a testes de habilidades, desempenho ou comportamento. Atualmente, nas mais diferentes áreas do conhecimento vêm crescendo o interesse na aplicação de técnicas derivadas da TRI, como nos ambientes de enfermagem e saúdee. O termo teoria do traço latente se refere a uma família de modelos matemáticos que relaciona variáveis observáveis e traços hipotéticos não observáveis, estes responsáveis pelo aparecimento da resposta ou comportamento emitido. O fundamental da TRI consiste em expressar numa fórmula matemática a relação existente entre variáveis observadas e variáveis hipotéticas (traço latente). Assim, se conhecemos as características das variáveis observadas estes se tornam constantes na equação, tornando-se solucionável, permitindo que se estime o nível do traço latente do indivíduo [1]. A TRI ainda permite a construção de uma escala que apresenta uma interpretação completa condizente com o traço latente investigado. No foco da mensuração dos fenômenos de enfermagem e saúde é imprescindível que as tecnologias desenvolvidas para o acesso aos traços latentes e sua consequente avaliação se faça por meio de métodos cientificamente aceitos, como as técnicas derivadas da Teoria da Resposta ao Item [2]. Pode-se citar como traços latentes o nível de autocuidado, o comportamento de adesão ao tratamento, a qualidade de vida em hipertensão arterial e a relação interpessoal no cuidado de enfermagem. Estes traços latentes que foram trabalhados por meio da TRI. Neste estudo discutiremos o potencial desta teoria através da apresentação dos resultados da pesquisa realizada para a mensuração da relação interpessoal no cuidado de enfermagem de pacientes em um hospital, uma clínica especializada e centros de atenção primária. Nesse sentido, foi desenvolvido e validado um questionário com 31 itens a partir de um modelo teórico de cuidado de enfermagem. Os itens possuem uma escala adjetival com até quatro alternativas de respostas e foram respondidos por 950 pacientes. Foi realizada uma análise fatorial, para verificar a dimensionalidade do instrumento e, em seguida, o ajuste do Modelo de Resposta Gradual (MRG) da TRI apropriado para a representação de itens com categorias ordinais. Utilizando o MRG as categorias dos itens foram posicionadas na escala. A escala foi desenvolvida em uma métrica (50,5), composta por seis níveis interpretativos, que medem essencialmente o quão efetiva está a relação interpessoal no cuidado de enfermagem. Esta interpretação dos níveis é uma propriedade dos modelos de TRI, que possibilita a criação de um plano de cuidados para o paciente conforme seu escore individual. Com a escala construída, foram obtidas as estimativas de relação interpessoal no cuidado de enfermagem em cada serviço pesquisado, permitindo a comparação entre os serviços.

Palavras chave: Teoria da Resposta ao Item. Modelo de Resposta Gradual.

Agradecimentos: Apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) com bolsa na modalidade Doutorado Sanduiche no País (PSW). Processo 300799/2015-2.

REFERÊNCIAS

[1] Pasquali, L; Primi, R. (2003). Fundamentos da Teoria da Resposta ao Item – TRI. *Avaliação Psicológica* 2(2),99-110.

[2] Araújo, E. A. C.; Andrade, D. F.; Bortollotti, S. L. V. (2009). Teoria da Resposta ao Item. *Rev. Esc. Enferm*, 43(esp),1000-8.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

Contraste da igualdade dun gran número de densidades baixo dependencia

Marta Cousido Rocha[1], Jacobo de Uña Álvarez[2] e Jeffrey D. Hart[3]

[1] Grupo de investigación SiDOR, Universidade de Vigo, Facultade de CC Económicas e Empresariais, Campus Lagoas-Marcosende, 36310 Vigo, España
[2] Departamento de Estatística e OR, e Grupo de investigación SiDOR, Universidade de Vigo, Facultade de CC Económicas e Empresariais, Campus Lagoas-Marcosende, 36310 Vigo, España
[3] Departamento de Estatística, Universidade de Texas A&M, College Station, Texas 77843, U.S.A.

RESUMO

Un tema recorrente na estatística moderna é traballar cun gran número de pequenos conxuntos de datos. Un exemplo deste tipo de datos son os producidos por microarrays. En moitas situacións é relevante comprobar que tódolos conxuntos de datos proveñen dunha única densidade. Por exemplo, en xenética cando cada densidade representa a distribución do nivel de expresión dun xen específico, que se mide para un pequeno número de suxeitos ao longo dun gran número (miles) de xenes, é interesante dispor da ferramenta para comprobar se podemos asumir a igualdade destas densidades.

Zhan e Hart (2012) propuxeron un contraste que permite verificar a hipótese nula de que todos os conxuntos de datos proveñen dunha única densidade cando o número de conxuntos tende a infinito, pero o tamaño mostral de cada un deles é fixo, asumindo que ditos conxuntos de datos son independentes. Porén, na práctica é posible a dependencia entre os diferentes conxuntos de datos, por exemplo se pensamos no campo da xenética é esperable a dependencia entre a expresión de varios xenes, e polo tanto neste tipo de situacións é necesaria unha adaptación do estatístico de Zhan e Hart (2012).

No traballo levouse a cabo unha adaptación do estatístico de Zhan e Hart (2012) ao caso de dependencia, a cal se basea en considerar un enfoque diferente no procedemento de estimación da varianza do estatístico, xa que a proposta de Zhan e Hart (2012) só é válida baixo a asunción de independencia. Posto que dito estatístico é un U-estatístico seguimos o enfoque de Bücher e Kojadinovic (2016a,b) definindo un estimador da varianza do mesmo baseado no multiplicador bootstrap dependente, seguidamente obtemos a distribución asintótica do estatístico baixo a hipótese nula de igualdade das densidades. Tras isto comprobamos o comportamento deste contraste en diferentes escenarios de simulación, observando que a taxa de rechazos baixo a hipótese nula é próxima ao nivel de significación establecido mentres que baixo hipóteses alternativas tende a 1 a medida que aumentamos o número de conxuntos de datos. Para rematar aplicamos no estudo de datos xenéticos o contraste deseñado por Zhan e Hart (2012) e a adaptación ao caso de dependencia proposta, comparando os resultados proporcionados por cada un dos contrastes.

AGRADECEMENTOS

O segundo autor agradece o financiamento do proxecto MTM2014-55966-P do Ministerio de Economía e Competitividade. Por outra parte deséxase recoñer a colaboración dos profesores Robert

Chapkin e Rymond J. Carroll permitindo o uso dos datos de microarrays. Para rematar a primeira autora desexa dar as grazas ao Ministerio de Economía e Competitividade polo seu soporte económico a través das axudas para contratos predoutorais para a formación de doutores concedida na convocatoria correspodente ao ano 2015, tamén agradecer a cofinanciación por parte do FSE (Fondo Social Europeo).

REFERENCIAS

Referencias

[1] Zhan, D., Hart, J. (2012). Testing equality of a large number of densities. Biometrika, 99, 1-17.

[2] Bücher, A., Kojadinovic, I. (2016a). A dependent multiplier bootstrap for the sequential empirical copula process under strong mixing. Bernoulli 22, 927-968.

[3] Bücher A., Kojadinovic, I. (2016b). Dependent multiplier bootstrap for non-degenerated U-statistics under mixing conditions with applications. Journal of Statistical Planning and Inference 170, 83-105

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

TEST DE MULTIMODALIDAD PARA DETERMINAR QUÉ GENES PERMITEN DISTINGUIR GRUPOS DE PACIENTES ONCOLÓGICOS

Jose Ameijeiras–Alonso[1], Rosa M. Crujeiras[1] y Alberto Rodríguez-Casal[1]

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

RESUMEN

El conocimiento de los perfiles de expresión génica de los tumores en pacientes oncológicos permite aplicar tratamientos óptimos, atendiendo a las características de cada tumor. Esto es, el uso de la llamada *medicina personalizada* en el tratamiento del cáncer se apoya en el conocimiento de los niveles de ARN de un conjunto de genes que están desregulados como consecuencia de la patología. Uno de los grandes desafíos a la hora de clasificar a los pacientes en base a los perfiles de expresión génica es, precisamente, encontrar genes cuya expresión esté modificada a causa de la enfermedad y que permitan clasificar claramente a los pacientes en distintos grupos que podrían tener, por ejemplo, diferente respuesta a un mismo tratamiento. En este sentido, aquellos genes que tienen una expresión alta respecto a una persona sana para determinados pacientes y una expresión baja para otros son claros candidatos a convertirse en marcadores pronóstico de la patología. Por este motivo, dentro del campo de la medicina personalizada, resulta de especial interés identificar qué genes presentan dos grupos de expresión (alta/baja) entre los pacientes. La identificación del número de grupos con distinta expresión génica dados unos ciertos datos, puede traducirse en el problema estadístico de determinar el número de modas (máximos relativos de la función de densidad) de la distribución subyacente a la muestra observada. Para poder lograr este objetivo, se presentará un nuevo procedimiento estadístico que permite contrastar el número de modas de una determinada muestra. Además, se aplicará este método a un conjunto de datos de expresión génica de pacientes con cáncer de mama.

Palabras e frases clave: Contraste de Modas, Estimación de la Densidad, Exceso de Masa, Expresión de ARN.

1. INTRODUCCIÓN Y MOTIVACIÓN DEL PROBLEMA

Dentro del campo clínico, el análisis de expresión de ARN se ha convertido en una técnica estándar para detectar qué genes tienen un comportamiento anómalo en un determinado tejido tumoral, permitiendo cuantificar las diferencias de expresión de un determinado gen entre un individuo enfermo y uno sano usado como referencia (véase, por ejemplo, [11]). Una de las técnicas que permite realizar estos análisis son los *microarrays* de expresión: superficies sólidas a las que se unen secuencias complementarias al ARN de los genes de interés y que permiten analizar simultáneamente los niveles de expresión de un gran número de genes. La unión entre el ARN de la muestra y la secuencia inmovilizada en el microarray se detecta gracias al uso de marcajes fluorescentes usando programas de imagen. Dentro de los genes que presentan distintos niveles de expresión en

el tejido tumoral respecto al usado como referencia, la búsqueda de aquellos con un comportamiento diferente en distintos grupos de pacientes que padecen una misma patología se ha vuelto crucial a la hora de decidir qué terapia emplear ante un determinado perfil de paciente. En este sentido, la situación ideal dentro de la denominada medicina personalizada sería la de encontrar un determinado gen para el cual los niveles de expresión de ARN permitiesen separar claramente a los pacientes en dos grupos: unos con expresión alta y otros baja respecto a la referencia.
En particular, se puede estudiar los genes cuya expresión génica presenta un comportamiento diferente en pacientes con cáncer de mama. Esto permitiría conocer que genes permiten diferenciar distintos grupos de pacientes y así a posteriori poder aplicar protocolos distintos en función de la información genética de un determinado individuo. Para ello, bastaría con saber el número de valores alrededor de los cuales se agrupa esta medida y no una completa caracterización de la variable aleatoria que modela la expresión génica. El determinar el número de valores alrededor de los que se agrupan las observaciones corresponde con el problema de determinar el número de modas de una variable aleatoria, que estas no son más que los máximos relativos de la función de densidad.
Si se conociese la distribución de la expresión génica subyacente a la muestra observada en cada gen el problema de determinar el número de grupos de pacientes estaría resuelto. El inconveniente surge cuando, como en el caso que se analizará, se poseen 200 observaciones de individuos con cáncer de mama y cada uno de estos genes (22283 en total) presenta un comportamiento distinto y complicado de modelar (véanse [9] y [6]). Se puede observar, en la Figura 1 (izquierda y centro), este efecto para dos de los genes que analizarán. Se representa la estimación la estimación tipo núcleo, la cual se define a partir de la muestra $(X_1, \ldots, X_n)$ como

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left((x - X_i)/h\right),$$

donde K es una función de densidad unimodal y simétrica y $h > 0$ el parámetro de ventana. En este caso, se ha empleado como núcleo K la densidad de una normal estándar y la ventana ha sido obtenida a través de la regla del pulgar (véase, por ejemplo, [12], Cap. 2 y 3). Observando la Figura 1 parece que la distribución subyacente al gen *202476_s_at* es unimodal y simétrica, mientras que la del gen *200960_x_at* parece ser bimodal y asimétrica. Esta situación complica la asunción de condiciones paramétricas, mientras que la realización de un estudio exploratorio empleando herramientas como el SiZer (véase [2]) implicaría realizar 22283 análisis gráficos. En este caso, una solución formal a estos inconvenientes es elaborar un test, de forma no paramétrica, para determinar si se rechaza que la distribución subyacente a la muestra es unimodal, es decir, se trataría de contrastar $H_0 : j \leq 1$ frente a $H_1 : j > 1$, donde j es el verdadero número de modas de la distribución. Esto permitiría saber para que genes sus expresiones genéticas se agrupan en torno a más de un grupo. Un segundo paso sería determinar qué genes tienen un comportamiento anómalo, esto es, que genes presentan distintos grupos de pacientes, pero la caracterización en enfermos con expresión alta y baja no sería suficiente. Para ello, se podría realizar el contraste $H_0 : j \leq 2$ frente a $H_1 : j > 2$.
En la Sección 2 se resumirá brevemente los distintos test que permiten realizar ambos contrastes, tanto el de analizar la hipótesis nula de unimodalidad, como la de (a lo sumo) bimodalidad. Dentro de estos test, los que presentan un buen calibrado son candidatos a poder analizar cómo de comportan los genes en el caso de la muestra de individuos con cáncer de mama que se presentará en la Sección 3.

2. TEST DE MULTIMODALIDAD

En primer lugar, poniendo el foco de atención en los test que permiten analizar qué genes tienen sus expresiones genéticas agrupadas en torno a un determinado valor, esto es cuando se quiere contrastar si la distribución subyacente a la muestra es unimodal, surgen en la literatura algunas alternativas basadas en la *ventana crítica*, ya sea empleándola como estadístico como es el caso de la introducida en [10] y estudiada en más detalle en [5], o para la construcción de un estadístico tipo Cramér–von–Mises, como la propuesta dada en [4]. Para el contraste de multimodalidad, surgen otro grupo de técnicas basadas en la idea del exceso de masa (véanse [7] y [3]). De todas ellas, solo las desarrolladas en [10] y [4] permiten su extensión al problema general de realizar el contraste

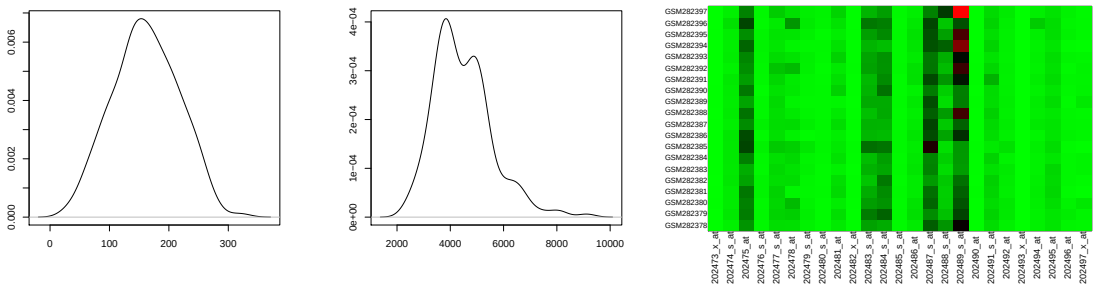


Figura 1: Izquierda: estimación tipo núcleo para la muestra del gen *202476_s_at*. Centro: estimación tipo núcleo para la muestra del gen *200960_x_at*. En ambos casos se ha empleado núcleo gaussiano y la regla del pulgar para la obtención de la ventana. Derecha: microarray asociado a una submuestra de los datos de pacientes con cáncer de mama con 20 pacientes (eje de ordenadas) y 25 genes (eje de abscisas). De rojo claro a verde claro: expresión genética alta a baja.

$H_0 : j \leq k$, cuando $k > 1$. El tener un test que analice este contraste permitiría, fijando $k = 2$, resolver el caso de determinar para qué genes la caracterización en enfermos con expresión alta y baja no es suficiente. El problema que surge con el caso general es que ninguna de estas propuestas presenta un buen calibrado, lo cual proporciona resultados insatisfactorios en la práctica (véase [1]).

Por tanto, para poder resolver el problema de determinar qué genes presentan más de dos grupos de expresión, surge la necesidad de desarrollar un procedimiento adecuado. Para elaborar el nuevo test que permita contrastar el caso general $H_0 : j \leq k$, se utilizará el estadístico de exceso de masa propuesto en [7]. Para obtener la expresión de este estadístico a partir de la muestra es necesario definir el exceso de masa empírico para k modas, este es,

$$E_{n,k}(P_n, \lambda) = \sup_{C_1,\ldots,C_k} \left\{ \sum_{l=1}^{k} P_n(C_l) - \lambda ||C_l|| \right\},$$

donde $\lambda > 0$ y $\{C_1, \ldots, C_k\}$ son intervalos cerrados y disjuntos y $P_n(C_l) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(X_i \in C_l)$ su probabilidad empírica. A partir de esta función, se define el estadístico de exceso de masa $\Delta_{n,k}$, para contrastar $H_0 : j \leq k$, como el máximo en λ de la diferencia entre el exceso de masa empírico para $k+1$ modas y k modas, esto es

$$\Delta_{n,k} = \max_{\lambda} \{E_{n,k+1}(P_n, \lambda) - E_{n,k}(P_n, \lambda)\},$$

rechazándose la hipótesis nula cuando el valor del estadístico sea elevado. Para la obtención del nuevo test que permita contrastar la hipótesis de si la distribución subyacente a la muestra posee (a lo sumo) k modas, será necesario aproximar la distribución de este estadístico bajo la hipótesis nula. En la realización de este nuevo procedimiento, con el objetivo de obtener buenos resultados cuando el tamaño muestral es "pequeño", se ha optado por el empleo de la metodología bootstrap (véase, por ejemplo, [5]), la cual requiere generar nuevas muestras bajo H_0. Para obtener estas remuestras, se emplea la estimación tipo núcleo utilizando una ventana para la cual se estima bien la verdadera función de densidad, suponiendo que esta tiene k modas, y después, se modifica esta estimación alrededor de las modas y antimodas. Esta modificación garantizará que el nuevo test de multimodalidad presentado tiene un comportamiento asintótico adecuado. Los resultados de este test, para tamaños muestrales finitos, han sido comparados con los de las otras propuestas presentes en la literatura a través de un estudio de simulación para comprobar su correcto funcionamiento en la práctica.

3. APLICACIÓN A DATOS DE PACIENTES CON CÁNCER DE MAMA

La aplicación del nuevo test propuesto será realizada sobre datos de pacientes con cáncer de mama. Para ello, se ha tomado la base de datos analizada en [9], esta consiste en el estudio de 200

pacientes, de entre 34 y 89 años, con cáncer de mama tratados en el Departamento de Obstetricia y Ginecología de la Universidad Johannes Gutenberg (Mainz, Alemania) entre los años 1988 y 1998. La base de datos se puede encontrar en el repositorio NCBI GEO con el código GSE11121 (véase [8]). Todos estos pacientes fueron tratados con cirugía, sin recibir ninguna terapia sistemática en el tratamiento realizado para disminuir el riesgo de que el cáncer vuelva. El proceso de selección de genes a estudiar, explicado en completo detalle en [9], dejó un total de 22283 genes cuya expresión de ARN es distinta a la referencia de un individuo sano. En cada paciente de la muestra, para cada uno de estos genes, se ha cuantificado como una variable continua la expresión génica. Sus valores han sido obtenidos a partir de la escala de colores, mostrados en el microarray, desde rojo (expresión alta) hasta verde (expresión baja), se muestra en la Figura 1 (derecha) el microarray para una submuestra de 20 pacientes y 25 genes. Finalmente, el objetivo de este trabajo es el de analizar para cada gen cuando se rechaza la hipótesis de que todos los pacientes pertenecen a un mismo grupo, y la de que pertenecen a lo sumo a dos. Para ello se aplicó el nuevo test propuesto a cada una de las 22283 muestras de tamaño muestral igual a 200.

AGRADECIMIENTOS

La realización de este proyecto ha sido financiada con la ayuda del Proyecto MTM2013–41383–P del Ministerio de Economía y Competitividad y de la red IAP P7/06 StUDyS del Gobierno Belga. Jose Ameijeiras–Alonso agradece también la ayuda para contratos predoctorales BES-2014-071006 del Ministerio de Economía y Competitividad.

REFERENCIAS

[1] Ameijeiras–Alonso, J., Crujeiras, R., Rodríguez–Casal, A. (2015).Mode testing, critical bandwidth and excess mass (Comunicación oral). Libro de actas: Actas del XXXV Congreso Nacional de Estadística e Investigación Operativa y IX Jornadas de Estadística Pública (ISBN 978-84-606-7906-6). (pp: 60). *XXXV Congreso Nacional de Estadística e Investigación Operativa y IX Jornadas de Estadística Pública*. Pamplona (España).

[2] Chaudhuri, P., Marron, J. S. (1999). SiZer for Exploration of Structures in Curves. *Journal of the American Statistical Association* 94, 807–823.

[3] Cheng, M. Y., Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society. Series B* 60, 579–589.

[4] Fisher, N.I., Marron, J. S. (2001). Mode testing via the excess mass estimate. *Biometrika* 88, 419–517.

[5] Hall, P., York, M. (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica* 11, 515–536.

[6] Hellwig, B., Hengstler, J. G., Schmidt, M., Gehrmann, M. C., Schormann, W., Rahnenführer, J. (2010). Comparison of scores for bimodality of gene expression distributions and genome–wide evaluation of the prognostic relevance of high–scoring genes. *BMC bioinformatics* 11, 1–18.

[7] Müller, D. W., Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *The Annals of Statistics* 13, 70–84.

[8] Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr H.–L., Hengstler J. G., Kölbl H., Gehrmann, M. (2015). *NCBI Gene Expression Omnibus (GEO) Series GSE11121.* The humoral immune system has a key prognostic impact in node–negative breast cancer. [fecha de consulta 08/04/2016]. Disponible en Web: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11121

[9] Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr H.–L., Hengstler J. G., Kölbl H., Gehrmann, M. (2008). The humoral immune system has a key prognostic impact in node–negative breast cancer. *Cancer research*, 68, 5405–5413.

[10] Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B*, 43, 97–99.

[11] Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastiee, T., Eisenh M.B., van de Rijni, M., Jeffreyj, S. S., Thorsenk, T., Quistl, H., Matesec, J.C., Brownm, P. O., Botsteinc, D., Lønningg, P. E., Børresen–Daleb, A.–L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98, 10869–10874.

[12] Wand M. P., Jones M. C. (1995). *Kernel Smoothing*. Chapman and Hall. Great Britain.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

O problema de triagem bayesiano do ponto de vista médico

Sandra Ramos[1], Antónia Amaral Turkman[2] e Marília Antunes[3]

[1]Instituto Superior de Engenharia do Porto e Centro de Estatística e Aplicações da Universidade de Lisboa, sfr@isep.ipp.pt
[2]Faculdade de Ciências da Universidade de Lisboa e Centro de Estatística e Aplicações da Universidade de Lisboa, maturkman@fc.ul.pt
[3]Faculdade de Ciências da Universidade de Lisboa e Centro de Estatística e Aplicações da Universidade de Lisboa, mcreis@fc.ul.pt

RESUMO

A discriminação entre os diferentes subtipos de uma determinada doença permite o uso de terapêutica diferenciada, contribuindo para o sucesso dos tratamentos médicos. O problema de triagem, cujo o objetivo é reter, geralmente para observação futura, indivíduos com um vetor ($d-$dimensional) de caraterísticas $\mathbf{x}$ pertencente a uma região $C_{\mathbf{x}}$, desempenha um papel importante nessa discriminação. Num campo de aplicação em contexto médico, a triagem consiste em determinar a categoria t (entre um conjunto de categorias mutuamemete exclusivas) a atribuir a um indivíduo que apresente um conjunto de características $\mathbf{x}$. Existe uma vasta literatura que trata o problema estatístico da triagem, tanto sob um ponto de vista clássico como sob um ponto de vista bayesiano. Do ponto de vista preditivo bayesiano, o problema estatístico consiste em obter a função de probabilidade de diagnóstico $p(t|\mathbf{x}, \mathcal{D})$, onde $\mathcal{D}= \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \ldots, (\mathbf{x}_n, t_n)\}$ representa os dados sobre n elementos cujas categorias são conhecidas, de modo definir a região $C_{\mathbf{x}}$, ótima em determinado sentido. O uso de distribuições paramétricas na obtenção da probabilidade de diagnóstico constitui a abordagem mais usual no problema de triagem [1, 2], já que esta especificação paramétrica permite frequentemente obter, com relativa facilidade, expressões analíticas para as quantidades preditivas necessárias para a adequada formulação do problema de triagem. Porém, em muitas situações práticas não é razoável esperar que as famílias de distribuições propostas ou conjeturadas descrevam apropriadamente o processo que gera as observações, sendo necessário considerar alternativas não paramétricas.

Neste trabalho, apresenta-se uma modelação flexível do problema de triagem baseada em modelos bayesianos não paramétricos, nomeadamente, em misturas por processos de Dirichlet como distribuição *a priori* para a distribuição desconhecida do vetor $\mathbf{X}$ [3]. Esta modelação não paramétrica permitiu aproximar as distribuições preditivas de interesse no problema de triagem e, por conseguinte, a aproximação da probabilidade de diagnóstico. Neste trabalho apresentam-se ainda resultados da aplicação do modelo proposto na discriminação, com base em informação genética, de indivíduos com diferentes tipos de cancro.

Palavras e frases chave: triagem em contexto médico, metodologia bayesiana, métodos não paramétricos, processo de Dirichlet, discriminação de diferentes subtipos de cancro.

AGRADECIMENTOS

Trabalho financiado pela Fundação para a Ciência e a Tecnologia, Portugal, através do projeto UID/MAT/00006/2013.

Referências

[1] Ramos S., Amaral Turkman M.A., Antunes M. (2010). Bayesian classification for bivariate normal gene expression. *Computational Statistics and Data analysis* 54, 2012–2020.

[2] Turkman K.F., Amaral Turkman M.A. (1989). Optimal screening methods. *Journal of the Royal Statistical Society* B, 51, 287–295.

[3] Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249–265.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

A new approach to predict changes in physical condition with application preventive medicine

Marcos Matabuena Rodríguez[1]

[1]Departamento de Análise Matemática, Universidade de Santiago de Compostela, 15782, Santiago de Compostela, Spain

1 Introduction

Athletes, in their quest to optimize performance, try to use mathematical tools [11], but nowadays there are no tools to know the effect of workout on the final performance [9]. A first attempt at the problem of modeling and predict the changes induced by fitness training appeared in 1975 with the contribution of the physiologist Banister [1], whose model has been frequently used in the literature to predict the effects of training on physical performance, see, for example, its validation in triathlon athletes [2], the case of a swimmer [4], studies in cross-country skiers [5], the convenience of exercise physiology practitioners and researchers to use mathematical models [6], the optimal design of a training strategy [8], the performance of runners [12], or the consideration of altitude training [15], among others.
Since the introduction of the Banister model, there have been proposed modified versions of the model, some of which are those introduced by Busso [3] or Allan and Coggan [7], among others, and some new models have been built through regression techniques, neural networks [14] or time series [13].
In this paper, we propose a new model based on techniques of functional differential equations and nonlinear regression that substantially improves the accuracy of the above methods, besides overcoming some of their limitations [10]. The model proposed has been successfully applied to study and predict the effect of training of a cyclist in the task of optimizing a workout plan, constituting what might be a future reliable and useful tool for coaches in differents sports differents of cycling .

2 Methods

The model proposed is based on the study of the properties of differential equations with delay and the estimation of parameters.
Let $g(t)$ and $h(t)$ be the positive and negative effects at time t. The physical condition at time t (denoted by $p(t)$) is given by the following identity:

$$p(t) = p(0) + k_1 g(t) - k_2 h(t), \tag{1}$$

where k_1 and k_2 are two positive constants and $p(0)$ is the initial physical condition. The values of the functions $g(t)$ and $h(t)$ are known by solving the following system of differential equations:

$$g'(t) = w(t) - \frac{1}{\tau_1} g(t) - \frac{1}{\tau_2} g(t-1), \tag{2}$$

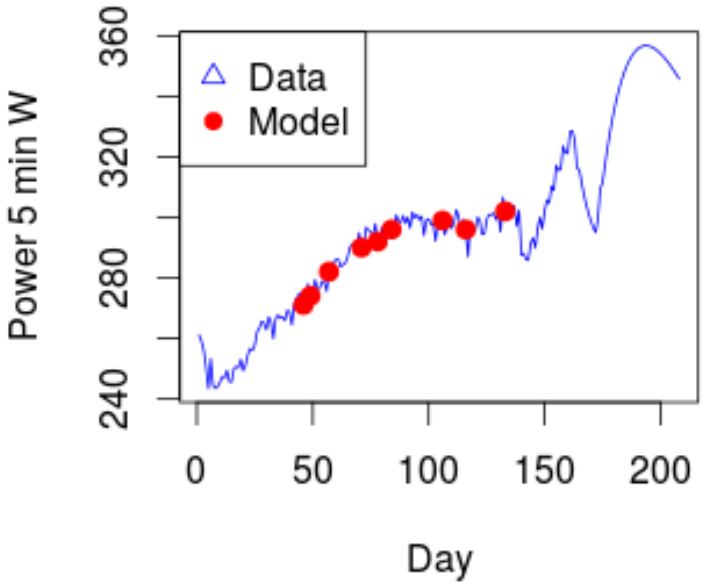


Figure 1: Example Banister model with delay

$$h'(t) = w(t) - \frac{1}{\tau_3}h(t) - \frac{1}{\tau_4}h(t-1), \tag{3}$$

where $w(t)$ represents the load at the instant of time t and τ_1, τ_2, τ_3, τ_4 are positive constants. Since, in practice, we work with measurements at fixed times, we consider a discrete model and approximate the solution to the differential system by considering the following difference equation:

$$p(n) = k_1 \sum_{i=1}^{n-1} \left[w(i) - \frac{1}{\tau_2} g(i-1) \right] e^{\frac{-1}{\tau_1}(n-i)} - k_2 \sum_{i=1}^{n-1} \left[w(i) - \frac{1}{\tau_4} h(i-1) \right] e^{\frac{-1}{\tau_3}(n-i)}, \tag{4}$$

where n denotes the nth instant of time.
At this point, we use nonlinear regression techniques or any heuristic algorithm, to estimate the constants k_1, k_2, τ_1, τ_2, τ_3, τ_4 by least squares method, using a sample of physical condition $p(1), \ldots, p(s)$ at different times. In this work, we use the experimental data of a cyclist extracted from the article [6]. To optimize the constants, we use R package and also we represent the results with the statistical package R.

3 Results

The result of the R^2 test between the experimental data and the mathematical model proposed is 0.9903. In Figure 1, we find a graph where the red points represent the data and the blue curve the solution given by the model. In Figure 2, we see the load executed by the athlete in his training, which determines the change in his physical condition. We observe that this graph presents high and low values, which correspond to the effect of the fatigue states and the assimilation of the training, what actually happens in reality.

4 Conclusions

In this work, we have shown that, using the model proposed, it is possible to obtain an excellent fit with experimental data, where moreover, we get calculate the positive and negative effects of the sport training with the ability to predict the effect of training loads in the sport performance. The improvement in the fit respect to other models, leads us definitely to an important tool for planning sports coach, and new gadgets and software that attempt to have control the effect that sports training has on the athlete individually [16].
Besides, this model could be adapted for weight control or to predict the risk of injury or the emotional state of the athlete, issues which play an important role in athletic performance. This model has also special interest to sedentary people when start physical a activity routine (for example, with the objective decrease his weight) or in the specific medical treatment after a one injury or medical problem in general. With all these applications of the model, we will be able to get important economic savings due to positive effects associated a one good training plan, moreover the immediate effects in our healthy.

References

[1] E.W. Banister, T.W. Calvert, M.V. Savage, and T. Bach. A systems model of training for athletic performance. *Australian Journal of Sports Medicine*, 7(3):57–61, 1975.

[2] E.W. Banister, J.B. Carter, and P.C. Zarkadas. Training theory and taper: validation in triathlon athletes. *European Journal of Applied Physiology and Occupational Physiology*, 79(2):182–191, 1999.

[3] T. Busso. Variable dose-response relationship between exercise training and performance. *Medicine and Science in Sports and Exercise*, 35(7):1188–1195, 2003.

[4] T.W. Calvert, E.W. Banister, M.V. Savage, and T. Bach. A systems model of the effects of training on physical performance. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-6(2):94–102, 1976.

[5] R. Candau, T. Busso, and J.R. Lacour. Effects of training on iron status in cross-country skiers. *European Journal of Applied Physiology and Occupational Physiology*, 64(6):497–502, 1992.

[6] D.C. Clarke and P.F. Skiba. Rationale and resources for teaching the mathematical modeling of athletic training and performance. *Advances in Physiology Education*, 37(2):134–152, 2013.

[7] A.R. Coggan. Training and racing using a power meter: an introduction, 2006.

[8] J.R. Fitz-Clarke, R.H. Morton, and E.W. Banister. Optimizing athletic performance by influence curves. *Journal of Applied Physiology*, 71(3):1151–1158, 1991.

[9] J.J. González Badillo and E. Gorostiaga. *Fundamentos del entrenamiento de la fuerza. Aplicación al alto rendimiento deportivo*. INDE, 1995.

[10] P. Hellard, M. Avalos, L. Lacoste, F. Barale, J.-C. Chatard, and G.P. Millet. Assessing the limitations of the Banister model in monitoring training. *Journal of Sports Sciences*, 24(5):509–520, 2006.

[11] V.B. Issurin. New horizons for the methodology and physiology of training periodization. *Sports Medicine*, 40(3):189–206, 2010.

[12] R.H. Morton, J.R. Fitz-Clarke, and E.W. Banister. Modeling human performance in running. *Journal of Applied Physiology*, 69(3):1171–1177, 1990.

[13] M. Pfeiffer. Modeling the relationship between training and performance - a comparison of two antagonistic concepts. *International Journal of Computer Science in Sport*, 7(2):13–32, 2008.

[14] M. Pfeiffer and A. Hohmann. Applications of neural networks in training science. *Human Movement Science*, 31(2):344–359, 2012.

[15] F.A. Rodríguez, X. Iglesias, B. Feriche, C. Calderón-Soto, D. Chaverri, N.B. Wachsmuth, W. Schmidt, and B.D. Levine. Altitude training in elite swimmers for sea level performance (Altitude Project). *Medicine and Science in Sports and Exercise*, 47(9):1965–1978, 2015.

[16] H.M. Toussaint, P. de Lange, and M.J. Truijens. Applying technology to optimise training: A look in the (near?) future. In M. Sidney, F. Potdevin, and P. Pelayo, editors, *Proceedings of the 4ème Journées Spécialisées de Natation*, pages 105–113. Faculté des Sciences du Sport et de l'Education Physique, Université de Lille2, Lille, 2008.

Comunicacións orais

Sesión V: Métodos estatísticos en Agricultura e Bioloxía

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

MODELO NÃO-LINEAR MISTO APLICADO À EVAPOTRANSPIRAÇÃO DA ALFACE

Omar Pereira[1], Altair Bertonha[2] e Isolde Previdelli[1]

[1]Departamento de Estatística - Universidade Estadual de Maringá
[2]Departamento de Agronomia - Universidade Estadual de Maringá

RESUMO

O aumento da população mundial trará um decréscimo na qualidade e disponibilidade de água e um aumento na demanda por alimentos. Considerada a tecnologia mais eficiente para o aumento da produtividade agrícola, a irrigação é também a maior consumidora de água no mundo. Portanto, a sustentabilidade dependerá de altas produtividades agrícolas com uso racional de água. Neste trabalho avaliou-se a evapotranspiração da alface *Lactuca sativa* em função de quatro níveis de irrigação, cultivada em vasos sob ambiente protegido ao longo de 23 dias consecutivos, por meio de um modelo não-linear de efeitos mistos. Quando a umidade do substrato nos vasos atingia 50 % da sua capacidade máxima de retenção, elevava-se o nível de água de acordo com os quatro tratamentos: $W_1 = 62.5\,\%$; $W_2 = 75.0\,\%$; $W_3 = 87.5\,\%$ e $W_4 = 100.0\,\%$. O modelo ajustou-se muito bem aos dados e mostrou que as estimativas de máxima evapotranspiração foram similares para os três primeiros tratamentos e menor para W_4. Isto indica a possibilidade de maximizar a produção de alimentos com menores níveis de água no solo, ou seja, o uso dos recursos hídricos com racionalidade e sem desperdício também foi a maneira mais eficientes de potencializar a produção agrícola.

Palavras e frases chave: Dados longitudinais, irrigação, manejo de água no solo.

1. INTRODUÇÃO

Modelos não-lineares, na maioria das vezes, são deduzidos a partir de suposições teóricas relacionadas a algum mecanismo físico (ou biológico, etc) inerentes ao fenômeno de interesse. Geralmente são soluções de equações diferenciais com parâmetros resultantes interpretáveis. Os modelos não-lineares mistos se propõem, ainda, a estimarem o comportamento médio de um indivíduo na população e a variabilidade entre e dentre indivíduos.

2. PROCEDIMENTOS EXPERIMENTAIS

O experimento foi conduzido em casa de vegetação na Universidade Estadual de Maringá, Brasil. Os tratamentos correspondem a quatro níveis de água no solo. Cada unidade experimental foi composta por um vaso de polietileno preto, sem dreno, com 2000 g de material obtido a partir de um Latossolo Vermelho, com 30 % de argila, com uma única planta de alface *Lactuca sativa* cv. Verônica. Esse material foi coletado de uma camada de 20 cm de profundidade de solo, passado em peneira de 4 mm e seco ao ar, e sua densidade nos vasos foi padronizada em 1.33 $g\ cm^{-3}$. A

capacidade máxima de retenção de água deste material no vaso foi de 27.4 % sobre a massa seca de substrato. Quando a umidade do substrato atingia 50 % da sua capacidade máxima de retenção, ocorria a suplementação de água elevando a umidade para os níveis $W_1 = 62.5\,\%$, $W_2 = 75.0\,\%$, $W_3 = 87.5\,\%$ e $W_4 = 100.0\,\%$. Em outras palavras, os tratamentos foram mantidos com déficit entre 50 e 37.5 %, 50 e 25 %, 50 e 12.5 %, e 50 e 0 %, respectivamente, com 9 repetições para os três primeiros tratamentos e 6 para o último, totalizando 33 unidades experimentais. A massa de água evapotranspirada de cada vaso foi determinada diariamente durante 23 dias consecutivos com uma balança de capacidade para 8 kg e precisão de 0.1 g. Ao longo do experimento, plantas adicionais foram destruídas para correção da massa dos vasos devido ao aumento da matéria fresca em cada tratamento. Ao final dos tratamentos avaliou-se a massa de matéria seca da parte aérea por vaso.

3. MODELO NÃO-LINEAR MISTO

O modelo não-linear misto proposto para o ajuste da curva de crescimento da evapotranspiração, E_{ij}, da planta i no tempo t_j, foi o modelo logístico de quatro parâmetros escrito como

$$E_{ij} = \phi_{2i} + \frac{\phi_{1i} - \phi_{2i}}{1 + \exp\left[(\phi_{3i} - t_j)/\phi_{4i}\right]} + \epsilon_{ij} \quad i = 1, \ldots, 33 \text{ e } j = 1, \ldots, 23, \tag{1}$$

em que

$$\boldsymbol{\phi}_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix} = \begin{bmatrix} \beta_1 + \gamma_1 x_{1i} + \delta_1 x_{2i} + \zeta_1 x_{3i} \\ \beta_2 + \gamma_2 x_{1i} + \delta_2 x_{2i} + \zeta_2 x_{3i} \\ \beta_3 + \gamma_3 x_{1i} + \delta_3 x_{2i} + \zeta_3 x_{3i} \\ \beta_4 + \gamma_4 x_{1i} + \delta_4 x_{2i} + \zeta_4 x_{3i} \end{bmatrix} + \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{bmatrix} \tag{2}$$

$$= \boldsymbol{\beta} + \boldsymbol{\gamma} x_{1i} + \boldsymbol{\delta} x_{2i} + \boldsymbol{\zeta} x_{3i} + \boldsymbol{b}_i. \tag{3}$$

Os parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$, $\boldsymbol{\zeta}$ e $\boldsymbol{\tau}$ representam os efeitos fixos do modelo e os $\boldsymbol{b}_i$ os efeitos aleatórios, os quais são considerados independentes entre as plantas. As covariáveis x_{ki}, $k = 1, 2, 3$ indicam o nível de água que a i-ésima planta recebeu. Por exemplo, se a planta i recebeu o terceiro tratamento, então $x_{1i} = 0$, $x_{2i} = 1$ e $x_{3i} = 0$. Os erros $\epsilon_{ij} \sim N(0, \sigma^2)$ são considerados independentes dos efeitos aleatórios, $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$, e para os diferentes i e j. O parâmetro ϕ_{1i} é a assíntota horizontal quando $t \to -\infty$, ϕ_{2i} é a assíntota horizontal quando $t \to +\infty$, ϕ_{3i} é o valor de t no ponto de inflexão (ponto médio entre as assíntotas), e ϕ_{4i}, o parâmetro de escala (sempre positivo) [1]. As análises foram realizadas no *software* estatístico **R** [2] com uso do pacote **nlme** [3].

4. RESULTADOS E DISCUSSÃO

O objetivo foi verificar a relação entre os parâmetros do modelo logístico para os dados da evapotranspiração acumulada da alface e o fator com quatro níveis de água. Primeiramente, ajustou-se o modelo (1) sem covariáveis e $\boldsymbol{\Psi}$ com estrutura geral definida-positiva. Em seguida, um modelo mais geral, que permite diferentes variâncias residuais para os tratamentos (heterocedasticidade), foi proposto para representar as dados. Além disso, todos os parâmetros do modelo variaram de acordo com os níveis de água, indicando que o efeito dos tratamentos é importante para o ajuste destes parâmetros. Em outras palavras, testando modelos aninhados com $\gamma_h = \delta_h = \zeta_h = 0$, $h = 1, 2, 3, 4$ ao nível de 5 %, resulta na rejeição de todos eles. Num processo de construção de modelos, em que testou-se a inclusão dos tratamentos e a eliminação de efeitos aleatórios [1] , o modelo final escolhido só não inclui efeito aleatório para o parâmetro de escala ϕ_4.

A Figura 1 apresenta três gráficos de diagnóstico para o modelo logístico de quatro parâmetros. Os dois primeiros mostram os resíduos padronizados *versus* valores ajustados e os valores observados em função dos valores estimados (a linha reta representa o ajuste perfeito). Estes gráficos não indicam grandes desvios do modelo não-linear, exceto por alguns pontos estimados. O outro gráfico, quantil-quantil, mostra que os resíduos atendem a suposição de normalidade, ou seja, a linearidade dos pontos não sugere uma violação grave desta suposição.

Os valores estimados para os efeitos fixos com os respectivos erros padrão estão na Tabela 1. O interesse principal recai sobre os parâmetros ϕ_2, a assintota superior que estima o potencial máximo de evapotranspiração, e ϕ_3, o ponto que inflexão. Nos tratamentos W_2 e W_3, tanto ϕ_2 como ϕ_3 não diferem, ao nível de 5 % de significância, de W_1, indicando que estes três níveis de água são

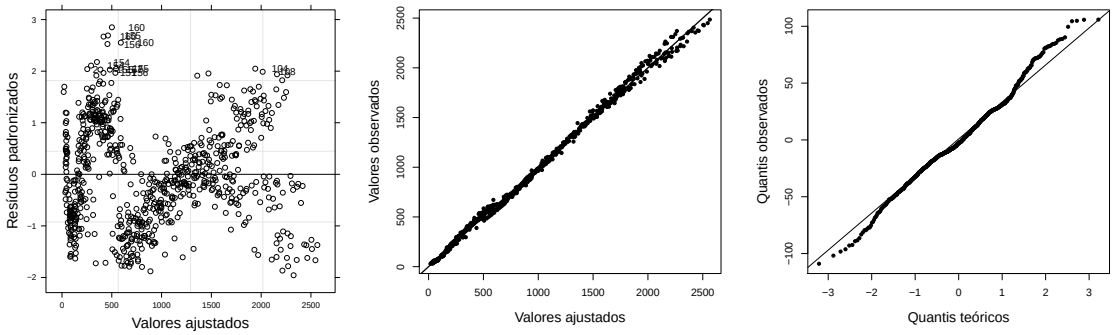


Figura 1: Gráficos de diagnósticos do modelo não-linear para evapotranspiração acumulada da alface em função de níveis de água. O primeiro e segundo gráficos mostram os resíduos padronizados e os valores observados em função dos valores ajustados. O último é o gráfico quantil-quantil.

	$\hat{\boldsymbol{\beta}}$ (W_1)	$\hat{\boldsymbol{\gamma}}$ (W_2)	$\hat{\boldsymbol{\delta}}$ (W_3)	$\hat{\boldsymbol{\zeta}}$ (W_4)
ϕ_1	-210 (20.93)**	-74 (32.31)*	-20 (30.3)	-94 (36.58)*
ϕ_2	3670 (159.71)**	-197 (241.53)	242 (239.1)	-1337 (197.69)**
ϕ_3	18.95 (0.52)**	0.43 (0.85)	0.43 (0.77)	-4.51 (0.69)**
ϕ_4	6.77 (0.22)**	1 (0.37)*	0.18 (0.32)	0.24 (0.37)

$^{**}p < 0.01$, $^{*}p < 0.05$

Tabela 1: Estimativas com erros padrão para os efeitos fixos do modelo.

equivalentes. No tratamento W_4 as estimativas destes dois parâmetros foram significativamente diferentes dos demais níveis de água, com ϕ_2 e ϕ_3 representando em torno de 60 % e 75 % de W_1. As curvas estimadas para os perfis individuais (efeitos aleatórios) e os dados observados podem ser visto na Figura 2 de acordo com os níveis de água. A linha mais grossa e preta representa o modelo marginal estimado, ou seja, a média populacional para cada tratamento, obtida quando os efeitos aleatórios são iguais a zero. Para todos os grupos, os modelos preditos estão muito próximos dos valores observados, indicando um ótimo ajuste das curvas pelo modelo não-linear de efeitos mistos. O tratamento W_4, quando comparado aos outros níveis de água, apresenta uma curva média claramente mais inferior.

A Figura 3 mostra a produção de matéria seca da parte aérea das plantas de alface em função do total de água evapotranspirada. Observa-se que, a medida que a evapotranspiração acumulada aumenta, a produção de matéria seca também aumenta. Além disso, parece haver uma relação linear entre estas variáveis, e um ajuste com inclinação igual a 0.003 foi adicionado, indicando que para cada quilograma de matéria seca produzido, há necessidade de 300 quilogramas de água evapotranspirada por planta.

5. CONSIDERAÇÕES FINAIS

O modelo não-linear de efeitos mistos ajustou-se bem aos dados da evapotranspiração ao longo dos 23 dias em que as plantas de alface foram submetidas à quatro níveis de água no solo. Com o modelo mostrou-se que o menor nível de umidade no solo, W_1, causou evapotranspirações tão elevadas quanto aos níveis W_2 e W_3, e que, o maior nível de água, W_4, teve o pior desempenho. Dessa forma, o aumento da produção agrícola está condicionada à disponibilidade de água no solo e o uso racional dos recursos hídricos também foi a forma mais eficiente para maximizá-la.

Referencias

[1] Pinheiro, J. C., Bates, D. M. (2002). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

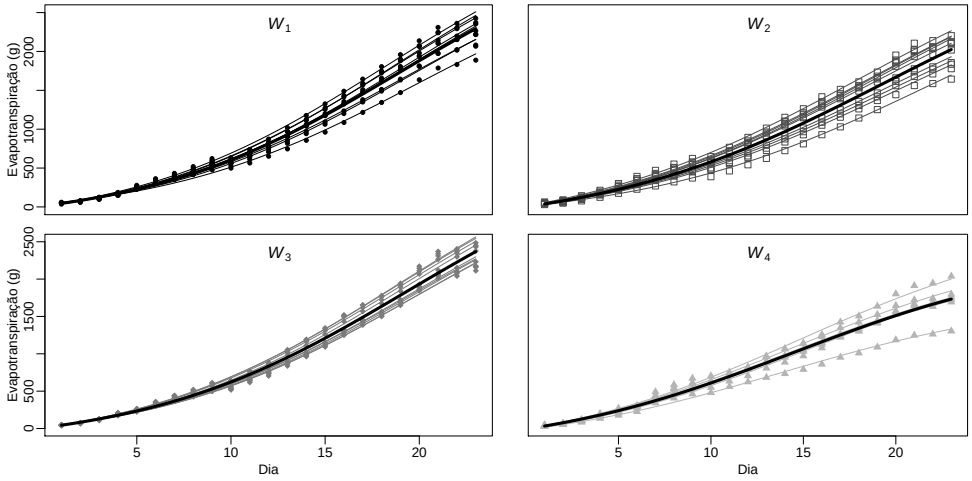


Figura 2: Curvas de crescimento observadas e preditas para a evapotranspiração acumulada de alface em função dos quatro níveis de água: $W_1 = 62.5\,\%$, $W_2 = 75.0\,\%$, $W_3 = 87.5\,\%$ e $W_4 = 100\,\%$. Em cada nível de água, a linha mais escura é o modelo marginal e as demais os perfis condicionais.

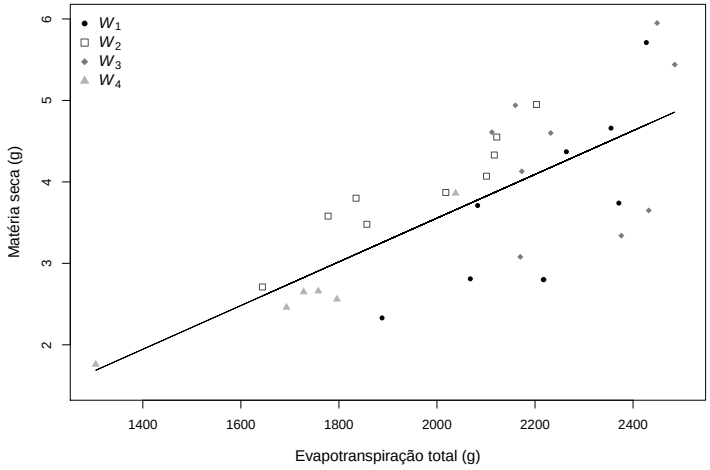


Figura 3: Massa de matéria seca da parte aérea das plantas de alface por vaso (g) em função da quantidade total de água evapotranspirada por tratamento (g) e o ajuste linear.

[2] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[3] Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-117.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de julio, 1 y 2 de julio de 2016

Optimización del esfuerzo taxonómico mediante la aplicación de técnicas estadísticas

Andrea Freijeiro[1], Carola Gómez-Rodríguez[1] y Andrés Baselga[1]

[1] Departamento de Zoología, Facultad de Biología, Universidad de Santiago de Compostela, c/ Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, España

RESUMEN

El número total de especies con las que compartimos el planeta es desconocido, pero se estima que puede oscilar entre los 2 y 5 millones. Tras siglos de estudio, el número de especies desconocidas es aún elevado ya que solamente se han descrito 1.6 millones. El proceso taxonómico, mediante el cual se describen especies nuevas para la ciencia, ha sido tradicionalmente muy lento y laborioso. Sin embargo, un análisis estadístico de cómo se ha producido la acumulación del conocimiento taxonómico permitirá realizarlo de forma más eficiente en un futuro. Este aspecto es clave ante las amenazas a la biodiversidad actuales, ya que nos arriesgamos a que se produzcan extinciones de especies que aún no conocemos.

En este trabajo analizamos cómo ha sido la descripción de especies de una familia (Chrysomelidae) de escarabajos (Coleoptera). Los coleópteros son un grupo hiperdiverso del que se estima que quedan muchas especies por descubrir, por eso son un taxón modelo para este tipo de estudios. La principal hipótesis de estudio es que la localización geográfica y el tamaño de la especie han condicionado su probabilidad de descripción (i.e. Fecha en la que fue descrita) así como el hecho de que se realizara correctamente y una única vez (i.e. número de sinonimias u ocasiones en las que una especie ha sido descrita de forma redundante bajo diversos nombres). Se ha construido una base de datos de los Chrysomelidae europeos en la que se recogen los nombres de especies, el número de sinonimias, el tamaño corporal y variables espaciales y del área para cada caso; abarcando 250 años de estudio taxonómico y 1599 especies y subespecies. Posteriormente se han aplicado OLS (Ordinary Least Squares regression) para estudiar la probabilidad de descripción y de sinonimia. Los modelos estadísticos muestran que las especies con distribuciones más amplias se han descrito antes mientras que el tamaño corporal de las mismas, a diferencia de otros estudios, no influye en la probabilidad de descripción. Además estas especies descritas con anterioridad y con distribuciones amplias acumulan un mayor número de sinónimos por lo que sería conveniente realizar revisiones de las descripciones de los últimos años que podrían esconder sinónimos no detectados todavía.

Queda patente por tanto cómo el empleo de técnicas estadísticas puede beneficiar a ciencias como la taxonomía, ayudando a focalizar futuras investigaciones. En esa línea, este trabajo sirve para informar a los taxónomos de crisomélidos sobre la importancia de dirigir su búsqueda de nuevas especies hacia aquellos casos con distribuciones más pequeñas y no tanto en especies de menor tamaño corporal, al mismo tiempo que se deberían realizar revisiones sobre la taxonomía más reciente.

Palabras y frases clave: Chrysomelidae, probabilidad de descripción, sinonimia, taxonomía

II Encontro Galaico-Português de Biometría
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

AN INTEGRATED POPULATION DYNAMICS MODEL FOR WILD TURKEYS

Russell Alpizar-Jara[1,2]

[1] Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, Rua Romão Ramalho 59, 7000-671, Évora, Portugal
[2] Departamento de Matemática, Escola de Ciências e Tecnologia, Universidade de Évora, Rua Romão Ramalho 59, 7000-671, Évora, Portugal

RESUMO

Integrated population dynamics models have become popularly used during the last decades [3]. These are models that jointly analyze data on population size and data on demographic parameters. Due to difficulties of incorporating data for parameter estimation in conventional population projection matrix-type models (i.e. Leslie and Lefkovitch), an integrated analyses with a state-space formulation has proven to be very useful [2]. This approach allows inferences about population dynamics accounting for parameter estimates and model uncertainties due to process variation, such as demographic and environmental stochasticity, and observational error. I will present an application of these models to wild turkey populations based on telemetry data from the Virginias [1], U.S.A.

Palabras e frases chave: *Meleagris gallopavo*, Population dynamics, Radiotelemetry data.

AGRADECIMENTOS

The author belongs to the Centro de Investigação em Matemática e Aplicações, Instituto de Investigação e Formação Avançada, Universidade de Évora, a researh center supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal, under project PEst-OE/MAT/UI0117/2014).

Referências

[1] Alpizar-Jara, R., Brooks, E., Pollock, K.H., Steffen, D., Pack, J.C. and Norman, G.W. (2001). An Eastern Wild Turkey Population Dynamic Model for the Virginias. *Journal of Wildlife Management*, 65(3): 415-424.

[2] Buckland, S. T., Newman, K. B., Thomas, L. and Koesters, N. B. (2004). State-space models for the dynamics of wild animal populations, *Ecological Modelling*, 171, **(1-2)**, pp. 157–-175.

[3] Newman, K., Buckland, S.T., Morgan, B., King, R., Borchers, D.L., Cole, D., Besbeas, P., Gimenez, O., Thomas, L. (2014). *Modelling Population Dynamics: Model Formulation, Fitting and Assessment using State-Space Methods.* Methods in Statistical Ecology. Springer Science, New York, pp. 215.

Comunicacións orais

Sesión VI: Epidemioloxía I

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Current statistical approaches to antibody data analysis of infectious diseases on the brink of elimination

Nuno Sepúlveda[1,2], Gillian Stresman[1], Michael White[3] and Chris Drakeley[1]

[1]London School of Hygiene and Tropical Medicine, United Kingdom
[2]Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[3]Division of Population Health and Immunity, Walter and Eliza Hall Institute, Australia

ABSTRACT

The human immune system has the capacity of responding to a virtually infinite number of pathogens. This capacity is derived from the so-called adaptive arm of the immune system composed of B and T cells that ultimately lead to the production of specific antibodies targeting the invading microorganisms. In some settings, the immune response is strong enough to induce the persistence of these specific antibodies over time, thus, providing the basis of an immunological memory of the pathogen. As a consequence, this antibody persistence quickens the immune response every time the same pathogen. An important practical application of this immunological principle is in the core of vaccine development where one seeks to identify pathogen-derived antigens able to trigger this type of immunological memory. A less-obvious application is to use these long-lasting antibodies for epidemiological purposes, specifically, in infectious diseases with low prevalence or on the cusp of elimination. The underlying rationale is that traditional infectivity measures have limited statistical power to tackle the epidemiology of these diseases due to the high chance of detecting only a few infected individuals in the population at a given time point. The persistence of pathogen-related antibodies over time leads to the definition of alternative measures (i.e., seroprevalence or seroconversion rate) related to disease exposure rather than presence of infection. However, the use of these measures in practice requires the knowledge of less-known statistical tools, such as (i) the Gaussian mixture models for classifying the individuals into the respective serological state that ultimately leads to the estimation of seroprevalence, (ii) the reversible catalytic or the superinfection models for describing the evolution of the corresponding serological state over time, which allows the estimation of the seroconversion rate, (iii) the antibody-acquisition models for tackling time changes of the antibody concentration in apparently stable serological classification of the individuals. This talk reviews these different statistical approaches to antibody data analysis. Real-world data will be used to illustrate the application of these approaches in practice, giving special emphasis to Malaria [1, 2] where disease elimination might be achieved in many different countries soon.

Keywords and key sentences: epidemiology, measuring elimination, nonlinear stochastic models.

References

[1] Corran P., Coleman P., Riley E., Drakeley C. (2007). Serology: a robust indicator of malaria transmission intensity? *Trends in Parasitology* 23, 575-582.

[2] Sepúlveda N., Stresman G., White M. T., Drakeley C. (2015). Current Mathematical Models for Analyzing Anti-Malarial Antibody Data with an Eye to Malaria Elimination and Eradication. *Journal of Immunology Research* 2015:738030.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

USO DOS MODELOS IDADE-PERÍODO-COHORTE NO ESTUDO DA MORTALIDADE E INCIDENCIA POR INFARTO AGUDO DE MIOCARDIO EN GALICIA

G Naveira Barbeito[1], M Pérez Ríos[1] e A Malvar Pintos[1]

[1] Servizo de epidemioloxía, Dirección Xeral de Saúde Pública, Xunta de Galicia

RESUMO

O modelo idade-período-cohorte é unha ferramenta descritiva que permite descompoñer as taxas dun evento como o produto dos efectos da idade do individuo, do período do evento e da cohorte de nacemento, co fin de describir a súa tendencia temporal. Neste traballo aplicáronse estes modelos ao estudo da mortalidade e incidencia por infarto agudo de miocardio en Galicia. Como resultado desta análise compre destacar o efecto estimado para a idade, xa que as taxas de mortalidade e incidencia aumentan a medida que se cumpren anos; para a cohorte, o risco relativo para a mortalidade é decrecente ao longo das cohortes de nacemento mentres que para a incidencia o risco flutúa; e para o período, o risco relativo aumenta ata aproximadamente a metade do período a estudo e logo decrece, aínda que no caso da incidencia obsérvase un período de estabilidade nos últimos anos.

Palabras e frases chave: Modelos idade-período-cohorte, mortalidade, incidencia, infarto agudo de miocardio.

1. INTRODUCIÓN

O uso dos modelos idade-período-cohorte está amplamente estendido na epidemioloxía xa que permiten describir a tendencia temporal dun evento, como pode ser a mortalidade ou a incidencia por infarto agudo de miocardio (IAM), en función de:

- A **idade** do individuo ao producirse o evento. A probabilidade de ocorrencia dun evento como pode ser o infarto tende a aumentar a medida que se cumpren anos.
- O **período** de tempo no que se produciu o evento. O seu efecto relaciónase con cambios producidos nun determinado momento e que afectan de modo similar a todas as idades, como poden ser os avances médicos.
- A **cohorte** de nacemento da persoa. O seu efecto asóciase con cambios producidos a nivel xeracional, como poden ser cambios no patrón de hábito tabáquico nas mulleres novas en comparación coas de idades máis avanzadas.

O coñecido como "diagrama de Lexis" é unha ferramenta gráfica de gran utilidade de cara a representar aos individuos e a súa evolución no tempo a través das tres variables mencionadas. Este diagrama presenta un formato matriz no que en horizontal se representa o período, en vertical a idade e en diagonal a cohorte. A modo de exemplo, segundo o enfoque baseado na análise da mortalidade, a liña horizontal marcada en verde na figura 1 representa ás persoas que faleceron aos 40 anos de idade para cada ano do período 1980-2012; a liña vertical verde representa a todas as persoas falecidas no ano 1982 con idades comprendidas entre 35 e 84 anos; e a liña diagonal verde representa as defuncións sucedidas na cohorte de persoas que no ano 1990 tiñan 35 anos de idade.

Como se deduce ao visualizar o diagrama de Lexis, as tres variables atópanse relacionadas (cohorte=período-idade). Esta dependencia lineal leva consigo o denominado "problema de

identificabilidade" que se dá ao incorporar conxuntamente as tres variables nun modelo de regresión. Este problema consiste en que non existe unha única solución posible para o axuste, o que impide interpretar correctamente o efecto da idade, do período e da cohorte. Para solucionar este problema empregouse a metodoloxía proposta por Holford no ano 1983[1] que permite eliminar do modelo a tendencia lineal indistinguible do efecto período e cohorte (denominada "drift").

Neste traballo aplicáronse os modelos idade-período-cohorte co obxecto de estudar a evolución da mortalidade (período de estudo: 1980-2012) e da incidencia (período de estudo: 1996-2014) por IAM en Galicia, analizando os efectos asociados á idade, ao período e á cohorte. A incidencia analizouse en global e por separado para aqueles pacientes que presentaron elevación do segmento ST ao ingreso (grupo STEMI) e para os que non presentaron elevación do segmento ST (grupo NSTEMI), por tratarse de dúas manifestacións diferenciadas do infarto.

2. MATERIAL E MÉTODOS

A análise idade-período-cohorte [2-4] realizouse por medio do axuste dun modelo de regresión de Poisson no que a variable resposta foi o logaritmo das taxas específicas (λ), de mortalidade ou de incidencia, e as variables explicativas foron a idade, o período e a cohorte. Este modelo pode expresarse do seguinte xeito:

$$\log\big(\lambda(I,P)\big) = f(I) + g(P) + h(C)$$

Onde a idade (*I*) foi considerada en grupos quinquenais, o período (*P*) foi considerado en anos e a cohorte (*C*) obtívose a partir das dúas anteriores. Estas tres variables foron engadidas no modelo como variables continuas e foron suavizadas por medio de funcións splines cúbicas naturais (*f*, *g* e *h*, respectivamente), con 5 nodos interiores.

Para solucionar o problema de identificabilidade, ademais da extracción do "drift", elixiuse como cohorte de referencia (c_0) a mediana das cohortes de nacemento (ano 1936 para a mortalidade e ano 1945 para a incidencia) de modo que $h(c_o)=0$, e $g(P)$ foi fixado a 0 en promedio e con pendente 0. Deste xeito o efecto da idade estimado polo modelo debe ser interpretado como as taxas específicas para a cohorte de referencia, o efecto da cohorte debe ser interpretado como o risco relativo respecto á cohorte de referencia e o efecto do período como o risco relativo respecto á predición para a idade e a cohorte.

Levouse a cabo un proceso de modelización cara diante, de xeito que se axustaron modelos aniñados por adición secuencial das variables explicativas anteriormente descritas xunto co "drift". A selección do modelo realizouse por medio do test de razón de verosimilitudes, que permite contrastar a hipótese nula de que os coeficientes asociados ás novas variables engadidas ao modelo, en comparación co modelo sen elas, sexan iguais a cero. Os modelos axustados e as comparacións realizadas foron as seguintes:

Modelo I: Idade
Modelo Id: Idade+drift (comparación: *modelo Id* vs. *modelo I*)
Modelo IP: Idade+Período (comparación: *modelo IP* vs. *modelo Id*)
Modelo IC: Idade+Cohorte (comparación: *modelo IC* vs. *modelo Id*)
Modelo IPC: Idade+Período+Cohorte (comparación: *modelo IPC* vs. *modelo IP*)

Para todos os contrastes realizados consideráronse estatisticamente significativos os valores p inferiores a 0,05. O programa empregado foi Stata 12.1[5].

As fontes de datos utilizadas foron as seguintes:

Mortalidade: Instituto Nacional de Estatística para o período 1980-1986 e Rexistro de Mortalidade de Galicia dende 1987. Seleccionouse a rúbrica 410 da Clasificación Internacional de Enfermidades, 9ª revisión (CIE-9) ata o ano 1998, e a rúbrica I21 da CIE-10 dende o ano 1999, para a poboación residente en Galicia durante o período 1980-2012. A idade foi restrinxida ao grupo de 35 a 84 anos.

Incidencia: Conxunto Mínimo Básico de Datos de Altas Hospitalarias (CMBD-AH). Obtívose a información relativa aos ingresos por IAM ocorridos nos hospitais da rede pública galega e POVISA, tanto para aqueles con diagnóstico principal de IAM (rúbrica 410 da Clasificación Internacional de Enfermidades, 9ª revisión, Modificación Clínica (CIE-9 MC)) como para aqueles con diagnóstico secundario de IAM cando o diagnóstico principal fora unha complicación de IAM que coincide con algunha das seguintes rúbricas da CIE 9-MC: 411.0, 411.1, 414.0, 414.10, 423.0, 424.0, 426, 427.1, 427.41, 427.42, 427.5, 428.1, 428.9, 429.5, 429.6, 429.71, 429.79, 429.71, 429.79, 429.81, 518.4, 780.2 e 785.51. A partir desta información identificouse o primeiro ingreso de cada paciente durante o período 1996-2014, antepoñendo sempre o ingreso con diagnóstico principal de IAM naquelas situacións nas que

o primeiro ingreso do paciente se dera por IAM como diagnóstico secundario e existira un reingreso por IAM como diagnóstico principal. O grupo STEMI identificouse a partir das rúbricas 410.0-410.6 e 410.8; e o grupo NSTEMI das rúbricas 410.7 e 410.9. A idade foi restrinxida ao grupo de 35 a 84 anos.
Poboación: Estimacións intercensais calculadas pola Dirección Xeral de Saúde Pública para o período 1980-1997 e do Padrón Municipal de Habitantes dende 1998. A idade foi restrinxida ao grupo de 35 a 84 anos.

3. RESULTADOS

Como resultado do proceso de modelización cara diante e de aplicar o test de razón de verosimilitudes para a selección do modelo, tanto para a análise da mortalidade como para a da incidencia (en global e para os grupos STEMI e NSTEMI), chegouse á conclusión de que o mellor modelo en todos os casos era o que incluía todas as variables. Tras o axuste dos modelos completos estimáronse os efectos asociados a cada unha das variables explicativas (Figura 2). Estes efectos descríbense a continuación.

Efecto da idade:
De forma xeral, tanto para mortalidade como para incidencia, obsérvase que as taxas aumentan a medida que aumenta a idade. No caso da incidencia para o grupo STEMI as taxas estabilízanse a partir dos 60 anos de idade.

Efecto da cohorte:
Neste caso o comportamento difire substancialmente ao comparar a mortalidade coa incidencia en global. O risco relativo de morrer por IAM foi en descenso ao longo das cohortes de nacemento, mentres que, no caso da incidencia, obsérvase un comportamento cíclico de aproximadamente 30 anos que crece nas cohortes nacidas ata mediados da década de 1920 e, a continuación decrece nas cohortes nacidas ata mediados da década de 1940. Este comportamento repítese para as cohortes nacidas nos 30 anos seguintes pero de forma lixeiramente máis acusada.
Ao comparar os grupos STEMI e NSTEMI o comportamento do risco relativo resulta contraposto. O grupo STEMI presenta unha tendencia decrecente no risco relativo ao longo de todas as cohortes de nacemento, aínda que se aprecia un tramo relativamente estable que se corresponde ás cohortes nacidas entre 1945 e 1960. No grupo NSTEMI a tendencia é crecente.

Efecto do período:
No caso da mortalidade, ata finais da década de 1990 obsérvase unha tendencia crecente no risco relativo de morrer por IAM, tendencia que se inverte para anos sucesivos.
No caso da incidencia, ata mediados da década de 2000 obsérvase unha tendencia crecente no risco relativo de ingresar por primeira vez por IAM, tanto en global como para os grupos STEMI e NSTEMI, tendencia que se volve decrecente para anos sucesivos. Dende 2010 mantense relativamente estable en global e lixeiramente crecente no grupo STEMI.

4. CONCLUSIÓNS

O risco relativo de morrer por infarto agudo de miocardio ou de ingresar por primeira vez por esta causa en Galicia foi en descenso durante aproximadamente a última década, acadando, no caso da incidencia, un período de estabilidade. O risco relativo de ingresar ao longo das cohortes de nacemento compórtase de xeito oposto nos grupos STEMI e NSTEMI, decrecente no primeiro dos grupos e crecente no segundo.

REFERENCIAS

[1] Holford TR. The estimation of age, period and cohort effects for vital rates. Biometrics. 1983;39:311–24.

[2] Carstensen B. Age-period-cohort models for the Lexis diagram. Stat Med. 2007;26:3018-45.

[3] Armitage P. Age-period-cohort analysis. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons; 1998. pp. 82-99.

[4] Rutherford MJ, Lambert PC, Thompson JR. Age-period-cohort modeling. Stata Journal. 2010;10(4):606-27.

[5] StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.

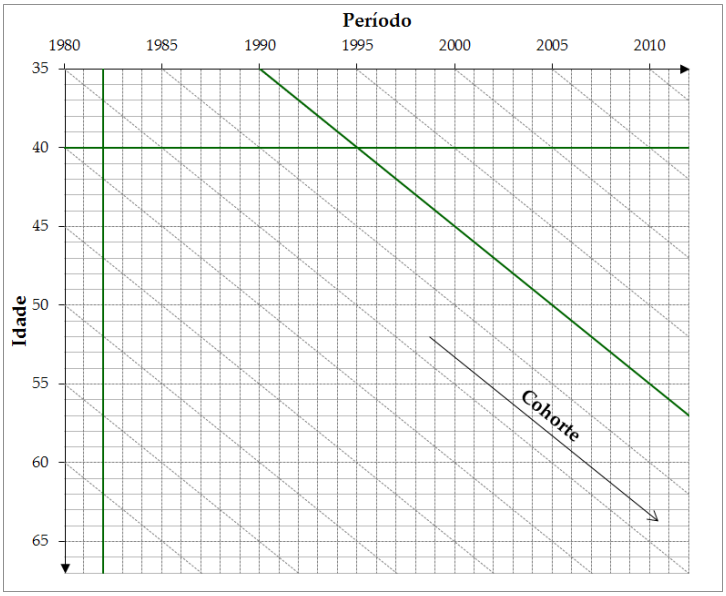


Figura 1: Diagrama de Lexis.

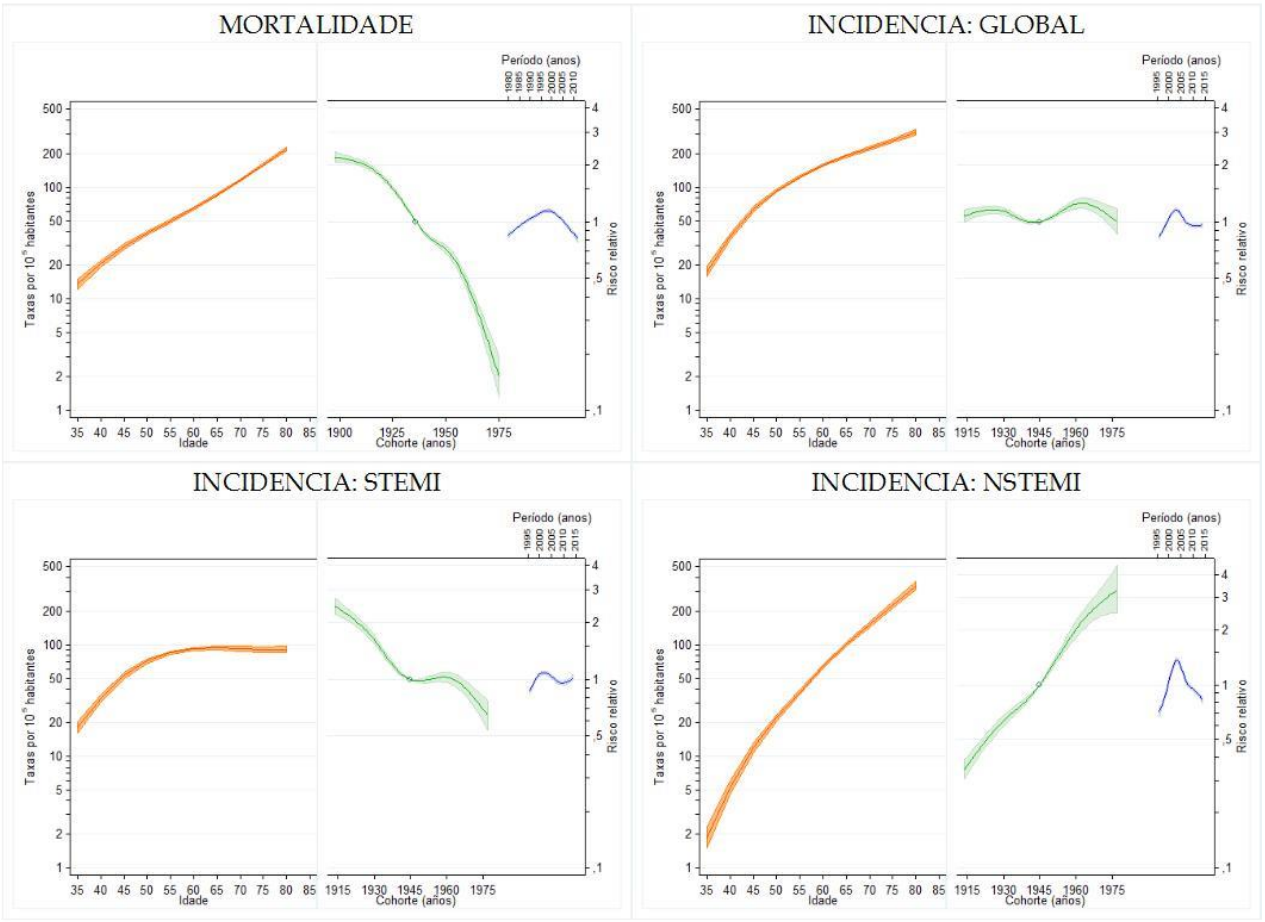


Figura 2: Análise idade-período-cohorte para a mortalidade e a incidencia, en global e para os grupos STEMI e NSTEMI, en Galicia: efectos estimados para a idade (taxa por 100.000 habitantes para a cohorte de referencia, representado en cor laranxa), a cohorte (risco relativo, representado en cor verde) e o período (risco relativo, representado en cor azul). Os valores preséntanse en escala logarítmica.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Analysis of pancreatic cancer in Spanish provinces combining Bayesian APC models and disease mapping approaches

Etxeberria J.[1,2,3], López-Abente G.[3,4], Riebler A.[5], Goicoa T.[1,2,6] and, Ugarte MD[1,2].

[1]Department of Statistics and O. R., Universidad Pública de Navarra.
[2]Institute for Advanced Materials (INAMAT), Universidad Pública de Navarra.
[3]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP).
[4]Cancer and Environmental Epidemiology Unit, National Center for Epidemiology, Carlos III Institute of Health.
[5]Norwegian University of Science and Technology.
[6]Research Network on Health Services in Chronic Diseases (REDISSEC).

SUMMARY

Pancreatic cancer is characterized by its high lethality. The reason is that during the early stages (when the tumor is more treatable) there are usually no symptoms and therefore this cancer is typically diagnosed at a late stage. This fact makes pancreatic cancer mortality a good surrogate of pancreatic cancer incidence.

It is generally known that pancreatic cancer mortality rates increase with age (as many other cancer locations) and that this cancer is more common in males than in females. Some other studies concluded that age-standardized mortality rates show an increasing trend in the last decades in most of the countries worldwide [2] and also in Spain [1]. However, the simple analysis of the evolution of pancreatic cancer mortality rates over different periods of time is not sufficient to identify possible determinants of the disease. In different time periods improved methods of diagnosis, treatments or, the disease classification may change disease trends. The cohort or the generation effects reflecting environmental and lifestyle risk factors should also be evaluated.

In this work, we provide a detailed analysis of pancreatic cancer mortality in Spanish provinces. In particular, we analyze the evolution of the geographical mortality pattern considering the effect of age, cohort, and gender in the analysis.

Based on a multivariate spatial age-period-cohort model proposed by Papoila, et al. (2009) [3] a set of models are considered as potential candidates to analyze pancreatic cancer mortality rates. The models combine features of APC models with disease mapping approaches. To ensure identifiability for gender-specific age, period, and cohort effects sum-to zero constraints were applied. A fully Bayesian approach based on integrated nested Laplace approximations (INLA) was considered for model fitting and inference [4]. Sensitivity analyses were also conducted.

Our results indicated that rates increase with age in both genders. The period effect showed a rising trend of pancreatic cancer mortality from 1990 to 2013, rates being higher in males than in females. The cohort effect showed an interesting pattern. The greatest difference between men and women were observed in those born before 1920. The younger the birth cohort, the smaller this difference was. The mortality showed a South-North pattern. The provinces in the North together with some in the

South-West were the ones with higher mortality rates during the studied period. The spatial geographical pattern of pancreatic cancer showed little changes along the period. Studies describing spatio-temporal evolution of pancreatic cancer mortality rates are scarce in our country and therefore this work could increase our understanding of the etiology of this disease and contribute to identifying target factors for primary prevention.

Key words: Disease mapping, Spatial Age-period-cohort models, Pancreatic mortality cancer.

ACKNOWLEDGMENTS

This work is supported by the Spanish Ministry of Economy and Competitiveness (Project MTM2014-51992-R) and by the Health Department of the Navarre Government (Project 113, Res.2186/2014)

REFERENCES

Referencias

[1] A. Cabanes, E. Vidal, N. Aragones, B. Perez-Gomez, M. Pollan, V. Lope, and G. Lopez-Abente. Cancer mortality trends in Spain: 1980-2007. *Annals of Oncology*, 21(Supplement 3):iii14–iii20, 2010.

[2] D. Hariharan, A. Saied, and H. M. Kocher. Analysis of mortality rates for pancreatic cancer across the world. *HPB*, 10(1):58–62, 2008.

[3] Ana L Papoila, Andrea Riebler, Antónia Amaral-Turkman, Ricardo São-João, Conceição Ribeiro, Carlos Geraldes, and Ana Miranda. Stomach cancer incidence in Southern Portugal 1998-2006: a spatio-temporal analysis. *Biometrical journal. Biometrische Zeitschrift*, 56(3):403–15, 2014.

[4] Havard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2):319–392, 2009.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

CRIMES EM PORTUGAL

Conceição Ribeiro[1], Paula Pereira[2]

[1]Instituto Superior de Engenharia-Universidade do Algarve/CEPAC/Centro de Estatística e Aplicações da Universidade de Lisboa
[2]Escola Superior de Tecnologia de Setúbal-Instituto Politécnico de Setúbal/Centro de Estatística e Aplicações da Universidade de Lisboa

RESUMO

O estudo da evolução da criminalidade, quer a nível temporal como espacial, apresenta-se de grande importância na definição de medidas que permitam melhorar o bem estar da população. Habitualmente, quando se pretende analisar a evolução da criminalidade numa dada região, recorre-se à comparação de taxas dos diversos anos. Com este trabalho pretende-se alargar essa análise e recorrer a modelos espaço-temporais que permitam tentar caracterizar a tendência da criminalidade, quer a nível espacial como temporal. Ou seja, tentar perceber se ao longo dos anos e em termos espaciais têm ocorrido alterações nos padrões de criminalidade. Para atingir estes objectivos, numa primeira fase será efectuada uma análise descritiva, temporal e espacial, da criminalidade em Portugal continental e, posteriormente, recorre-se-á aos modelos bayesianos hierárquicos. Para implementar estes modelos recorreu-se à metodologia INLA ("Integrated Nested Laplace Approximations").
Os dados analisados referem-se ao número de crimes detetados pelas autoridades policiais ou levado ao seu conhecimento por meio de denúncia ou queixa. Considera-se crime todo o facto descrito e declarado passível de pena criminal por lei anterior ao momento da sua prática. Das diversas vertentes do crime, este trabalho irá centrar-se em dois tipos de crimes: crimes contra pessoas e crimes contra o património. Nos crimes contra pessoas incluem-se, por exemplo, os homicidios, lesões corporias e maus tratos. Nos crimes contra o património incluem-se, por exemplo, furto, roubo, extorsão, violação de direitos de autor ou marca. Em relação a cada crime tem-se informação sobre a localização (concelhos de Portugal continental) e ano (2011 a 2015) em que ocorreram.

Palabras e frases chave: Modelos bayesianos hierárquicos, crimes, pequenas áreas, espaço-temporal.

AGRADECIMENTOS

O trabalho é financiado pela FCT - Fundação para a Ciência e a Tecnologia, Portugal, através do projecto UID/MAT/00006/2013.

REFERÊNCIAS

Referencias

[1] Knorr-Held L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics In Medicine*, 19(17–18), 2555–2567.

[2] Knorr-Held L., Besag J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, 17(18), 2045–2060.

[3] Law J., Quick M., Chan P. (2014). Bayesian Spatio-Temporal Modeling for Analysing Local Patterns of Crime Over Time at the Small-Area Level. *Journal of Quantitative Criminology*, 30, 57–78, DOI 10.1007/s10940-013-9194-1.

[4] Law J., Quick M., Chan P. (2015). Analyzing Hotspots of Crime Using a Bayesian Spatiotemporal Modeling Approach: A Case Study of Violent Crime in the Greater Toronto Area. *Geographical Analysis*, 47(1), 1–10, DOI 10.1111/gean.12047.

[5] Lawson A.B. (2009). *Bayesian Disease Mapping Hierarchical Modeling in Spatial Epidemiology*. CRC Press, New York.

[6] Papoila A.L., Riebler A., Amaral Turkman, M.A., São João R., Ribeiro C., Geraldes C., Miranda A. (2014). Stomach cancer incidence in Southern Portugal 1998–2006: a spatio-temporal analysis. *Biometrical Journal* 56 (3): 403–415.

[7] Ribeiro C., Amaral Turkman M.A., Cardoso J.C. (2011). Bayesian Hierarchical Models: An analysis of Portugal Road Accident Data. *Proceedings of the Spatial Data Methods for Environmental and Ecological Processes, 2nd Edition*, 101-104 (B.Cafarelli, ed.), Italy.

[8] Ribeiro C., (2013) *Modelação de Dados EspaçoTemporais em Segurança Rodoviária*, Phd tesis, Universidade de Lisboa, Portugal.

[9] Rue H., Martino S., Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B* 71(2), 319–392.

Comunicacións orais

Sesión VII: Métodos estatísticos en Medicina II

II Encontro Galaico-Português de Biometría
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

PROFILING PROTEÓMICO DE ISOLADOS DE *ESCHERICHIA COLI* E *ENTEROCOCCUS* SPP. ATRAVÉS DA ESPECTROMETRIA DE MASSA

Tiago Santos [1-3], Irene Oliveira[4], José Luis Capelo[5,6] , Hugo M. Santos[5,6], Catarina Marinho[1-3], Alexandre Gonçalves[1-3], José Eduardo Araújo[6], Patrícia Poeta[3,6], Gilberto Igrejas[1,2,6]

[1] Unidade de Genómica Funcional e Proteómica, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
[2] Departamento de Genética e Biotecnologia, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
[3] Departamento de Ciências Veterinárias, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
[4] Departamento de Matemática e Centro de Investigação e Tecnologias Agro-ambientais e Biológicas, Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal
[5] Sociedade Científica ProteoMass, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[6] UCIBIO-REQUIMTE, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

RESUMO

A espetrometria de massa com dessorção/ionização da amostra assistida pela ação da matriz/laser e leitura do tempo de voo (MALDI-TOF MS) é um dos métodos usados em proteómica para identificação e caracterização de bactérias e surge atualmente como uma opção eficiente na identificação rápida de espécies bacterianas comensais e patogénicas por avaliação de picos de massa específicos presentes ao nível do género, espécie e subespécie bacteriana.
Por meio desta técnica caracterizaram-se isolados fecais de *Enterococcus* spp. e *Escherichia coli* de aves selvagens do Arquipélago dos Açores. Para a análise proteómica foram usadas células bacterianas intactas provenientes de 60 isolados de *Enterococcus* spp. (18 *Enterococcus faecalis*, 15 *E. faecium*, 15 *E.durans* e 12 *E. hirae*) e 60 isolados de *Escherichia coli*, de amostras fecais de aves dos Açores.
Procedeu-se ao processamento dos espectros de massa (MS) para identificação dos picos de massa discriminantes na separação dos isolados das espécies distintas. Usaram-se metodologias multivariadas de redução de dimensão, Análise em Componentes Principais (ACP), e de classificação não supervisionada, Análise de *Clusters* (AC), e de classificação supervisionada, Análise Discriminante Linear (ADL), para a identificação de agrupamentos de isolados com caracterização similar a nível da informação dos picos de massa.
Foram avaliados os picos com potencial para biomarcadores específicos destes géneros bacterianos.

Palavras-chave: *Enterococcus* spp., *Escherichia coli*, espectrometria de massa, MALDI-TOF MS, proteómica.

1. INTRODUÇÃO

As análises laboratoriais de diagnóstico exigem a pesquisa constante e atualização de ferramentas bioinformáticas que sejam cada vez mais rápidas e precisas na caracterização e classificação de microrganismos, quer tenham como base a informação genómica, proteómica ou metabolómica.
No caso de estudos direcionados para a identificação de espécies bacterianas, a análise dos espetros de massa (MS) pode ser realizada pela comparação de espécies iónicas obtidas na amostra e os picos de massas referenciados nas bases de dados já existentes (biomarcadores), ou comparando os perfis de espetros de massa de todas as amostras analisadas[1]. Por exemplo, uma das metodologias que tem vindo a ser cada vez mais usada como um método expedito e preciso na identificação de bactérias intactas, em comparação com os métodos convencionais de microbiologia e/ou biologia molecular, é a espetrometria

de massa com dessorção/ionização da amostra assistada pela ação da matriz/laser e leitura do tempo de voo (MALDI-TOF MS). Quintela-Baluja e colaboradores [2] realizaram o estudo de isolados de espécies bacterianas via MALDI-TOF MS para separação de isolados bacterianos de espécies distintas do género *Enterococcus* spp. e em Santos e colaboradores [3] descreve-se em maior pormenor a análise realizada para separação de isolados bacterianos de espécies distintas do género enterococos de origem animal. Neste estudo o foco foi a separação a nível espetrometria de massa de isolados de *Enterococcus* spp. e *Escherichia coli* sem o recurso a bases de dados de referência, mas sim a um protocolo extenso de tratamento bioinformático e análise estatística.

2. MATERIAL E MÉTODOS

Para a análise proteómica, via MALDI-TOF MS, foram utilizadas células bacterianas intactas provenientes de 60 isolados de *Enterococcus* spp. e 60 isolados de *Escherichia coli* de aves selvagens do Arquipélago dos Açores.

Os espectros de todas amostras foram analisados usando o pacote MALDIquant do software R [4]. No pacote MALDIquant procedeu-se a análise de dados de espectrometria de massa, possibilitando uma implementação de rotina para o tratamento dos espetros de massa não processados (mzXML), desde a importação, passando pelo pré-processamento, até ao alinhamento e calibração.

A análise de 120 espectros de massa tratados foi obtida pelo pré-processamento de 1186 espectros, provindos da análise de espetrometria de massa com 10 réplicas de cada um dos 120 isolados. Procedeu-se ao processamento tendo em conta os seguintes passos: transformação das intensidades pela raiz quadrada; remoção da linha de base pelo método de Snip; normalização dos espectros pelo método TIC (*total ion current*); alisamento pelo algoritmo de *Savitsky-Golay* e deteção de picos tendo em conta a razão sinal-ruído (*signal to noise*) de 3.

Os espectros alinhados e processados foram de seguida sujeitos a metodologias de análise estatística multivariada, nomeadamente com Análises em Componentes Principais (ACP) e de classificação não supervisionada, Análise de *Clusters* (AC) por espécie e no total e de classificação supervisionada, Análise Discriminante Linear (ADL), para a identificação de agrupamentos de isolados com caracterização similar a nível da informação dos picos de massa. Consideraram-se, ainda, análises alternativas aos dados com base em preditores binários [5].

3. DISCUSSÃO E CONCLUSÃO

Através de uma análise prévia, sem processamento de espetros de massa dos isolados, foram claramente identificados, ao nível de espécie, biomarcadores (picos de massa) conhecidos previamente nos perfis de estirpes referência. Os resultados da análise de *clusters* e da ACP evidenciaram a separação dos dois géneros em estudo, *Escherichia e Enterococcus* quer na análise dos dados processados, quer na análise de dados binários (Fig. 1). Contudo não foi visualizada a separação de espetros de massa provindos das 4 diferentes espécies do género *Enterococcus*. Adicionalmente, na ACP sem a presença dos isolados de *Escherichia coli*, não foi obtida a separação entre os isolados de espécies diferentes do género enterococos .

No total foram obtidos 215 picos distintos, entre eles o pico de massa *m/z* 4428 ± 3, encontrado unicamente em isolados de enterococos, tal como em [2], e outros dois picos (*m/z* 5379 ± 3 e 6253 ± 3) detetados apenas em espetros dos isolados de *Escherichia*, sugerindo que estes picos possam representar biomarcadores específicos destes géneros bacterianos, respetivamente. No caso de três das quatros espécies de enterococos em análise, foi, também, possível identificar picos de massa característicos das mesmas.

Os resultados obtidos demonstram o potencial do MALDI-TOF MS para a rápida e eficaz identificação de isolados de *Enterococcus* spp. e *Escherichia coli*, ao nível de género e até na possibilidade de distinguir ao nível das diferentes espécies bacterianas pertencentes ao grupo enterococos (*Enterococcus faecium*, *Enterococcus faecalis*, *Enterococcus durans* e *Enterococcus hirae*) tendo detetado picos de massa específicos dos dois géneros estudados e picos de massa caracterizadores das diferentes espécies do género enterococos. Existem múltiplas vantagens na aplicação da espectrometria de massa em identificação bacteriana associada a um rigoroso processamento e análise estatística dos espetros de massa.

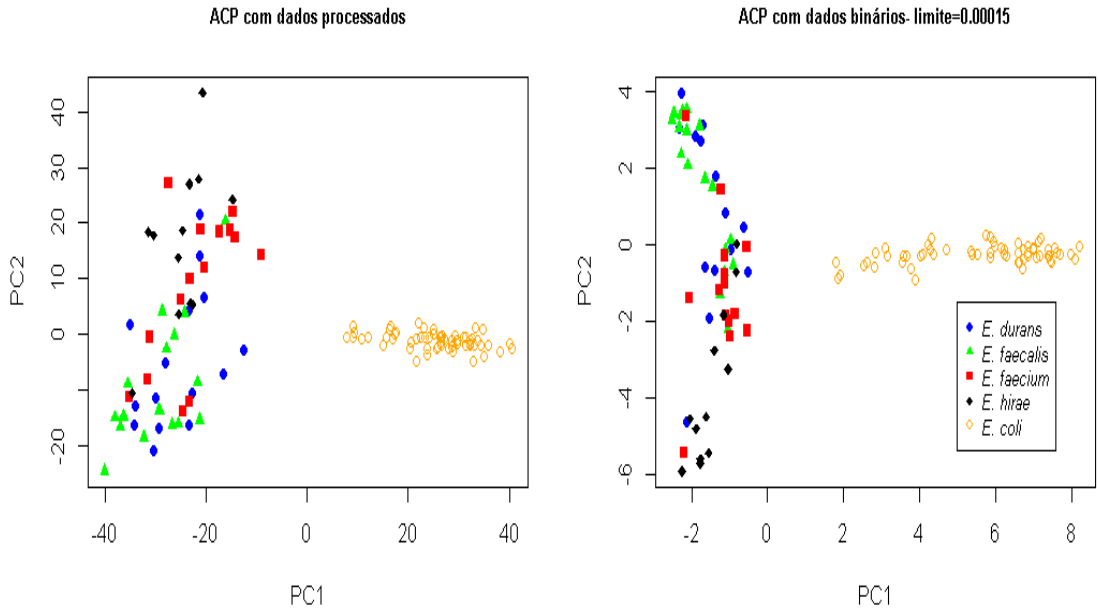


Figura 1: Scores das componentes PC1 vs PC2 das análises ACP realizada aos dados.

AGRADECIMENTOS

Os autores gostariam de agradecer à Unidade de Ciências Biomoleculares Aplicadas-UCIBIO financiada através do projeto FCT/MEC (UID/Multi/04378/2013) e co-financiada pelo ERDF através do PT2020 Partnership Agreement (POCI-01-0145-FEDER-007728). E ainda agradecer ao CITAB financiado por Fundos Europeus de Investimento através do FEDER/COMPETE/POCI - Programa de Competitividade e Internacionalização, no âmbito do projeto POCI-01-0145-FEDER-006958 e Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto PEst-OE/AGR/UI4033/2013.

REFERÊNCIAS

[1] Fenselau, C., Demirev, P.A. (2001) Characterization of intact microorganisms by MALDI mass spectrometry. *Mass spectrometry reviews* 20, 157-71.
[2] Quintela-Baluja M, Bohme K, Fernandez-No IC, Morandi S, Alnakip ME, Caamano-AnteloS, et al. (2013) Characterization of different food-isolated Enterococcus strains by MALDITOF mass fingerprinting. *Electrophoresis* 34, 2240-50.
[3] Santos, T., Capelo, J. L., Santos, H., Oliveira, I.; Marinho, C., Gonçalves, A., Araújo, J.E., Poeta, P.; Igrejas, G. (2015) Use of MALDI-TOF mass spectrometry fingerprinting to characterize Enterococcus spp. and Escherichia coli isolates, *Journal of Proteomics* 127 B, 321-31
[4] Gibb S, Strimmer K. (2012) MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*. 28(17) 2270-1.
[5] Gibb, S., K. Strimmer (2015) Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis. *Bioinformatics* 31, 3156-62.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

ESTIMAÇÃO DE EQUAÇÕES DE REFERÊNCIA PARA PARÂMETROS ESPIROMÉTRICOS PARA A POPULAÇÃO DE VILA FRANCA DE XIRA

Carina Silva 1[1], Sandra Carvalho 2[2], Lisete Sousa 3[3]

[1]Escola Superior de Tecnologia da Saúde de Lisboa e CEAUL
[2]Escola Superior de Tecnologia da Saúde de Lisboa e Hospital de Vila Franca de Xira
[3]Faculdade de Ciências da Universidade de Lisboa e CEAUL

RESUMO

A espirometria é um exame de rotina utilizado no diagnóstico e controlo de doenças pulmonares. Os parâmetros espirométricos dependem de múltiplos fatores nomeadamente da altura, da idade e do género. A interpretação dos resultados da espirometria é normalmente realizada com base na comparação entre os valores espirométricos e os valores de referência que derivam de equações de referência. Estas equações são obtidas a partir de grupos populacionais representativos, constituídos por indivíduos saudáveis (sem diagnóstico de doença pulmonar e não fumadores). Não existem equações adaptadas às características da população portuguesa, normalmente são utilizados parâmetros de referência para a população europeia. Esta necessidade também se justifica face ao progresso inerente às técnicas, equipamento e protocolos. Nesse sentido, pretende-se estimar equações de referência para os parâmetros espirométricos adaptadas à população de Vila Franca de Xira. As equaçõees foram estimadas utilizando o método GAMLSS (Generalized Additive Models for Location Scale and Shape). Foram consideradas como variáveis explanatórias a idade, altura e género para além das possíveis interações. Consideraram-se as distribuições normal e Box-Cox-Cole-Green. As equações de referência foram comparadas com as mais utilizadas na Europa e na América do Norte. A análise foi realizada com recurso às bibliotecas `gamlss` e `gamlss.tr` do `R`.

Palabras e frases chave: Espirometria, GAMLSS, equações de referência.

AGRADECIMENTOS

Trabalho parcialmente financiado por fundos nacionais através da Fundação para a Ciência e a Tecnologia, Portugal, através do projeto UID/MAT/00006/2013.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Modeling long-memory and volatility in Heat Rate Variability

Argentina Leite[1], Ana Paula Rocha[2] e Maria Eduarda Silva[3]

[1]Escola de Ciências e Tecnologia, Universidade de Trás-os-Montes e Alto Douro & C-BER & INESC TEC, Portugal `tinucha@utad.pt`
[2]Faculdade de Ciências, Universidade do Porto & CMUP, Portugal `aprocha@fc.up.pt`
[2]Faculdade de Economia, Universidade do Porto & CIDMA, Portugal `mesilva@fep.up.pt`

ABSTRACT

This work focus on the application of time series models in the analysis of Heart Rate Variability (HRV), considering Fractionally Integrated AutoRegressive Moving Average (ARFIMA) models with heteroscedastic errors. To model the heterocedascity nonlinear Generalized Autoregressive Conditionally Heteroscedastic (GARCH) and Exponential Generalized Autoregressive Conditionally Heteroscedastic (EGARCH) models are considered. These models are used to capture and remove long memory and characterize conditional volatility in 24 hour HRV recordings from the Noltisalis database.

Keywords and key sentences: HRV, ARFIMA, GARCH, EGARCH

1. INTRODUCTION

Heart Rate Variability (HRV) reflects the interaction between perturbations to the cardiovascular variables and the corresponding response of the cardiovascular regulatory systems [1]. The modeling of such variability can provide a quantitative and non-invasive method to assess the integrity of the cardiovascular system. HRV data display non stationary characteristics and exhibit long memory and time-varying conditional variance (usually designated by volatility) which may contain indicators of current disease or warnings about impending diseases. Traditionally, HRV data can be characterized by linear AutoRegressive (AR) models, which describes only short memory in the mean. These models combined with recursive least squares have been used to estimate the volatility in HRV data. An alternative approach to describe the dynamics in HRV data was proposed by Leite *et al* [2, 3] using Fractionally Integrated AutoRegressive Moving Average (ARFIMA) models with heteroscedastic errors. These models which are an extension of the AR models usual in the analysis of HRV and may be used to capture and remove long memory and estimate the volatility in 24 hour HRV recordings [3], satisfy the following equations

$$\begin{aligned} \phi(B)(1-B)^d x_t &= \theta(B)\epsilon_t && (1)\\ \epsilon_t &= \sigma_t z_t && (2)\\ \sigma_t^2 &= \mathrm{Var}(\epsilon_t|H_{t-1}) && (3) \end{aligned}$$

where B is the backward-shift operator. Equation (1) describes the conditional mean of the process with serially uncorrelated residuals ϵ_t and is said an ARFIMA(p, d, q) where d, the long-memory

parameter, determines the long-term behaviour in mean. Equations (2) and (3) describe the conditional variance of the process which varies over time as in time-varying AR models. In (2), ϵ_t are called shocks and z_t, independent and identically distributed random variables with zero mean and unit variance are the standardized shocks.
When

$$\sigma_t^2 = u_0 + \sum_{i=1}^{P} v_i \, \sigma_{t-i}^2 + \sum_{i=1}^{Q} u_i \, \epsilon_{t-i}^2 \tag{4}$$

with $u_0 > 0$, $v_1, ..., v_P$, $u_1, ..., u_Q \geq 0$, $\sum_{i=1}^{P} v_i + \sum_{j=1}^{Q} u_j < 1$, the model is called GARCH(p, q). The parameters, u_i and v_i characterize the volatility clustering phenomena observed in many data sets. Alternatively to the GARCH model, we consider the EGARCH(P, Q) [4] defined as follows:

$$\log\sigma_t^2 = u_0^* + \sum_{i=1}^{P} v_i \log\sigma_{t-i}^2 + \sum_{i=1}^{Q} u_i \, |z_{t-i}| + \sum_{i=1}^{Q} \xi_i z_{t-i} \tag{5}$$

where $u_0^* = u_0 - \sum_{i=1}^{Q} u_i \sqrt{\frac{2}{\pi}}$ and $z_t = \epsilon_t/\sigma_t$. The parameters u_i and v_i characterise the volatility clustering phenomena and the parameters ξ_i describe the leverage effects. The impact of positive shocks, $\epsilon_{t-i} > 0$ is $(u_i + \xi_i)\frac{\epsilon_{t-i}}{\sigma_{t-i}}$, while for negative shocks it is $(u_i - \xi_i)\frac{\epsilon_{t-i}}{\sigma_{t-i}}$. If $\xi_i = 0$ $\log\sigma_t^2$, responds symmetrically to ϵ_{t-i}.
In this work, we apply ARFIMA$(p, d, 0)$-GARCH$(1, 1)$ and ARFIMA$(p, d, 0)$-EGARCH$(1, 1)$ models to 24 hour HRV recordings from the Noltisalis database [5]: an healthy subject-N5 (N), a patient suffering from congestive heart failure-C2 (C) and a heart transplanted patient-T3 (T).

2. MODELING HRV

For illustrative purposes we first model two short records: an healthy patient (segment of RR series with 1200 beats for subject-N5) and a patient suffering from congestive heart failure (segment of RR series with 1100 beats for subject-C2). The results are presented in Table 1.
The higher value for the long-memory parameter for the patient C is already reported in the literature [3]. The results also indicate that both patients present heteroscedasticity, indicated by significant values for $\hat{u}_1$ and $\hat{v}_1$. These results are in accordance with results in [3] which indicate that heterocedasticity is stronger in normal patients. Furthermore, the estimates for $\hat{\xi}_1$ indicate asymmetric response to positive and negative shocks which is, also, stronger in normal patients.

Parameter	**Healthy-N5 (N)**		**Patient-C2 (C)**	
	ARFIMA GARCH	ARFIMA EGARCH	ARFIMA GARCH	ARFIMA EGARCH
$\hat{d}$	0.31		0.62	
$Q(20)$ of $\hat{\epsilon}_t$	101.13		36.76	
$CCF(1)$ of $\hat{\epsilon}_t$,$\hat{\epsilon}_t^2$	0.25		0.14	
$\hat{u}_0$	$0.24\text{e}^{-4}(0.00\text{e}^{-5})$	-2.83 (0.30)	$0.42\text{e}^{-6}(0.81\text{e}^{-7})$	-9.05 (1.85)
$\hat{u}_1$	0.36 (0.03)	0.32 (0.05)	0.16 (0.04)	0.17 (0.06)
$\hat{v}_1$	0.44 (0.05)	0.70 (0.03)	0.07(0.16)	0.08(0.19)
$\hat{\xi}_1$	—	0.32 (0.03)	—	0.19(0.04)
$Q(20)$ of $\hat{z}_t$	6.80	10.34	5.98	9.74
$CCF(1)$ of $\hat{z}_t$,$\hat{z}_t^2$	0.14	0.010	0.07	0.01
AIC of $\hat{z}_t$	-7784.90	-7881.60	-76771.2	-7.686.8

Table 1: Estimates (standard deviation) for two short HRV records: healthy-N5 (N) with 1200 beats and patient-C2 (C) with 1100 beats. $Q(20)$ stands for McLeod-Li' s test statistic for the residuals. $CCF(1)$ is the lag one cross-correlation between the residuals and their squared values.

In the case of long recordings, such as ambulatory 24 hour HRV data (approximately 100000 beats), exhibiting several non stationary characteristics, ARFIMA-EGARCH modeling is combined with adaptive segmentation [3]: long records are decomposed into short records of variable length and the break points, which mark the end of consecutive short records, are identified by AIC criterion. The short records thus obtained have a minimum length 512 and are subsequently modeled using ARFIMA-EGARCH models.

The results for the 24 hour HRV series of 3 subjects are summarized in Fig. 1. Long memory characteristic is stronger in the sick subjects as found previously in the literature, [3]. The McLeod-Li test statistic for conditional heteroscedasticity, $Q(20)$ indicates that the percentage of segments with volatility is 98.9% for the normal subject, N, 88.7% for C and 82.2% for T. The volatility parameter u_1 is lower for sick subjects C and T. The parameter v_1 presents high variability for the sick subjects and requires further study. The values of the lag one cross-correlation between the residuals of ARFIMA and their squared values, $CCF(1)$, indicate presence of asymmetry in the volatility. For patient N there are 96.8% such segments while for C and T there are only 43.8% and 46.5%, respectively. Thus, the sick patients present lower values for the leverage parameter ξ_1 and consequently less asymmetry in HRV.

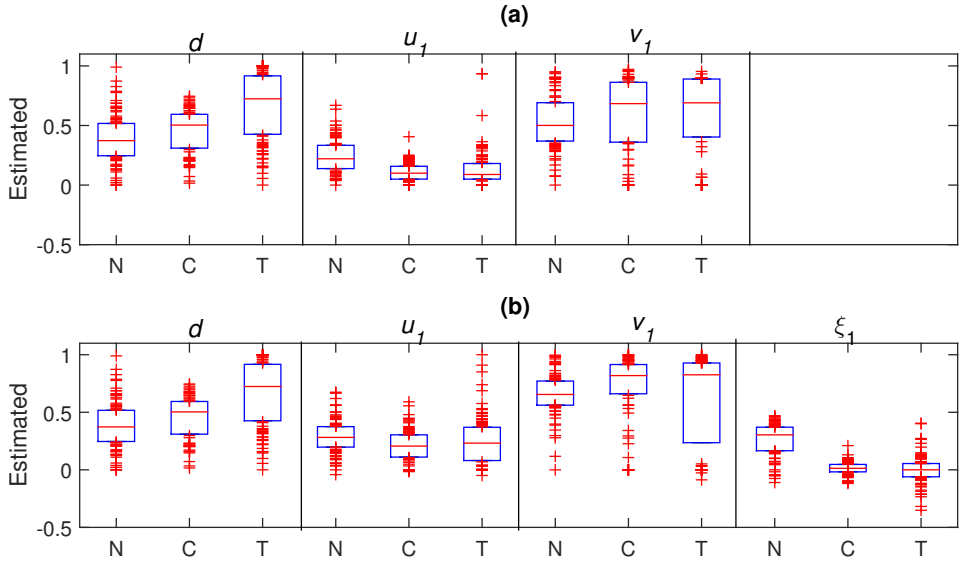


Figure 1: Boxplot for estimates $\hat{d}$, $\hat{u}_1$, $\hat{v}_1$ and $\hat{\xi}_1$ using (a) ARFIMA-GARCH and (b) ARFIMA-EGARCH models and segmentation for 24 hour HRV data of the three subjects: an healthy subject-N5 (N), a patient suffering from congestive heart failure-C2 (C) and a heart transplanted patient-T3 (T).

ACKNOWLEDGMENT

This work is partially funded by ERDF and COMPETE 2020 Programme project POCI-01-0145-FEDER-006961, CIDMA UID/MAT/04106/2013, CMUP UID/MAT/00144/2013, and by FCT (Portugal) with national (MEC) structural funds, under the partnership agreement PT2020.

REFERENCES

References

[1] Task Force of the European Society of Cardiology and North American Society of Pacing Electrophysiology (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation* 93, 1043-1065.

[2] Leite A., Rocha A. P., Silva M. E. (2013). Long memory and volatility in HRV: an ARFIMA-GARCH approach. In: *Computing in Cardiology* Vol 36, 165-168.

[3] Leite A., Rocha A. P., Silva M. E. (2013). Beyond long memory in heart rate variability: an approach based on fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity. *Chaos*, 23, 023103.

[4] Tsay R. S. (2005). *Analysis of Financial Time Series (2nd ed.)*. Wiley-Interscience.

[5] Signorini M. G., Sassi R., Cerutti S. (2001). Working on the NOLTISALIS Database: Measurement of nonlinear properties in heart rate variability signals. In: *Proceedings of IEEE-EMBS*, 547-550, International Conference, Istanbul, Turkey (IEEE, Piscataway).

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Performance of adaptive designs in dose finding experiments

Nancy Flournoy[1], José Moler[2], Fernando Plo[3]

[1] University of Missouri. Department of Statistics. USA
[2] *Corresponding author:* Department of Statistics and Operations Research, Universidad Pública de Navarra, Pamplona, Spain. jmoler@unavarra.es
[3] Department of Statistical Methods, Universidad de Zaragoza, Zaragoza, Spain.

RESUMEN

In this work we present several tools to study the performance of adaptive designs in phases I or II of Clinical Trials. In the early phases of a clinical trial, the main statistical target is to estimate percentiles in response variables in order to control toxicity and efficiency of new drugs or, even, new medical procedures. We also analyze the way to combine inferential and ethical criteria to evaluate the performance of some relevant adaptive procedures presented in the specialized literature.

Palabras e frases clave: clinical trials, adaptive designs, optimal designs.

1. INTRODUCCIÓN

Dose finding experiments are statistical experiments that focus on the estimation of the dose for which a targeted percentage of individual responses are successful. In the context of clinical trials, to develop a new drug, the first step consists in establishing the maximum tolerated dose for which the rate of toxicity reaches a targeted value. These experiments are called Phase I clinical trials. In Phase II clinical trials the target is to establish the minimum efficacy dose for which the rate of efficacy is over a targeted value, more details can be seen, for instance, in [9] or [3]. Even though early phases of clinical trials are a relevant context where dose finding experiments are carried out, many other areas use them. For instance, in the context of psychometrics and education, item response theory (IRT) calibrates the difficulty of an item in a test or questionnaire. Subjects with different ability give responses to the item and depending on the rate of successful responses the difficulty is estimated. A detailed explanation is given in [2], where the subject ability takes the role of doses in clinical trials.
In what follows, we will assume the context of clinical trials, and the terms will be used in this framework. We assume that L different treatments or dosages can be applied upon the arrival of each patient, $d_1 < d_2 < \ldots < d_L$. For the nth patient in the trial we will consider the following notation

$$Y_n = \begin{cases} 1, & \text{success response;} \\ 0, & \text{failure response.} \end{cases} \qquad \delta_{nj} = \begin{cases} 1, & \text{dose } d_j \text{ applied;} \\ 0, & \text{otherwise.} \end{cases} \qquad j = 1, \ldots, L.$$

So that, Y_n and $\delta_n = (\delta_{n1}, \cdots, \delta_{nj})$ are, respectively the response and the dose received by the nth patient. Under the assumption of a model for the response, there exists a function F such that:

$$E[Y_n|\delta_n] = F(\beta, \delta_n) \tag{1}$$

where β are the coefficients of the model.
For economic reasons and, in clinical trials, also for ethical reasons, small sample sizes are searched. Besides, the number of participants with doses far from the target dose should be as small as possible. This considerations call for adaptive designs, which use the information obtained during the experiment to allocate the following patient as close as possible to the target dose. More formally, let

$$\mathcal{F}_n = \sigma(Y_j, \delta_j : j \leq n)$$

the accrued information up to the nth patient, then, a randomized adaptive allocation rule is

$$\pi_{ni} := P(\delta_{nj} = 1 | \mathcal{F}_{n-1}). \tag{2}$$

Many randomized adaptive allocation rules have been proposed in the literature. Depending on the criterion chosen, several classifications can be made. For instance, when only the information of the k most recent participants in the trial, the allocation rule is said of *short memory*; otherwise is said of *long memory*. Another classification can be made depending on whether the design assumes a parametric model for the responses and use it for future allocations, then the rule is said *model based design*; otherwise is said *algorithm based design*. Observe that the design can be used just for allocating patients but, independently, a parametric model can be assumed for the responses for estimation purposes. A deep review is provided in [8].
There are three problems, clearly interrelated, that have been recently raised in the specialized literature.
The first one deals with the tradeoff between ethics and estimation. Ethical issues appear when the design try to avoid the exposition of patients to undesired dose levels but this affects the amount of information (quality of the sample, for instance) and, so, the right estimation of the target percentile. In [1] they consider the dilemma between *treatment*, that is, last patient in the experiment receives the updated estimation of the target dose, and *experimentation* that is, good estimation of the target percentile. *Treatment* is an ethical issue because this approach is concerned about the assignment of patients to the dose closest to the target percentile. They prove that there are dose-finding designs that will concentrate on the right dose from some time on. In contrast, they observe that algorithm based designs are not concerned by *treatment* purposes and can provide consistent estimator although many patients will be allocated in irrelevant doses. Finally, they conclude that more research on how to balance *treatment* and *experimentation* is needed. In this paper we tackle this problem by proposing several criteria to measure *treatment* and *experimentation*. That is, ethical issues and quality of estimations.Then, a procedure to merge these criteria is given with the goal of giving a sole measure which contains ethical measures along with inferential ones. The proportion of patients exposed to doses far from the target is a measure of the ethics and the mean square error (MSE) measures the ability of one estimator to make right inferences. Both measures are natural candidates in practice to measure ethical and inferential criteria, respectively.
The second problem refers to the consistency of the designs. Sometimes, dose-finding designs are presented without convergence studies because they use small sample sizes and, so, their asymptotic results are considered useless. However, as stated in [6], most of the dose-finding experiments rely in a self-correcting mechanism which concentrates assignments of treatments around the target but if the design has not a stationary behavior with desirable properties, the self-correcting mechanism is of limited utility. In [4], they prevent the users against the best-intention designs, those that allocate the present patient to the dose believed to be best, because they lead to allocations that converge to the wrong point with non-zero probability and, in some cases, almost surely. In this work, we first summarize asymptotic results for dose-finding experiments and we make a comparative study among the consistency of the estimators.
Finally, the third problem deals with the choice of a good estimator given a design. Unfortunately, there is not a criterion to determine which is the best one because the performance of an estimator depends on the design chosen. Besides this, two different estimators can be compared and their performances can be quite different depending on the design applied so that, one of them can not be chosen as the best uniformly. Comparative studies have been carried out to choose estimators for a percentile estimation, for instance, [7] and [5], but conclusive results can not be obtained because there is not a whole study that combines all the available designs with all the candidate

estimators. Besides this, these comparative studies are not based on theoretical and asymptotic properties but on simulation studies for small sample sizes.
In this paper, we tackle the first problem combining measures of ethics and inference. But, the study of these measures in the limit is clearly related with the other two problems. The asymptotic behavior of the ethical measure is the study of the consistency of the design, whereas the asymptotic behavior of the estimation measure is related with the quality of the estimator chosen.

Referencias

[1] Azriel, D., Mandel, M. and Rinnot, Y. 2011. The treatment versus experimentation dilemma in dose finding studies. *Journal of Statistical Planning and Inference,* 141, 2759–2768.

[2] Chang, H. and Ying, Z. 2009. Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics*, 37 (3), 1466–1488.

[3] Chevret, Sylvie (Editor) 2006. *Statistical methods for dose-finding experiments* . West Sussex: Wiley.

[4] Fedorov, V., Flournoy, N., Wu, Y. and Zhang, R. 2011. Best intention designs in Dose-Finding studies. *Technical Report, Isaac Newton Institute for the Mathematical Sciences, Cambridge, UK.* http://www.newton.ac.uk/preprints/NI11065.pdf.

[5] Ivanova, A., Montazer-Haghighi, A., Mohanty, S.G. and Durham, S. D. 2003. Improved up-and-down designs for phase I trials. *Statistics in Medicine*, 22, 69–82.

[6] Oron, A., Azriel, D. and Hoff, P. 2011. Dose-Finding designs: the role of convergence properties. *The International Journal of Biostatistics,* 7, 1, 1–17.

[7] Stylianou, M. and Flournoy, N. 2002. Dose finding using the biased coin up-and-down design and Isotonic Regression. *Biometrics*, 58, 171–177.

[8] Sverdlov, O., Wong, W.K. and Ryeznik, Y. (2014) Adaptive clinical trial designs for phase I cancer studies *Statistical surveys*, 8, 2–44

[9] Ting, N. (Editor) 2006. *Dose finding in drug development.* New York: Springer.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Economical evaluation of computed tomography angiography (CTA) versus conventional angiography (CA) to diagnose Coronary ischemia.

Reyes-Santías, Francisco[1], Cadarso-Suárez, Carmen[2], Espasandín, Jenifer[3]

[1] Profesor asociado Universidad de Vigo, Unidad de Epidemiología e Investigación Clínica Complejo Hospitalario Universitario de Santiago
[2] Catedrático, Departamento de Estadística, Universidad de Santiago (USC)
[3]Departamento de Estadística, Universidad de Santiago (USC)

SUMMARY

Purpose/Objective: In this study three decision analysis models are compared to evaluate the cost-effectiveness of these two imaging technologies used in coronary ischemia diagnose, the CTA (computed tomography angiography) and the conventional angiography (hemodynamics).

Introduction: Health technology assessment is becoming more and more important, especially with the economic environment affecting society nowadays. One of the areas that have been having more progression is the technology to diagnose, as imaging technics that can help diagnose and treat patient's condition. Most of the times, the decision between a procedure and another one can be affected by its costs. Nevertheless dealing with public health the cost cannot be weighted alone, as we have other aspects that are of major importance, as the condition diagnose and resolution, and the patient commodity and quality of life. In a first approach, the cost-effectiveness is a good compromise to take a decision that can take in to account the interests of institutions, physicians and patients.

Materials and Methods: Data was obtained from a population of 69 individuals, both men and women, with ages between 35 and 88 years old. The variables collected for this study were patient age, sex, imaging procedure used and its cost. Three decision analysis models were used: Decision tree (Gold standard), Discriminant Analysis and Regression analysis - Logit model.

Results: The discriminant analysis shows the variable age as significant while the variable costs is not significant, and 65,2 % of the cases being obtained correctly classified. The regression MCP nevertheless, shows significance for variable cost with a positive coefficient in using the CT angiocardio. In the same way, the logistic regression shows significance for the variables costs and age and not for variable sex, being positive the coefficients of the significant variables.

Conclusion: Logistic regression shows the highest capacity of explanation to choose technology for angiocardio, with the biggest weight for the variable costs on the remaining ones of the study.

Key Words: Mathematical Modelling in Medicine, Economic evaluation; Health technology, Coronary Ischemia, Medical imaging

1. INTRODUCTION

Health technology assessment [1] is becoming more and more important, especially given the economic environment affecting society nowadays. One of the areas that have seen more progress is the technology in diagnosis, such as imaging techniques that can help diagnose and treat a patient's condition. Most of the times the decision between one procedure and another can be affected by its costs.

Nevertheless, dealing with public health the cost cannot be weighed alone, as we have other aspects that are of major importance, such as the diagnose and resolution of the condition, and the patient's comfort and quality of life. In a first approach, cost-effectiveness is an appropriate factor in making a decision that can take in to account the interests of institutions, physicians and patients.

In the case of diagnosis of coronary ischemia we need to choose between multidetector computed tomography (CTA) and conventional angiography (CA). Multidetector computed tomography for coronary angiography [2] is non-invasive and allows visualization of both the lumen and vessel walls of coronary arteries and to discriminate between calcified and noncalcified atherosclerotic plaque. Conventional angiography [3-5] by catheter is an invasive technique which only allows visualisation of the coronary lumen, but does allow a therapeutic invasive procedure.

The aim of this article is to assess four classification statistical techniques: comparing discriminant analysis, and logistic regression with the gold standard decision tree to identify the best technology option for detecting relevant coronary artery lesions.

2. Materials and Methods

Data was obtained from a population of 81 individuals, aged between 35 and 88 years old. The variables collected for this study were patient age, sex, imaging procedure used and its cost.

Non-invasive studies were performed on 41 with a helical computer tomography system. A contrast agent was injected into a peripheral vein, and cross-sectional images were reconstructed with a slice thickness of 0.5 mm or 1.0 mm. The other 40 studies were performed with Invasive coronary angiography. Findings from both techniques were analyzed according to a predetermined segmented anatomical model of the coronary artery. The detection and relevance of coronary artery lesions were evaluated.

Cost was estimated using four scenarios each combining alternative allocation of human resources. For scenario one the time assigned was 30, 60, 120 and 90 minutes for the specialist, radiologist, nurse and technician respectively. Time for scenario two was 60 for each professional. The third scenario considerers the following times 120, 60, 120 and 120. Finally the fourth scenario supposed time of 90, 60, 90 and 90.

The total cost calculation was made assigning the human resources cost plus the cost for medicines material and technologies to each scene.

Decision tree was taken as gold standard to compare with three decision analysis models: Discriminant Analysis, Logistic regression.

Decision tree [6] is a flow diagram depicting the logical structure of a choice under conditions of uncertainty.

It was calculated from the following expression,

$$p(E \mid R) = \frac{p(E)p(R \mid E)}{p(E)p(R \mid E) + p(\bar{E})p(R \mid E)} \quad (1)$$

$E = event;\ \bar{E} = complement;\ R = \mathrm{Re}sult\ \ of\ \ research$

Discriminant Fisher Function

$$D = \mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_k x_k \quad (2)$$

An explanation of logistic regression begins with an explanation of the logistic function, which, like probabilities, always takes on values between zero and one:

and $$\Pi(x) = \frac{e^{(\beta 0+\beta 1X)}}{e^{(\beta 0+\beta 1X)}+1} = \frac{1}{e^{-(\beta 0+\beta 1X)}+1'}$$

and $$g(x) = \ln \frac{\Pi^{(X)}}{1-\Pi^{(X)}} = \beta_0 + \beta_1 X$$

$$\frac{\Pi(x)}{1-\Pi(x)} = e^{(\beta_0+\beta_1 X)}$$

Results

Decision tree

The decision tree shows the results for the four scenarios. The total cost for all scenarios is over 200 for computed tomography, and rising from 156 for conventional angiography for scenario one.

We are faced with two alternatives in the diagnosis of Coronary Ischemia. The first is the technology of computed tomography angiography (CTA), the second possibility, conventional angiography (CA).

The problem arises when choosing the most suitable technology. Using conventional angiography we can be sure that this is the technology of finally use, the probability is 100%. The results for the four scenarios were 156, 191, 206 and 225 euros respectively.

We compare this alternative to computed tomography angiography of, but in this case the probability of using this alternative ultimately not be 100%, but 79% and will have a 21% chance of conducting a conventional angiography additionally ie, for a given patient, there will be two interventions, whereby the cost for this case is calculated as the sum of conducting both interventions. The results show that for all scenarios, it is more advantageous to make a conventional angiography from the start, as we incur lower costs.

Discriminat analysis

The discriminant analysis shows the variable age as significant while the variable costs is not significant, and 65,2 % of the cases being obtained correctly classified.

Logistic regression

With this model we can estimate the probability of an event based on a set of predictor variables, which can er qualitative or quantitative. In our case, sex and age are qualitative while the costs would be the only quantitative variable.

The results of the logistic regression shows significance for the variable age and not significance for variable sex and costs, being positive the coefficients of the significant variable. The percentage of correct classification was 63.8% of cases.

3. **Conclusions**

Sensitivity of 64-multislice CT is high enough to rule out significant coronary artery disease in patients with low probability pre test of coronary artery disease though the Positive Predicted Value is low.

Diagnosis power of Non-invasive coronary artery angiography by multidetector-row spiral computed tomography although is good it is still inferior to invasive coronary angiography.

Moreover, each methodology used in this study provides different significant variables for the detection of the Coronary Ischemia.

Referencias

[1] Berger ML, Bingefors K, Hedblom EC, Pashos CL, Torrance GW. *Health Care Cost, Quality, and Outcomes: ISPOR Book of Terms*. Lawrenceville, NJ: ISPOR, 2003.

[2] Sones MS, Shirey EK. *Cine coronary arteriography*. *Med Concepts Cardiovasc* Dis. 1962;31:735.

[3] Rodríguez-Palomares JF, Cuellar H, Martí G, García B, González-Alujas MT, Mahia P, et al. *Coronariografía mediante tomografía computarizada de 16 detectores antes de la cirugía de recambio valvular*. Rev Esp Cardiol. 2011;64:269-76.

[4] Mendoza-Rodríguez V, Llerena LR, Milián-García V, Linares-Machado R, Hernández-Cañero A, Llerena LD, et al. *Precisión de la tomografía de 64 cortes en el diagnóstico de la cardiopatía isquémica*. Arch Cardiol Mex. 2008;78:162-70.

[5] Leta, R, Garcia Picart, J, Carreras, F, et al. (2004). *Non-invasive coronary angiography with 16 multidetector-row spiral computed tomography: A comparative study with invasive coronary angiography.* Revista española de cardiología, 57(3), 217-24.

[6] Bedding, A. and Lilly, E. (2005), *Decision making in health and medicine Hunink M, Glasziou P, Siegel J, Weeks J, Pliskin J, Elstein A, Weinstein MC* (2001) ISBN 0521770297; 404 pages; £38; $60 Cambridge University Press

Comunicacións orais

Sesión VIII: Epidemioloxía II

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Adjusted Wald confidence intervals in the meta-analysis of prevalence proportions

Vera Afreixo[1], Sara Escudeiro[2] and Adelaide Freitas[3]

[1]CIDMA/IBIMED/Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
[2]CIDMA/Polytechnic Institute of Coimbra, ESAC, 3040-316, Coimbra, Portugal
[3]CIDMA/Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

ABSTRACT

Meta-analysis could be used in many fields of research, having high importance in clinical context. In particular, it could be used to estimate the prevalence and incidence of a disease or a multi-resistant bacteria in epidemiological context. In this paper, we discuss the meta-analysis methods for the proportion effect size (e.g. prevalence, incidence), pointing our interest in rare events (analogous for the abundant events) that take into account the practical problem of estimating one low or very low prevalence/incidence. Several approximate adjusted Wald confidence intervals (CIs) for the meta-analysis of prevalence proportions are proposed. These CIs are developed using a parametric family of shrinkage estimators for estimating the proportions. Two popular statistical models, the fixed-effect model and the random-effects model, are used and discussed in this work. A simulation study is carried out to compare these CIs amongst themselves and compare them with those obtained by the best-known transformation: logit and double arcsine. The proposed comparison is based in three measures: bias; mean square error (MSE) and coverage probability (CP).
Based on these comparisons, the CIs proposed can be recommended as an alternative to the arcsine transformation (the preferred option) in the context of rare events, for the random-effect model. It was noticed that the choice of method for estimating the proportions should take into account the number of studies that are included in the meta-analysis study.

Keywords and key sentences: Meta-analysis, Proportion, Approximate confidence intervals, Linear combination of binomial proportions, Rare events.

1. INTRODUCTION

The meta-analysis of proportions is usually carried out via three well known methods: the classic Wald method (Wald-0) [1], the logit transform and the arcsine transform [2]. The inverse of the Freeman- Tukey double arcsine transformation is preferred over the logit and classic Wald methods [2], in particular for large or small proportions.
There are several variants of the Wald method to estimate the linear combinations of binomial proportions [5, 6]. In this work, we present a comparison between transformed and direct methods to perform proportion meta-analysis of rare events.
To perform the simulation studies we use the prevalence meta-analysis of the mecC gene [3] as a motivation example. The overall estimated prevalence of the mecC gene, according to the double arcsine transformation and through a random-effect model, was 0.009% (95% CI = 0.005–0.013).

2. PROPORTION META-ANALYSIS BY WALD VARIANTS

Let us consider k independent binomial studies, $X_1, \ldots, X_k$, and let us denote by p_i the effect size and n_i the size in the ith study. In each study an adjusted estimate of the effect size $p_i, i = 1, \ldots, k$, denoted by $\tilde{p}_i$, is calculated using the parametric family of shrinkage estimators

$$\tilde{p}_i = \frac{X_i + h_i}{n_i + 2h_i}, \quad h_i > 0\,. \tag{1}$$

Depending on the particular h_i parameter, several different estimators can be considered.

Variants:	Wald-0	Wald-1	Wald-2	Wald-3
h_i	0	$\frac{2}{k}$	$\frac{z^2_{\alpha/2}}{2k}$	$\frac{z^2_{\alpha/2}}{2}\left(\mathbf{1}_{A_{i_-}} + \mathbf{1}_{A_{i_+}} + \frac{1}{k}\right)$

Table 1: Parameters h_i for the Wald CI methods in the estimation of p_i based on the parametric family (1). The sets associated to the indicator functions $\mathbf{1}_{A_{i_\mp}}$ are given by $A_{i_-} = \{x_i : (x_i = 0 \wedge \beta_i < 0) \vee (x_i = n_i \wedge \beta_i > 0)\}$ and $A_{i_+} = \{x_i : (x_i = 0 \wedge \beta_i > 0) \vee (x_i = n_i \wedge \beta_i < 0)\}$.

Table 1 shows the values of h_i suggested by Price and Bonett [4] and Andrés et al. [5]. When $h_i = 0$, we have the maximum likelihood estimator (MLE) of p_i, denoted by $\hat{p}_i$.
A linear combination of binomial proportions may be defined as $L = \sum_{i=1}^k \beta_i\, p_i$, where β_i is a known coefficient value. The MLE of L is equal to $\hat{L} = \sum_{i=1}^k \beta_i\, \hat{p}_i$. When suitably centred and scaled, the statistic $\hat{L}$ converges asymptotically in distribution to a standard normal distribution, i.e., $(\hat{L} - L)/\sqrt{v(\hat{L})} \underset{d}{\to} N(0,1)$, where $v(\hat{L})$ denotes the variance of $\hat{L}$ and is given by

$$v(\hat{L}) = \sum_{i=1}^k \frac{\beta_i^2\, p_i\,(1-p_i)}{n_i}\,. \tag{2}$$

An approximate $100(1-\alpha)\%$ adjusted Wald CI for p has the form

$$\tilde{L} \mp z_{\alpha/2}\sqrt{\tilde{v}(\hat{L})}\,, \tag{3}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution and $\tilde{v}(\hat{L})$ an variance estimator.

In meta-analysis the pooled effect size (proportion) is the weighted average of the individual effects. The process used to assign the weights to the classic or adjusted pooled proportion depends on the model used. Under the fixed-effect model, we assume that all studies in the analysis share the same true effect size p, that is, $p_1 = \cdots = p_k = p$, and the pooled effect size is given by

$$\tilde{p}_{\text{IV}} = \frac{\sum_{i=1}^k w_i \tilde{p}_i}{\sum_{i=1}^k w_i} = \sum_{i=1}^k \frac{w_i}{\sum_{i=1}^k w_i}\tilde{p}_i = \sum_{i=1}^k \beta_i \tilde{p}_i, \tag{4}$$

where w_i is the weight assigned to each study. According to the inverse variance of method, it is defined as

$$w_i = \frac{1}{\tilde{v}(\hat{p}_i)} =, \; i = 1, \ldots, k$$

where $\tilde{v}(\hat{p}_i) = \frac{\tilde{p}_i(1-\tilde{p}_i)}{n_i + 2h_i}$ is the within-study adjusted variance estimate.
Under the random-effects model, we assume that the true effect size could vary from study to study and the pooled proportion is given by

$$\tilde{p}_{\text{RE}} = \frac{\sum_{i=1}^{k} w_i^* \tilde{p}_i}{\sum_{i=1}^{k} w_i^*}, \tag{5}$$

where $w_i^* = \frac{1}{1/w_i + \hat{\tau}^2}$, with τ^2 being the variance between-studies [7]. The $\hat{\tau}^2 = \begin{cases} 0, & Q \leq k-1 \\ \frac{Q-(k-1)}{U}, & Q > k-1 \end{cases}$, with $Q = \sum_{i=1}^{k} w_i \tilde{p}_i^2 - \frac{\left(\sum_{i=1}^{k} w_i \tilde{p}_i\right)^2}{\sum_{i=1}^{k} w_i}$ the Cochran's Q-test statistic and

$$U = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}.$$

2. EXPERIMENTAL RESULTS

We used Monte Carlo simulation to compare the performance of the different CI methods for meta-analysis of prevalence[2]. We assessed the performance of each CI using the fixed-effect and random-effects models. Our motivation example, the meta-analysis on the prevalence of mecC MRSA (a small proportion), based on previously published results, has a total of 25 studies with significant heterogeneity. We assume three distinct small proportions, 0.01, 0.05 and 0.001; under the heterogeneity hypothesis we assume normal distribution with standard deviations 0.001, 0.005 and 0.0001, respectively. We implement 200000 the simulation procedure.
Table 2 shows the performance of the five estimation methods under analysis for the random-effects model, in the context of small proportions. The analysis was performed at a nominal value of 5%.

p=0.05						
	Wald-0	Wald-1	Wald-2	Wald-3	Logit	Double arcsine
Mean	0.04085	0.04725	0.04718	0.04922	0.05053	0.05016
Bias	-0.00915	-0.00275	-0.00282	-0.00078	0.00053	0.00016
MSE	0.00009	0.00002	0.00002	0.00001	0.00001	0.00000
CP	0.20762	0.73983	0.735565	0.87773	0.83867	0.86806
p=0.01						
	Wald-0	Wald-1	Wald-2	Wald-3	Logit	Double arcsine
Mean	0.00771	0.00897	0.00895	0.00965	0.01029	0.01010
Bias	-0.00230	-0.00103	-0.00105	-0.00036	0.00029	0.00010
MSE	0.00001	0.00000	0.00000	0.00000	0.00000	0.00000
CP	0.20187	0.68179	0.67770	0.90508	0.76152	0.79993
p=0.001						
	Wald-0	Wald-1	Wald-2	Wald-3	Logit	Double arcsine
Mean	0.00092	0.00081	0.00080	0.00107	0.00111	0.00105
Bias	-0.00008	-0.00019	-0.00020	0.00007	0.00011	0.00005
MSE	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
CP	0.80365	0.58707	0.57046	0.93620	0.83416	0.89840

Table 2: Results of the evaluation measures for the CIs obtained through the different methods, for small prevalence with the random-effects model. MSE - mean squared error; CP - coverage probability.

In the example under study, we observe that variant-3 of the Wald method provides the best coverage probability. We also performed a Monte Carlo simulation study using the fixed-effect model for the same small proportion and verified that the method based in double arcsine transform has the best performance (results not shown).

The meta-analysis for the mecC MRSA example was performed with the random-effects model due the presence of significant heterogeneity. Motivated by our simulation study, we reestimated the pooled prevalence using variant-3 of the Wald method: $\tilde{p}$=0.0039 with 95% CI = 0.0018–0.0057.

		Confidence Interval	
	$\tilde{p}$	Lower Bound	Upper Bound
Wald-0	0.0027	0.0011	0.0042
Wald-1	0.0028	0.0012	0.0044
Wald-2	0.0028	0.0012	0.0044
Wald-3	0.0039	0.0018	0.0057
Logit	0.0133	0.0054	0.0321
Double arcsine	0.0090	0.0054	0.0135

Table 3: Results by method for the mecC MRSA example [3], in the context of random-effects model.

We should note that the estimated prevalence is lower than that presented by [3] (see Table 3).

3. CONCLUSIONS

The double arcsine transformation is the recommended method for the fixed-effect model. However, in the case of estimating small proportions with the random-effects model, variant-3 of the Wald method provides better results than the methods based on the double arcsine transformation. In the context of the random-effects model, there is still room for improvement, as the non-coverage probability is still far from the nominal value.

ACKNOWLEDGMENT

This work was supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, the iBiMED - Institute for Biomedicine the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013 and UID/BIM/04501/2013.

References

[1] Wald, A. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society* 1943, 54: 426–482.

[2] Barendregt, Jan J and Doi, Suhail A. and Lee, Yong Yi and Norman, Rosana E. and Vos, Theo. Meta-analysis of prevalence. *Epidemiol Community Health* 2013, 67:974–978.

[3] Diaz, Raquel and Ramalheira, Elmano and Afreixo, Vera and Gago Bruno. Methicillin-resistant Staphylococcus aureus carrying the new mecC gene—a meta-analysis. *Diagnostic Microbiology and Infectious Disease*2016, 84:135–140.

[4] Price, Robert M. and Bonett, Douglas G.. An improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis* 2004, 45:449–456.

[5] Andrés, A. Martín and Hernández, M. Álvarez and Tejedor, I. Herranz. Inferences about a linear combination of proportions. *Statistical Methods in Medical Research* 2011, 20:369–387.

[6] Andrés, A. Martín and Tejedor, I. Herranz and Hernández, M. Álvarez. The optimal method to make inferences about a linear combination of proportions. *Journal of Statistical Computation and Simulation* 2012, 82:123–135.

[7] Borenstein, Michael and Hedges, L. V. and Higgins, J. P. T. and Rothstein, H. R. Introduction to Meta-Analysis. *John Wiley and Sons, Ltd.* 2009.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

MODELOS DE REGRESSÃO LOGÍSTICA SOB DELINEAMENTOS AMOSTRAIS COMPLEXOS

Isabel Natário[1], Sheyla Rodrigues Cassy[2]

[1] CMA, Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, 2829-516, Caparica, Portugal. *icn@fct.unl.pt*
[2] CMA, Faculdade de Ciências da Universidade Eduardo Mondlane

Palabras e frases chave: Delineamentos amostrais complexos; regressão logística

RESUMO

A grande maioria dos estudos de prevalência de doença é realizado a partir de dados obtidos por inquérito, em que a informação é extraída a partir de unidades amostrais selecionadas de acordo com delineamentos frequentemente complexos, que mais usualmente resultam da estratificação da população em grupos relativamente homogéneos, dos quais se amostram um número pré-determinado de unidades, e da aglomeração, em que se divide a população em grupos e se amostra aleatoriamente um seu subconjunto. A aglomeração pode ser considerada em vários estádios, como, por exemplo, no delineamento usado nos Demographic and Health Surveys (DHS) descrito em [4] para Moçambique: alojamentos dentro de unidades de enumeração da partição estatística da área a considerar, até as unidades elementares, os indivíduos, serem amostradas no último estádio.

Assim, nos delineamentos amostrais complexos as unidades amostrais são selecionadas com diferentes probabilidades, sendo necessário fazer refletir esse facto nas análises estatísticas que se vierem com eles a fazer, atribuindo-lhes pois correspondentes pesos diferentes, normalmente tomados como sendo o inverso das correspondentes probabilidades de seleção. As complicações metodológicas daqui decorrentes devem ser levadas em consideração na análise estatística destes dados e na estimação das variâncias dos estimadores obtidos, sob pena de se sobrestimar ou subestimar os erros padrão associados aos estimadores empregues [11].

O modelo de regressão logística detém um dos papéis centrais em estudos de prevalência de doenças [8], sendo dos modelos mais empregues neste tipo de análises. Esta trabalho é sobre como a regressão logística deve ser empregue nestas situações em que os dados provêm de delineamentos amostrais complexos: reveem-se e detalham-se quais os pontos que já estão metodologicamente resolvidos e quais os aspetos em que ainda há questões em aberto. Note-se que muitos dos desenvolvimentos metodológicos não foram propostos apenas no contexto do modelo de regressão logística, sendo transversais a outros modelos.

Relativamente à estimação do modelo de regressão logística sob amostragens complexas, a abordagem que se faz é considerar uma função de pseudo-verosimilhança que incorpora os pesos de amostragem que se devem usar na função de verosimilhança [17]. É a partir desta função que se obtêm os estimadores de máxima verosimilhança [3], que

permitem a determinação das razões de chances ajustadas, representando o impacto multiplicativo do aumento de uma unidade numa certa variável preditora nas chances da variável resposta tomar o valor 1 (constituir um sucesso, ter a doença), controlando os efeitos das outras variáveis.

Os erros padrão podem ser obtidos a partir de um método de estimação da variância por linearização da série de Taylor, que leva em consideração as características do delineamento amostral [18].

Os testes de Wald para testar a significância dos coeficientes de regressão neste contexto foram propostos por [9]. Testes da razão de verosimilhanças e testes *score*, que constituem uma alternativa aos testes de Wald, para modelos de regressão foram desenvolvidos por [14].

Para avaliar a qualidade do ajuste dos modelos de regressão logística para dados recolhidos com base em delineamentos amostrais complexos, o conhecido teste de Hosmer-Lemeshow foi adaptado por [1].

Ainda a seleção de modelos através de critérios de seleção de modelos tais como os AIC e BIC foram recentemente adaptados para dados de delineamentos amostrais complexos por [15].

A necessidade destes desenvolvimentos metodológicos é relativamente recente, movidos em grande parte pela necessidade da aplicação, sendo que ainda há questões da análise dos modelos de regressão logística para dados obtidos por delineamentos amostrais complexos que ainda não estão completamente encerradas. Por exemplo, as análise de resíduos, observações influentes e potencialmente influentes são pontos que não têm ainda sido abordados neste tipo de estudos. Outras técnicas da qualidade de ajuste que não as atrás referidas, comuns na aplicação dos modelos de regressão logística, como as descritas em [8], não estão ainda desenvolvidas.

O modelo de regressão logística para dados obtidos de delineamentos amostrais complexos, bem como as possíveis análises descritivas ou outro tipo de análises, encontram-se já implementados em vários programas de *software* estatístico, como sendo o SAS, o STATA e o SPSS. Também para o `R-project`, [13] desenvolveu o pacote `survey`, que permite ligar os necessários meta-dados do delineamento considerado aos dados, tal que os ajustes corretos na análise possam ser implementados de forma automática e correta. Estes *softwares* têm-se vindo a modernizar e completar à medida que os desenvolvimentos metodológicos vêm sendo alcançados.

Vários são já os exemplos que aplicam as metodologias específicas a dados obtidos por delineamentos amostrais complexos, em particular considerando o modelo de regressão logística, ainda que por vezes as análises sejam ainda pouco estatisticamente detalhadas e mais do tipo descritivo. Salientam-se aqui alguns estudos de modelação do uso de rede mosquiteira no combate à Malária em Moçambique [4] (*software* `R`), no Gana [2] (*software* SPSS), estudo de fatores para o uso individual de redes mosquiteiras no combate à Malária na Etiópia [7] (*software* STATA), estudo sobre as condições físicas e de manutenção das redes mosquiteiras no combate à Malária no Quénia [16] (*Software* `R`), estimação da prevalência do HIV no Zimbabwe [5] (*Software* `R`), estimação da prevalência e seus fatores entre mulheres jovens com dois ou mais parceiros sexuais na Beira, Moçambique [19] (*Software* STATA). Em muitos destes estudos não há ainda preocupações sobre a avaliação da qualidade do ajuste nem da seleção de modelos, por exemplo, sendo pois importante a disseminação de textos que clarificam a sua aplicabilidade, como [12] e [10], ou trabalhos que a divulguem, como o presente.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado por fundos nacionais através da FCT - Fundação para a Ciência e Tecnologia - sob o projeto UID/MAT/00297/2013.

Referencias

[1] Archer, K.J., Lemeshow, S., Hosmer, D.W. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics and Data Analysis* 51, 4450–4464.

[2] Baume, C.A., Franca-Koh, A.C. (2011). Predictors of mosquito net use in Ghana. *Malaria Journal* 10, 1–6.

[3] Breslow, N.E., Holubkov, R. (1997).Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* 16, 103–116.

[4] Cassy, S.H., Natário, I, Martins, M.R. (2016). Aplicação da Regressão Logística na Modelação do uso de Redes Mosquiteiras em Moçambique. *Atas do XXII Congresso da Sociedade Portuguesa de Estatística* - submetido.

[5] Chinomona, A., Mwambi H.G. (2015). Estimating HIV Prevalence in Zimbabwe Using Population-Based Survey Data. *PLoS ONE* 10: e0140896.

[6] Chromy, J.R., Abeyasekera, S. (2005). Statistical Analysis of Survey Data. In *Household Sample Surveys in Developing and Transition Countries*, Chapter XIX, 389-417. United Nations, NY. Available for download at http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf.

[7] Graves, P.M. *Etal.* (2011). Factors associated with mosquito net use by individuals in households owning nets in Ethiopia. *Malaria Journal* 10, 1–12.

[8] Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition.* Wiley Series in Probability and Mathematical Statistics, New York.

[9] Koch, G.G., Freeman, D.H., Freeman, J.L. (1975). *Strategies in the Multivariate Analysis of Data from Complex Surveys.* International Statistical Review / Revue Internationale De Statistique, 43, 59–78.

[10] Lee, E.S., Forthofer, R.N. (2006). *Analyzing Complex Survey Samples, Second Edition.* Quantitative Applications in the Social Sciences Series. Sage Publications, Thousand Oaks.

[11] Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software* 9, 1548–7660.

[12] Lumley, T. (2010). *Complex Surveys: a Guide to Analysis Using R.* Wiley, New Jersey.

[13] Lumley, T. (2014). *Survey: Analysis of Complex Survey Samples.* R Package version 3.30-3.

[14] Lumley, T., Scott, A. (2014). Tests for Regression Models Fitted to Survey Data. *Australian and New Zealand Journal of Statistics* 56, 110–123.

[15] Lumley, T., Scott, A. (2015). AIC and BIC for Modeling with Complex Survey Data. *Journal of Survey Statistics and Methodology* 3, 1–18.

[16] Mutuku, F.M. *Etal.* (2013). Physical condition and maintenance of mosquito bed nets in Kwale County, coastal Kenya. *Malaria Journal* 12, 1–14.

[17] Roberts, G., Rao, N.K., Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* 74, 1–12.

[18] Woodruff, R.S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association* 66, 411-414.

[19] Zango, A. *Etal.* (2013). Determinants of Prevalent HIV Infection and Late HIV Diagnosis Among Young Women with Two or More Sexual Partners in Beira, Mozambique. *PLoS ONE* 8, e63427.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

demogal: paquete para proyectar flujos y stocks de población

Calvo Ocampo, Mª Esther[1], Iglesias Patiño, Carlos[1], Gómez Duaso, Juan[1], López Vizcaíno, Mª Esther[1], Martín Vila, María[1] y Silveira Calviño, Solmary[1]

[1]Instituto Galego de Estatística

RESUMEN

El objetivo de este trabajo es el desarrollo de un paquete en el lenguaje de programación R, **demogal**, que permita la elaboración de proyecciones de la población para uno o más espacios geográficos, basándose en el método clásico de las componentes y utilizando el modelo de proyección multirregional. El paquete demogal incorpora en la elaboración de la proyección la posibilidad de consistencia con un espacio superior o no.

Palabras y frases clave: proyección de población, paquete R, demografía.

1. Introducción

La elaboración de indicadores demográficos y proyecciones de población es una tarea habitual de las oficinas de estadística pública porque, en todo tipo de planificación, es fundamental conocer la estructura demográfica actual y las perspectivas de evolución de la población. Hoy en día existen herramientas informáticas que permiten el cálculo de indicadores y de las tablas de mortalidad de manera sencilla, como pueden ser **Epidat**[2], hojas de cálculo, etc. Estas herramientas, en algunos casos, no tienen la flexibilidad necesaria para adaptarse a cualquier metodología de cálculo, en otros casos, no permiten procesar grandes conjuntos de datos y, lo que es más importante, cuando se maneja gran cantidad de información, la mayor parte de ellas no permiten automatizar el proceso de cálculo, etc. Teniendo en cuenta todas estas dificultades que plantean las herramientas existentes, desde el Instituto Galego de Estatística (IGE) se pensó en aprovechar la flexibilidad de R para desarrollar un paquete que permitiese automatizar el cálculo de todas las funciones demográficas que periódicamente se utilizan en el IGE.

2. Metodología

La elaboración de estas proyecciones de población está basada en el método clásico de componentes[3], el cual, a partir de una población residente en un cierto espacio geográfico y de los datos observados para cada una de las componentes demográficas básicas (mortalidad, fecundidad y migración), permite obtener la población en periodos futuros bajo ciertas hipótesis sobre la evolución de las componentes, que son las que determinan el crecimiento de la población y su estructura por edades. Se utiliza el modelo de proyección multirregional propuesto por Willekens[1], que proporciona las cifras de población residente y las cifras proyectadas de nacimientos, defunciones y movimientos migratorios que tendrán lugar en cada espacio considerado y para cada uno de los

periodos del horizonte proyectivo, guardando todo ello la necesaria coherencia entre flujos y stocks demográficos y la debida consistencia interterritorial. En el paquete demogal se contemplan dos enfoques:

- Proyección de la población para uno o más espacios geográficos aplicando el modelo multirregional.

- Proyección de la población para varios espacios geográficos mediante el modelo multirregional, pero imponiendo que los datos agregados del conjunto de espacios coincidan con los de una proyección preexistente del área formada por el conjunto de los espacios (consistencia con el espacio superior).

El método empleado proyecta cada uno de los fenómenos demográficos por separado, para después unirlos en la obtención de la proyección de la población. Las hipótesis que se hayan planteado sobre cada uno de estos fenómenos demográficos se aplican a la población en el periodo de partida t_0, obteniendo la población en el periodo t_{0+1}. Actuando así, de forma iterada, se llega al horizonte de la proyección, t_{0+nn}. Para la obteción de la proyección de la población será necesario disponer de información de nacimientos, defunciones y movimientos migratorios correspondiente a los últimos n periodos: $t_{0-n}, t_{0-(n-1)}, \ldots, t_{0-1}$.

3. El paquete demogal

En el paquete demogal hay un total de 12 nuevas funciones que pueden ser de interés para las personas que se dedican a hacer indicadores y proyecciones de distintos tipos de poblaciones. La funciones se presentan de manera resumida en la Tabla 1.

Function	**Description**
mortable	Permite calcular las funciones de la tabla de mortalidad completa o abreviada por sexo, para diferentes espacios geográficos y diferentes periodos de tiempo a partir de los datos de defunciones y población.
project	Proyecta diferentes magnitudes de la fecundidad en base a una relación log-lineal de las mismas y el tiempo. Se utilizan n periodos para hacer la proyección.
tr	Calcula el Índice sintético de fecundidad y el Índice sintético de emigración.
PMortality	Proyecta las tasas de mortalidad y los supervivientes de la tabla de mortalidad para el espacio de proyección.
PFertility	Proyecta el Índice sintético de fecundidad y las tasas de fecundidad para el espacio de proyección.
ExtEm	Proyecta las tasas de emigración al exterior del espacio de proyección.
PMortalityReg	Proyecta y regionaliza las tasas de mortalidad para el espacio de proyección, sexo, edad y todo el periodo de proyección.
PFertilityReg	Proyecta y regionaliza las tasas de fecundidad para el espacio de proyección, edad y todo el horizonte de proyección.
ExtEmReg	Proyecta y regionaliza las tasas de emigración al exterior del espacio de proyección.
ExtInReg	Regionaliza y proyecta el número de inmigrantes procedentes del exterior. En ella se parte de la hipótesis de que se considera un número constante de inmigrantes durante todo el periodo de proyección para el espacio superior (v_{inmig}). Si no se introduce el parámetro, toma la media de los que hubo en los últimos n años.
IntEm	Proyecta las tasas de emigración interna en todo el periodo de proyección.
projection	Proyecta la población para el espacio de proyección, sexo, edad y todo el periodo de proyección.

Tabla 1: Nuevas funciones del paquete demogal.

El flujo de utilización de las funciones se muestra en la Figura 1. En lo que sigue se presentan un par de ejemplos ilustrativos del uso de las funciones descritas en la Tabla 1.

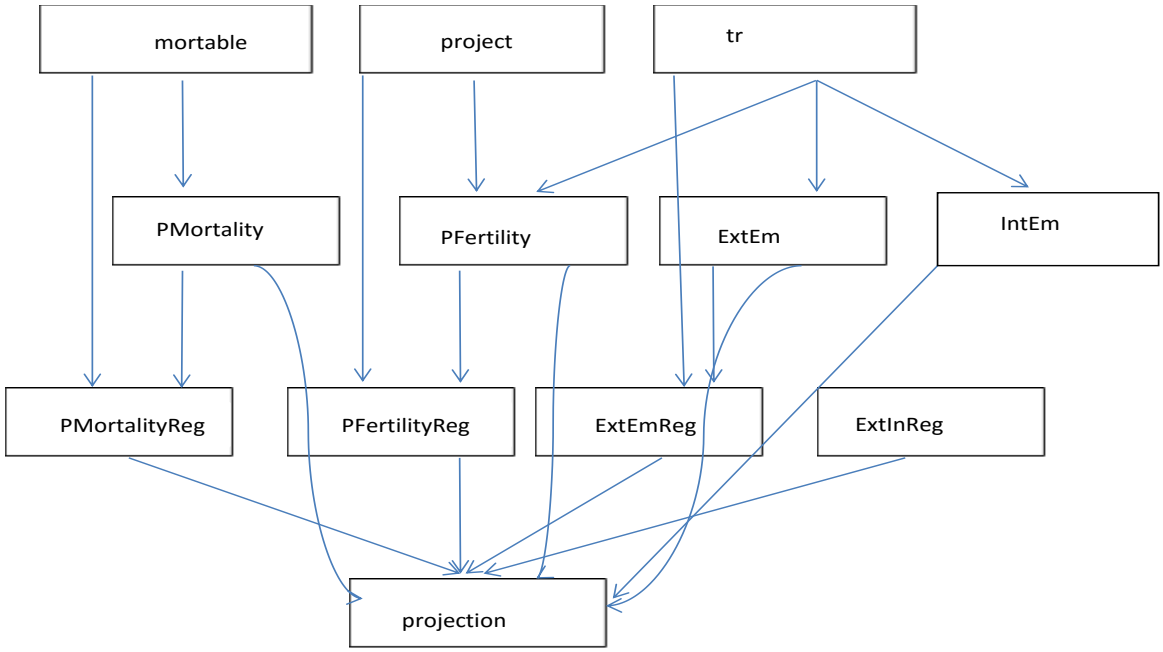


Figura 1: Flujo de ejecución del paquete.

4. Ejemplo de uso de demogal

El paquete demogal permite realizar proyecciones de la población para uno o más espacios geográficos. En el caso de que se considere más de un espacio geográfico se puede optar por la consistencia con un espacio superior (unión de cada uno de los espacios de proyección) o no.
Se presenta en primer lugar un ejemplo de la proyección de Galicia del año 2016 al 2030. El paquete *demogal* requiere el input de dos parámetros: $year = 2015$ (periodo de partida de la proyección, que se corresponde con el t_0) y $nn = 15$ (número de años que se van a proyectar). Para llevar a cabo esta proyección se necesita además información de los fenómenos demográficos en periodos anteriores, así como la población del periodo base. Se utiliza por tanto la siguiente información:

- Población del año 2015 por sexo y grupos quinquenales de edad (*poboacion*).
- Nacimientos por sexo y grupos quinquenales de edad de la madre para los 10 años anteriores (*data.nac.ga*).
- Mortalidad, población y defunciones por sexo y grupos quinquenales de edad para los 10 años anteriores (*data.ga*).
- Emigrantes al exterior por sexo y grupos quinquenales de edad, para los últimos 3 años (*emig.out.ga*).
- Inmigrantes procedentes del exterior por sexo y grupos quinquenales de edad, para los últimos 3 años (*imig.out*).

A continuación se presenta un breve resumen de las funciones utilizadas:

```
R> ###PROYECCIÓN DE GALICIA
R> year=2015
R> nn=15
```

```
R> #PROYECTAR MORTALIDAD
R> m.rate.pr=PMortality(data.ga, nn)[[1]]
R> #PROYECTAR FERTILIDAD
R> f.rate.pr=PFertility(data.nac.ga,nn)[[2]]
R> #PROYECTAR EMIGRACIONES
R> u=ExtEm(emig.out.ga,nn)
R> #PROYECTAR INMIGRACIONES
R> prov=12
R> inmig=27000
R> v=as.data.frame(cbind(prov,inmig))
R> uin=ExtIn(imig.out,relation,nn,v_inmig=v)

R> #Lista con la emigraciones y las inmigraciones. En este caso no hay migraciones internas
R> migration.ga=list(u[[1]],uin.ga,NULL)
R> ### SE PROYECTA LA POBLACIÓN DE GALICIA
R> m=projection(nn,r,poboacion,m.rate.pr,f.rate.pr,migration.ga,relation=NULL)
```

Se considera ahora un segundo ejemplo en el que se pretende hacer una proyección de la población de las provincias gallegas que sea consistente con la proyección hecha en el primer ejemplo para Galicia. La consistencia con el espacio superior se indica en la función de proyección con el dataframe *relation* y, al igual que en el primer ejemplo, se requiere también el input de los parámetros $year = 2015$ y $nn = 15$. Por lo que respecta a la información necesaria para realizar la proyección, en este caso se necesitará además de la misma información de partida que en el primer ejemplo pero relativa a cada provincia, la siguiente información:

- Población, nacimientos y defunciones por sexo y grupos de edad proyectados para Galicia.
- Fichero que relacione los espacios geográficos de proyección (provincias) con el espacio superior al que pertenecen (Galicia) (*relation*).

A continuación se presenta un breve resumen de las funciones utilizadas:

```
R> prox_m_ga <- PMortality(data.ga, nn)
R> prox_m_proReg <- PMortalityReg(data.prov, prox_m_ga[[2]], relation,
+     reg = "W")
R> tfr.sup = PFertility(data.nac.ga, nn)[[1]]
R> prox_n_pro <- PFertilityReg(data.nac.prov, tfr.sup, nn, relation)
R> internas = IntEm(emig.int_prov, emig.intc_prov, nn)
R> u = ExtEm(emig.out.ga, nn)
R> iset = u[[2]]
R> ise = u[[3]]
R> uu = ExtEmReg(emig.out, iset, ise, relation, nn)
R> prov = 12
R> inmig = 27000
R> v = as.data.frame(cbind(prov, inmig))
R> uin = ExtIn(imig.out, relation, nn, v_inmig = v)
R> m_prov = projection(nn, r, pop_prov, prox_m_proReg, prox_n_pro,
+     migration = list(uu, uin, internas), relation, pop.sup.pr,
+     b.sup.pr, m.sup.pr)
```

Referencias

[1] Willekens, F.J. and Drew, P. (1984), A Multiregional Model for Regional Demographic Projection. *Demographic Research and Spatial Policy*, Ch.15, pp. 309-331 - Academic Press, London.

[2] Epidat: programa para análise epidemiolóxica de datos. Versión 4.0, xuño 2012. Consellería de Sanidade, Xunta de Galicia, España; Organización Panamericana de la Salud (OPS-OMS). Disponible en: http://dxsp.sergas.es.

[3] Vinuesa, Julio y Puga, Dolores (2007), Técnicas y ejercicios de Demografía

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

ESPERANZAS DE VIDA SEN LIMITACIÓNS EN GALICIA

Iglesias Patiño, C.L.[1], López Vizcaíno, M.E.[1], Martín Vila, M.[1], Calvo Ocampo, E.[1] e Santiago Pérez, M.I.[2]

[1] Instituto Galego de Estatística
[2] Dirección Xeral de Saúde Pública

RESUMO

Estímase a esperanza de vida sen limitacións para Galicia no ano 2013 polo método de Sullivan, empregando unha enquisa por mostraxe para determinar as prevalencias de limitacións por problemas de saúde na poboación de 16 anos e máis, por sexo e idade. As prevalencias non dispoñíbeis impútanse. Estímase tamén o seu erro de mostraxe baseado no deseño por varios métodos.

Palabras e frases chave: esperanza de vida, método de Sullivan, modelo logit, remostraxe.

1. INTRODUCIÓN

A esperanza de vida ten sido durante moitos anos o indicador sanitario por excelencia. Unha das súas vantaxes é que permite as comparanzas transversais dun xeito doado, ata os sistemas estatísticos menos desenvolvidos adoitan ter resultados sobre a poboación e os óbitos.

Este indicador ten acrecentado o seu nivel dun xeito notábel, froito da mellora das condicións de vida nas xeracións pasadas. Aínda que, quizais, o seu crecemento estea a se desacelerar nos países máis desenvolvidos, debido ás mudanzas de hábitos, é indubidábel que o pasamento das persoas se produce agora a unha idade máis avanzada que hai algunhas décadas.

No entanto, non está claro que este alongamento da vida vaia acompañado dun estado de saúde aceptábel. Por iso convén introducir outros indicadores que "desconten" o efecto duns anos vividos nunhas condicións peores. Neste senso, as expectativas de saúde son indicadores sintéticos que, alén de reflectir o impacto da mortalidade nunha poboación, recollen tamén as condicións de saúde presentes nela [1]. En concreto, a esperanza de vida sen limitación (EVL, en diante) estima o número medio de anos que pode esperar vivir unha persoa sen limitación das actividades que realiza na vida cotiá. Para calcular este indicador precísanse as táboas de mortalidade da poboación e as prevalencias de limitación por grupos de idade, os mesmos que os da táboa. Frecuentemente, a fonte das prevalencias é unha enquisa, polo que están afectadas por un erro de mostraxe que se debe ter en conta á hora de estimar as EVL.

Por todo isto, o obxectivo desde traballo foi calcular esperanzas de vida sen limitacións para a poboación galega no ano 2013 e propoñer métodos para estimar o seu erro de mostraxe.

2. DESCRICIÓN DA METODOLOXÍA

Para o cálculo das EVL empregamos a técnica de Sullivan [2] cun procedemento de estimación lixeiramente diferente acorde co que se emprega adoito na literatura baixo esta denominación. A construción da táboa de mortalidade, uns dos obxectivos da Biometría humana, realizámola segundo o proxecto técnico consultábel na web do IGE na epígrafe Poboación, e mediante a función `mortable` cuxo precedente xa comunicabamos no primeiro destes encontros [3]. Esta función forma parte do futuro paquete **demogal**, que estamos a presentar neste mesmo encontro noutro relatorio [4]. É unha táboa

transversal, do momento, debido a que o tempo biométrico, a idade, pode resultar moi longo para algúns individuos da cohorte.

Sexa $\mathcal{I}_x$ o conxunto dos estremos inferiores dos chanzos etarios $[z, z+n_z)$ cuxo mínimo é x, por exemplo, $\mathcal{I}_0 = \{0, 1, 5, 10, 15, 20, \cdots, \omega\}$ onde ω é o estremo inferior do chanzo derradeiro. Entón a EVL para a idade x e cada sexo é

$$EVL_x = \frac{\sum_{i \in \mathcal{I}_x} L_i \, (1 - \widehat{P}_i)}{l_x}$$

onde L_i é a serie biométrica de anos vividos 'years lived', l_x os sobreviventes á idade x, ambas as dúas funcións da táboa de mortalidade abreviada e $\widehat{P}_i$ son as prevalencias de limitación no chanzo $[i, i+n_i)$.

Para dar una medida de variabilidade das EVL, pode empregarse un estimador da varianza do estimador segundo o deseño da mostra. Aínda que as $\widehat{P}_i$ non son independentes so deseño, adoito están correlacionadas negativamente polo que ten senso, aínda que de forma aproximada, a seguinte igualdade para o cálculo da varianza da EVL

$$VAR_p[EVL_x] = \frac{\sum_{i \in \mathcal{I}_x} L_i^2 \, VAR_p(1 - \widehat{P}_i)}{l_x^2}$$

que pode ser estimada mediante

$$var_p[EVL_x] \coloneqq \frac{\sum_{i \in \mathcal{I}_x} L_i^2 \, var_p(1 - \widehat{P}_i)}{l_x^2}$$

sendo $var_p(1 - \widehat{P}_i)$ un estimador inesgado de $VAR_p(1 - \widehat{P}_i)$ cando menos aproximadamente.

A fonte de datos para estimar as prevalencias é a Enquisa estrutural a fogares (EEF) 2014 [5]. Considérase que unha persoa está limitada se declara levar seis meses ou máis limitada debido a un problema de saúde para realizar as actividades que se fan a cotío (1-gravemente limitada, 2-limitada non gravemente, 3-nada limitada). A EVL pode referirse entón aos anos vividos sen ningunha limitación, ou sen limitación grave (EVLG), segundo P_i inclúa as opcións 1+2 ou só a 1, respectivamente.

No entanto, a EEF ten un deseño complexo e mesmo as prevalencias de limitacións son un estatístico complexo, estimáronse mediante as estatísticas da enquisa coas probabilidades de inclusión calibradas [5]. Isto aconsellaba procurar un método alternativo para dar unha medida de variabilidade da EVL.

Por outra banda, a EEF non pregunta a menores de 16 anos sobre a limitación, polo que é preciso imputar as prevalencias para os chanzos máis novos co gallo de ofrecer a EVL para as idades correspondentes. No IGE, estabamos a realizar a imputación mediante unha técnica directa, apriorística, na que considerabamos 0 a prevalencia ao nacer, asignamos a de [16, 20) a [15, 20) e para os chanzos anteriores a este aplicabamoslle unha prevalencia constante e igual á metade da correspondente a [16, 20).

Por todo o anterior e tendo en conta a variabilidade das estimativas das prevalencias específicas, optouse por construír o seguinte modelo 'logit'

$$\ln \frac{\widehat{P}_i}{1 - \widehat{P}_i} = \alpha + \beta_1 * Idade + \beta_2 * Sexo + \varepsilon_i,$$

onde consideramos que ε_i ten media 0 e variancia, a herdada da sondaxe. Como variábel Idade consideramos os centros dos chanzos, marcas de clase, correspondentes a $\mathcal{I}_{16} = \{16, 20, \cdots, 75\}$ e no derradeiro, neste caso $\omega = 75$, considerouse 87'5 dado que a esperanza aos 75 anos é superior a 13 anos. O Sexo é unha binaria con valor 0 para os homes e 1 para as mulleres.

Con este modelo dispomos doutro método para estudar a variabilidade da EVL herdada da enquisa, mediante unha remostraxe tipo bootstrap paramétrico, e mais suavizar e imputar doutro xeito as prevalencias. Axustamos o modelo por cadrados mínimos ponderados, isto é, cunha matriz de covariancias diagonal. Para estudar a variancia da loxita, aproximámola do seguinte xeito

$$\ln \frac{\widehat{P}_i}{1 - \widehat{P}_i} = \ln \frac{P_i}{1 - P_i} + \frac{\widehat{P}_i - P_i}{P_i(1 - P_i)} - \frac{1 - 2P_i'}{P_i'(1 - P_i')} \left(\widehat{P}_i - P_i\right)^2$$

onde P_i' é un punto intermedio entre $\widehat{P}_i$ e P_i e, xa que logo,

$$VAR\left[\ln\frac{\widehat{P}_i}{1-\widehat{P}_i}\right] \cong \frac{VAR_p\left[\widehat{P}_i\right]}{P_i^2(1-P_i)^2}$$

Para comparar estes dous métodos, estimouse $VAR_p\left[\widehat{P}_i\right]$ coas probabilidades de inclusión xa calibradas, que era a información dispoñíbel, mediante o paquete en R **survey**. Para a loxita, no denominador substituíuse tamén a prevalencia específica pola súa estimativa respectiva.

Para o cálculo dos intervalos de confianza da EVL polo método de Sullivan empregouse a aproximación normal e considerouse un nivel de confianza do 95%. Na remostraxe, en cada reiteración, para cada chanzo e sexo, calculouse a loxita predita dando valores ás variábeis Idade e Sexo e mais xerando un valor aleatorio gaussiano correspondente ao erro. Mediante a transformación inversa da 'logit', a loxística, obtivemos a prevalencia específica. A partir da réplica de serie de prevalencias, calculouse a EVL para as diferentes idades de $\mathcal{I}_{16}$. Reiteramos este procedemento ata $B = 999$ e consideramos os cuantís correspondentes ás pormilaxes 25 e 975.

3. RESULTADOS

Os resultados do axuste do modelo para as prevalencias son $\hat{\alpha} = -3{,}093; \widehat{\beta_1} = 0{,}0396; \widehat{\beta_2} = 0{,}2669$. Os gráficos da figura 1 representan, para cada sexo, as prevalencias específicas da enquisa, $\widehat{P}_i$, pola poligonal e as preditas polo modelo, $\widehat{\widehat{P}}_i$, mediante a curva suave, respectivamente. A necesidade de suavización resulta máis evidente nas mulleres que nos homes á vista dos gráficos. No caso das prevalencias de limitación grave (figura non mostrada), a necesidade é aínda maior.

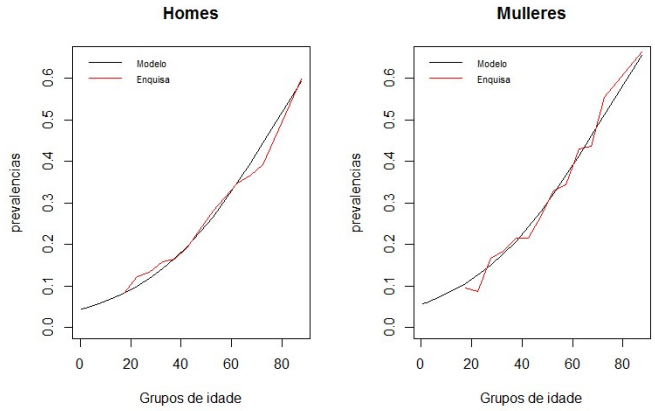


Figura 1: Prevalencias orixinais e suavizadas polo modelo segundo o sexo

Representamos na figura 2, os intervalos de "confianza" das EVL polos métodos de remostraxe e de Sullivan.

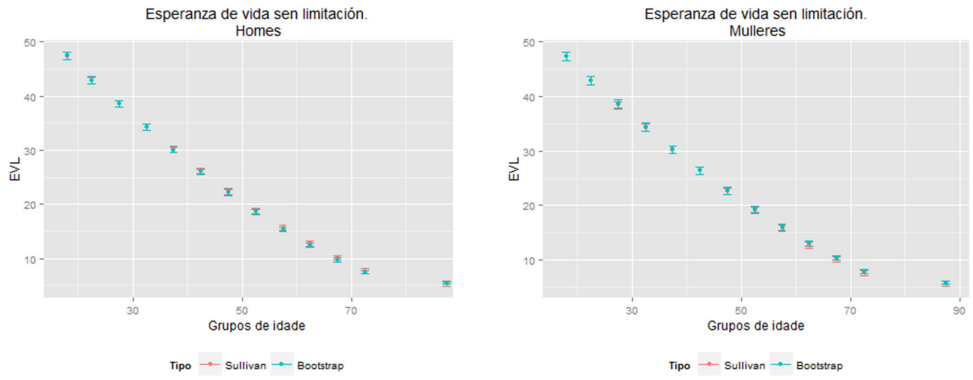


Figura 2: EVL e intervalos de "confianza" do 95% mediante os métodos tradicional e de remostraxe

Na seguinte táboa podemos comparar os resultados das dúas imputacións para o cálculo da EVLG e da EVL. Tamén se presenta a esperanza de vida.

	Homes	Mulleres
Esperanza de vida ao nacemento	79,5	85,6
Esperanza de vida ao nacemento sen limitación grave		
Imputación apriorística	77,2	83,3
Imputación modelo	77,1	83,1
Esperanza de vida ao nacemento sen limitación		
Imputación apriorística	61,5	61,4
Imputación modelo	61,3	61,0

Táboa 1: Esperanza de vida e mais EVL e EVLG con imputacións apriorística e polo modelo segundo sexo

3. CONCLUSIÓNS

As técnicas estatísticas poden complementar satisfactoriamente as fontes, p.ex. a enquisa por mostraxe, sen un grande custo adicional. Neste caso permitiu dispor de prevalencias en chanzos etarios non incluídos nos resultados da enquisa, imputándoos dun xeito total, como nos correspondentes á infancia, ou parcial, como a imputación do chanzo [15, 20) do que só se dispuña da prevalencia para [16, 20) e, logo, calcular as EVL des do nacemento, con $\mathcal{I}_0 = \{0, 1, 5, 10, 15, \cdots, 75\}$. Tamén había permitir proxectar os subchanzos iniciais do chanzo derradeiro, p.ex. [75, 80) e [80, 85) co que se podería alongar o cálculo da EVL.

Obtivemos unha medida da variabilidade da EVL alternativa ao método de Sullivan e unhas prevalencias suavizadas que facilitan a súa interpretación sanitaria. Unha característica da fonte consiste en que a limitación é autodeclarada polo entrevistado, isto podería explicar que exista unha diferenza de 6 anos na esperanza de vida e na EVLG entre os dous sexos e non hai diferenza na esperanza de vida sen limitación. As prevalencias da limitación percibida polas mulleres son superiores, por termo medio, ás dos homes.

AGRADECEMENTOS

Ao Servizo de Estatísticas Sociais do IGE por procuraren a información e a Solmary Silveira Calviño e mais Juan Gómez Duaso, copartícipes no 'project' **demogal**.

REFERENCIAS

[1] Consellería de Sanidade (2014) *Epidat 4: Ayuda de Demografía*. In http://www.sergas.es/Saude-publica/Documents/1896/Ayuda_Epidat_4_Demografia_Octubre2014.pdf

[2] Sullivan, D.F. (1971) A Single Index of Mortality and Morbidity. *HSMHA Health Rep*. 86(4): 347–354.

[3] Martín-Vila, M. et al. (2013) Las tablas de mortalidad de Galicia y la Región Norte de Portugal empleando el software R. *Programa e resumos dos Encontros de Biometría,* Braga

[4] Calvo, E. et al. (2016) **demogal**: paquete para proyectar flujos y stocks de población *II Encontros galaico-portugueses de Biometría,* Santiago de Compostela

[5] IGE Enquisa estrutural a fogares. Metodoloxía e cuestionario. In http://www.ige.eu/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0205021

Comunicacións orais

Sesión IX: Métodos estatísticos en Ecoloxía

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

UM MODELO PARA A ABUNDÂNCIA DE PEIXES COM DEPENDÊNCIA ESPACIAL BASEADA NO NÚMERO DE CONFLUÊNCIAS ENTRE RIACHOS

Júlio César Pereira[12], Maurício Cetra[1] e Martin Ridout[2]

[1]Universidade Federal de São Carlos - Brasil
[2]University of Kent - UK

RESUMO

A modelagem quantitativa da abundância e distribuição das espécies de peixes de riachos pode apresentar dificuldades devido a particularidades dos dados, tal como correlação espacial entre as observações nos diferentes pontos de coleta. No presente trabalho propôs-se um modelo Poisson log-linear com uma componente espacial que depende do número de confluências para contagens de *Phalloceros reisi*. Dois casos particulares desse modelo, em que não se leva em conta a dependência espacial, também foram ajustados. O primeiro modelo se revelou mais realista apresentando assim melhor performance na predição da abundância.

Palavras e frases chave: Contagens de peixes, correlação espacial, rede de rios e afluentes, abordagem bayesiana.

1. INTRODUÇÃO

Frequentemente estudos sobre a fauna de peixes em água doce buscam entender a relação das espécies com as características do ambiente onde vivem [2]. Uma vez que a variação nos fatores que constituem o ecossistema provoca variação na composição das espécies, tais estudos são importantes, pois permitem levantar os fatores que determinam a ocorrência bem como conhecer o tipo de habitat favorável a cada espécie [6]. Além disso, o conhecimento advindo desses estudos pode ser usado na predição da distribuição e abundância das espécies, que por sua vez são utilizadas como indicadores da integridade do ecossistema e também servem de subsídio para medidas de manejo, políticas de conservação, proteção ambiental e manutenção da biodiversidade [1].

Um dos obstáculos quando se busca modelar a abundância de peixes de riachos e suas relações com o ambiente, está na dificuldade de acomodar em um modelo estatístico as particularidades de dados que são coletados nesses complexos ecossistemas [1]. Uma dificuldade na modelagem da abundância de peixes de riachos se dá pela correlação que pode estar presente nas contagens e o fato de que ela não é simplesmente mensurada por modelos tradicionais de correlação espacial, mas podem depender da calha do curso d'água e suas confluências [5, 8].

Nesse sentido, o objetivo deste trabalho foi propor um modelo para a abundância de peixes levando-se em conta a correlação espacial baseada na distância dada pelo número confluências entre os diferentes locais de amostragem.

2. DADOS

Neste trabalho foi modelada a distribuição e abundância em número de indivíduos de *Phalloceros reisi*, uma pequena espécie, cujo gênero tem ampla distribuição geográfica no leste e sul do Brasil. Os dados foram coletados nas bacias dos Rios Sorocaba e Paranapanema no estado de São Paulo, Brasil. As coletas ocorreram entre os meses de Julho e Novembro (período seco) do ano de 2010. Foram amostradas a ictiofauna em seções de 100 m de dez afluentes do Rio Sorocaba e oito na bacia do Rio Paranapanema. Os locais de coleta, bem como as contagens de *Phalloceros reisi* são representados na Figura 1. Essa figura ilustra também a maneira como são conectados os afluentes, e a partir dela é possível se obter o número de confluências entre quaisquer dois pontos de coleta. Dentre os locais amostrados, o número de confluências variou entre 1 e 9. Além do número indivíduos obtidos nas capturas, foram mensuradas variáveis ambientais para cada localização. Essas variáveis são referentes a tipo de substrato (em porcentagem/m^2), estabilidade do barranco, uso do solo ao entorno e morfologia do canal.

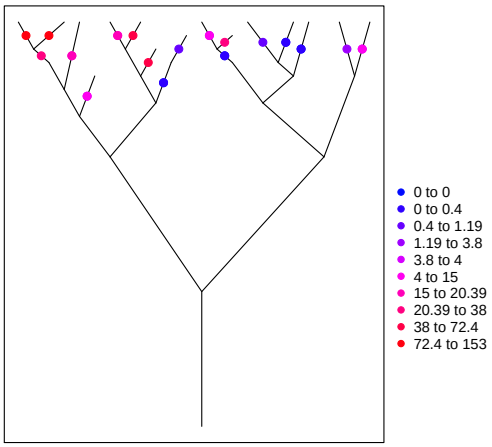


Figura 1: Riachos das bacias dos rios Sorocaba e Paranapanema, pontos de coleta representados pelos círculos e contages de peixes identificadas pelas cores dos círculos.

3. MODELOS

Denotando por $Y(s)$ a contagem de indivíduos observados no local s, considerou-se que $Y(s_i)$, $s_1, ..., s_n$ segue uma distribuição de Poisson com valor esperado $\mu(s)$. Porém, a fim de se incluir as covariáveis que poderiam afetar as contagens, bem como a possível dependência espacial entre os locais de coleta, adotou-se a função de ligação logarítmica para relacionar o preditor linear às contagens. Dessa forma tem-se o assim chamado modelo Poisson-espacial definido pelas equações 1 e 2.

$$Y(s) \sim Poisson(\mu(s)) \tag{1}$$

$$\boldsymbol{\ln \mu(s) = X(s)\beta + v(s)} \tag{2}$$

em que $X(s)\beta$ é uma componente fixa do preditor linear, $X(s)$ é um vetor linha de dimensão $p \times 1$ contendo os níveis das covariáveis observadas na localização s e β é o correspondente vetor de parâmetros; v é uma componente aleatória espacialmente estruturada, assumida seguir um processo gaussiano, com média zero, variância σ^2 e função de correlação exponencial $\rho(h;\alpha) = \exp\left(\frac{-3h}{\alpha}\right)$. A parametrização assumida para a função de correlação de acordo com [9] permite interpretar o parâmetro α como o alcance efetivo, isto é, a distância para a qual a autocorrelação é aproximadamente igual a $0,05$.

Um aspecto peculiar no modelo adotado aqui, concentra-se na distância utilizada na função de correlação. Como foi conjecturado que a dependência entre as contagens de duas localizações quaisquer poderia depender do número de confluências entre elas, a distância h na função de correlação foi medida em termos do número de confluências entre as localizações, sendo, portoanto, uma distância discreta.

Um segundo modelo também foi ajustado a fim de se fazer uma comparação com o modelo definido pelas equações 1 e 2. Neste segundo modelo considerou-se o valor do parâmetro α da função de correlação tendendo a zero. Dessa maneira, a respectiva função de covariâncias assume a forma:

$$\lim_{\alpha \to 0} Cov(h) = \begin{cases} \lim\limits_{\alpha \to 0} \sigma^2 \exp\left(\dfrac{-3h}{\alpha}\right) = \sigma^2 & \text{para } h = 0 \\ \lim\limits_{\alpha \to 0} \sigma^2 \exp\left(\dfrac{-3h}{\alpha}\right) = 0 & \text{para } h > 0. \end{cases} \tag{3}$$

Observa-se pela equação (3) que a covariância entre duas localizações quaisquer, separada por uma distância $h > 0$, torna-se nula, enquanto que quando $h = 0$ tem-se uma variância σ^2. Em outras palavras, a matriz de covariâncias, torna-se diagonal com elementos σ^2. Nesse segundo modelo os efeitos aleatórios $v(s_1), \ldots v(s_n)$ são mutuamente independentes, normalmente distribuidos com variância comum igual a σ^2. O modelo Poisson log-linear assim definido é usualmente dito incorporar variação extra ou super dispersão [4]. Esse modelo é referenciado nestre trabalho como Poisson com variância extra.

Um terceiro modelo ajustado aos dados foi o modelo de Poisson, considerando-se somente efeitos fixos no preditor linear, chamado de Poisson com efeitos fixos. Esse modelo é definido por:

$$Y_i \sim Poisson(\mu_i) \tag{4}$$

$$\mathbf{\ln \mu_i = X_i \beta} \tag{5}$$

em que X_i são as covariáveis medidas na localização i e β o vetor de coeficientes dessas covariáveis.

3.1 Processo de inferência

Do ponto de vista da inferência bayesiana, a especificação do modelo se completa com a escolha da distribuição a priori dos parâmetros. Para os componentes do vetor de parâmetros β assumiu-se uma priori vaga ($\pi(\beta) \propto 1$) para representar a falta de informação sobre como cada covariável influência a média da distribuição de Poisson. Para o alcance α assumiu-se uma priori exponencial $\alpha \sim Exp(\lambda)$. E de acordo com [7] considerou-se que o alcance se concentra em torno da metade da máxima distância entre as observações, assim tomou-se como média a priori para α a metade da máxima distância observada. Para a variância σ^2 assumiu-se uma distribuição qui-quadrado escalonada inversa, $Scale - Inv - \chi^2(\nu_1, \nu_2)$. De acordo com [4] os hiperparâmetros ν_1 e ν_2 podem ser interpretados como uma adivinhação e o grau de certeza sobre o valor de σ^2, sendo quanto menor ν_2 menor o grau de confiança. Foram assumidos $\nu_1 = 2$ e $\nu_2 = 10$.

Assumindo independência a priori entre os parâmetros, a distribuição a posteriori conjunta é dada por $\pi(\beta, \sigma^2, \alpha \mid y) \propto L(\beta, \sigma^2, \alpha \mid y)\pi(\beta)\pi(\sigma^2)\pi(\alpha)$, em que $L(\beta, \sigma^2, \alpha \mid y)$ denota a função de verossimilhança.

Para amostrar da distribuição a posteriori dos parâmetros foi utilizado o algoritmo Langevin-Hastings, implementado no pacote geoRglm. Esse algoritmo usa informação local da distribuição alvo para fazer proposta e tem apresentado bom desempenho na prática quando comparado aos algoritmos Metropolis-Hastings com passeio aleatório [3]. Como o pacote geoRglm não permite outras formas de distâncias além da distância euclidiana, foram necessárias adaptações nas funções do pacote a fim de se poder ajustar e predizer o modelo Poisson-espacial usando a distância dada pelo número de confluências.

Uma comparação da performance dos modelos foi realizada quanto à capacidade de predição. Para tanto foram computados os intevalos de credibilidade 95% da distribuição preditiva a posteriori e foi analisado sua performance em cobrir a abundância observada em cada localização.

4. RESULTADOS E DISCUSSÃO

Os modelos descritos na seção anterior foram aplicados às contagens de *Phalloceros reisi*. Em uma preliminar análise foram selecionadas como covariáveis a serem inclusas no modelo, aquelas cujos coeficientes foram significativos quando ajustado um modelo linear generalizado clássico. Dessa forma, as covariáveis selecionadas foram porcentagem de vegetação por m^2 como substrato, pH e velocidade da água.

Na Figura 2 são apresentadas as predições obtidas em cada local de amostragem por cada um dos três modelos ajustados. Comparando as predições, é possível observar para o modelo Poisson-espacial somente três dos intervalos não contêm as contagens observadas nos locais de coleta. Para o modelo Poisson com variância extra, cinco intervalos não cobrem as contagens observadas. Por outro lado, considerando-se o modelo Poisson, incluindo-se somente efeitos fixos, onze intervalos não cobrem o valor observado. Foi conjecturado, assim, que o modelo Poisson incluindo-se somente efeitos fixos não teve boa performance nas predições por não considerar nenhum tipo de variação extra presente nos dados. Enquanto que o modelo de Poisson com extra variância, leva-se em conta variação adicional presente nos dados, apesar de não se levar em conta correlação espacial. Por fim o modelo Poisson-espacial obteve o melhor desempenho dentre os três modelos por levar em conta variação extra e, além disso, por considerar tal variação na forma de uma componente aleatória espacialmente estruturada, que é quantificada talvez de uma forma mais realista, em termos de distância que depende da paisagem (numero de confluencias) da área em estudo.

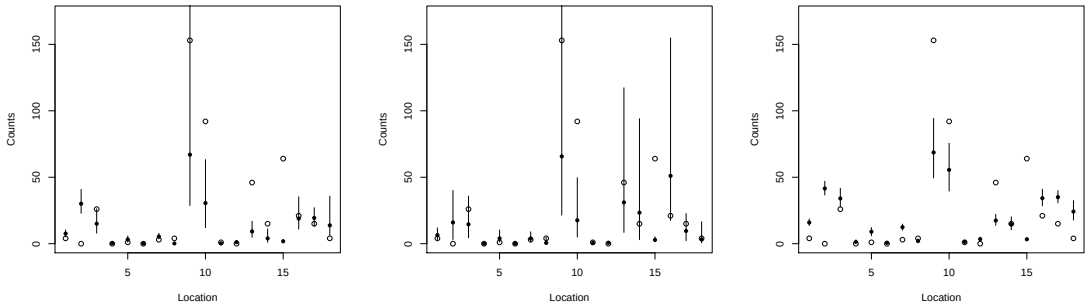


Figura 2: Contagens observadas (○), intervalo de credibilidade 95% (linha vertical) e mediana da distribuição preditiva (●), modelo Poisson espacial (à esquerda), Poisson com variância extra (centro) e Poisson com efeitos fixos (direita).

AGRADECIMENTOS

O primeiro autor é grato ao CNPq - Brasil pelo apôio concedido na forma de bolsa de pós-doutorado. O segundo autor é grato à FAPESP (processo: 2009/53056-8).

Referências

[1] Boone, E. L., Stewart-Koster, B., Kennard, M. J.(2011). A hierarchical zeroinflated Poisson regression model for stream fish distribution and abundance. *Environmetrics* 23, 207–218.

[2] Cruz B.B., Miranda, L. E., Cetra, M. (2013). Links between riparian landcover, instream environment and fish assemblages in headwater streams of south-eastern Brazil. *Ecology of Freshwater Fish*, 1–10.

[3] Christensen, O. F., Moller, J., Waagepetersen, R. (2001). Geometric ergodicity of Metropolis-Hastings algorithms for conditional simulation in generalised linear mixed models. *Methodology and Computation in Applied Probability* 3, 309–327.

[4] Diggle, P. J., Ribeiro, P. J. (2007). *Model-based Geostatistics*.Springer Series in Statistics, Springer, New York.

[5] Ganio, L. M., Torgensen, C. E., Gresswell, R. E. (2005). A geostatistical approach for describing spatial pattern in stream networks. *Frontier in Ecology and the Environment* 3, n. 3, 138–144.

[6] Kanaiwa, M., Inque, T., Yamamoto, A. (2011). Determinants of fish species composition in Abashiri River. Scientific Report. *Biological processes/Disturbances by oil and gas development* 36, 243 – 245.

[7] Schmidt, A. M., Rodriguez, M. A., Capistrano, E. S. (2015) Population counts along elliptical habitat contours: Hierarchical modeling using Poisson lognormal mixtures with nonstationary spatial structure. *Ann. Appl. Stat.* 9 (3), 1372–1393.

[8] Ver Hoef, J. M., Peterson, E. E. (2010) A Moving Average Approach for Spatial Statistical Models of Stream Networks. *Journal of the American Statistical Association* 105, n. 489, 6–18.

[9] Ver Hoef, J. M., Peterson, E. E., Clifford, D., Shah, R. (2014). SSN: An R Package for Spatial Statistical Modeling on Stream Networks. *Journal of Statistical Software* 56(3), 1–43.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de julio, 1 y 2 de julio de 2016

Un nuevo método para separar componentes de la disimilitud entre múltiples comunidades biológicas basada en datos de abundancia: variación equilibrada vs. gradientes de abundancia

Andrés Baselga[1]

[1] Departamento de Zoología, Facultad de Biología, Universidad de Santiago de Compostela, c/ Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela

RESUMEN

Cuando se toma en cuenta la variación en la abundancia de la especies, la disimilitud entre comunidades biológicas puede ser dividida en dos componentes que representan (i) la variación equilibrada en la abundancia, en la que los individuos de algunas especies en un sitio se sustituyen por el mismo número de individuos de otras especies en otro sitio, y (ii) los gradientes de abundancia, por lo que algunas la abundancia de las especies decrece unidireccionalmente de unos sitios a otros [1-2]. Dicho método estaba disponible para disimilitud entre pares de comunidades biológicas, pero no para la disimilitud de múltiples comunidades. En este trabajo se presentan nuevas extensiones de los índices de disimilitud de Bray-Curtis y Ruzicka para múltiples sitios. Estos nuevos índices se pueden dividir en componentes de (i) variación equilibrada y (ii) gradientes de abundancia. Se muestra un ejemplo evaluando la disimilitud entre múltiples sitios para comunidades de aves y mariposas en un mosaico de hábitats, ilustrando la utilidad de la comparación de la disimilitud entre múltiples sitios basada en datos de abundancia con su correspondiente medida basada en datos de ocurrencia [3] para caracterizar la heterogeneidad biótica a través de múltiples sitios. En general, el nuevo método puede ser útil cuando la cuestión de interés es la disimilitud entre un conjunto de múltiples comunidades biológicas (en el espacio, el tiempo u otro), ya que la separación de los componentes de (i) variación equilibrada de la abundancia (es decir, patrones de sustitución) y de (ii) los gradientes de abundancia (es decir, patrones anidados o de subconjuntos) puede arrojar luz sobre los procesos que subyacen a la variación en la abundancia de especies.

.

Palabras y frases clave: abundancia, diversidad beta, índice de Bray-Curtis, disimilitud entre múltiples sitios, sustitución , subconjunto.

REFERENCIAS

[1] Baselga A. (2013) Separating the two components of abundance-based dissimilarity: balanced changes in abundance vs. abundance gradients. *Methods in Ecology and Evolution* 4: 552–557.

[2] Legendre P. (2014) Interpreting the replacement and richness difference components of beta diversity. *Global Ecology and Biogeography* 23: 1324–1334.

[3] Baselga A. (2010) Partitioning the turnover and nestedness components of beta diversity. Global Ecology and Biogeography 19: 134-143.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

OS PATRÓNS NA VARIACIÓN LATITUDINAL DA DIVERSIDADE BETA EN VERTEBRADOS SON SEMELLANTES ENTRE CLADOS E SUXIREN CAUSAS HISTÓRICAS COMÚNS

Adrián Castro-Insua[1], Carola Gómez-Rodríguez[1] e Andrés Baselga[1]

[1]Departamento de Zooloxía, Facultade de Bioloxía, Universidade de Santiago de Compostela, Rúa Lope Gómez de Marzoa, 15782, Santiago de Compostela

RESUMO

Entender o efecto potencial dos factores abióticos na variación na composición de especies (a denominada diversidade beta [1]) é fundamental para entender por que a diversidade rexional varía entre latitudes. Porén, usar análises correlativas para tentar entender os procesos que controlan a biodiversidade pode ser insuficiente, porque as variábeis que se corresponden con diferentes hipóteses (ex. o clima actual, mudanzas climáticas no pasado, limitacións na dispersión post-glacial) están mutuamente correlacionadas. Por esta razón, suxerimos aquí que pode ser útil procurar puntos de inflexión na variación latitudinal da diversidade (aplicando técnicas estatísticas de uso pouco común en ecoloxía, neste caso, mediante regresións por partes ou *piecewise regressions*). Estes puntos de inflexión poderían dar pistas sobre outros procesos que non aparecerían se se usasen unicamente análises correlativas. Neste traballo, analizamos se existen puntos de inflexión similares nos patróns latitudinais da diversidade beta en distintos grupos de vertebrados con ecoloxías e capacidades de dispersión moi diversas. En aves, morcegos e mamíferos non voadores aparecen puntos de inflexión na variación latitudinal da diversidade beta parecidos aos que se observan en anfibios [2]. Estes puntos de inflexión tamén se poden encontrar en variábeis como a altitude media e as diferenzas de temperatura entre o último máximo glacial (LGM) e o clima actual, pero non nas variábeis climáticas actuais, polo que parece que as glaciacións no LGM puideron ter algún papel nestes patróns na diversidade beta. Este exemplo mostra como o uso de técnicas estatísticas novas ou diferentes pode proporcionar perspectivas que permitan avanzar tamén no campo da macroecoloxía.

Palabras e frases clave: regresións por partes, puntos de inflexión, diversidade beta, glaciacións no Pleistoceno, vertebrados

REFERENCIAS

[1] Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs,* **30**, 280–338.

[2] Baselga, A., Gómez-Rodríguez, C. & Lobo, J.M. (2012) Historical legacies in world amphibian diversity revealed by the turnover and nestedness components of beta diversity. *PLoS One,* **7**, e32341.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

PREDICCIÓN DEL PATRÓN DE ASENTAMIENTO LARVARIO EN LAS RÍAS GALLEGAS CON GAM PARA DATOS REALES Y FUNCIONALES

Isabel Fuentes-Santos [1], Uxio Labarta [2], X. Antón Álvarez-Salgado [3] y Mª José Fernández-Reiriz[4]

[1]Instituto de Investigaciones Marinas - Consejo Superior de Investigaciones científicas.
isafusa@iim.csic.es
[2]Instituto de Investigaciones Marinas - Consejo Superior de Investigaciones científicas.
labarta@iim.csic.es
[3]Instituto de Investigaciones Marinas - Consejo Superior de Investigaciones científicas.
xsalgado@iim.csic.es
[4]Instituto de Investigaciones Marinas - Consejo Superior de Investigaciones científicas.
mjreiriz@iim.csic.es

RESUMEN

La captación de semilla en sustratos artificiales es crucial para garantizar la sostenibilidad del cultivo de mejillón en las rías gallegas. Esta necesidad ha motivado el estudio del ciclo de asentamiento larvario del mejillón y su relación con condiciones ambientales y meteorológicos. Este trabajo muestra la utilizad de los modelos GAM para datos reales y funcionales en el desarrollo de un sistema de predicción del patrón de asentamiento larvario a partir de variables meteorológicas medidas a diario por fuentes oficiales.

Palabras e frases clave: asentamiento larvario, datos funcionales, modelos aditivos generalizados, radiación solar.

1. INTRODUCCIÓN

El cultivo de mejillón en Galicia consiste en el engorde de juveniles en batea y depende de la magnitud del reclutamiento larvario, ya sea sobre sustratos naturales (litoral rocoso) o artificiales (colectores artificiales).

La necesidad de obtener semilla mediante cuerdas colectoras para mantener la sostenibilidad de la miticultura en Galicia ha puesto de manifiesto la importancia de conocer los patrones de asentamiento larvario y los factores ambientales que lo regulan, para poder evaluar y mejorar la estrategia actual de captación de semilla. Con este fin en el año 2006 se implantó un programa de monitorización del asentamiento larvario y captación de semilla de mejillón en la Ria de Ares-Betanzos.

El objetivo de este trabajo es predecir el ciclo e intensidad del asentamiento larvario en la Ría de Ares-Betanzos en función de condiciones meteorológicas. Para ello hemos utilizado modelos aditivos generalizados (GAM) [1], [2] para datos reales, que combinan la interpretabilidad de los modelos lineales y la flexibilidad de la inferencia no paramétrica, y su extensión al marco de los datos funcionales (FGKAM, functional generalizad kernel additive models) [3].

2. METODOLOGÍA

Se ha seguido el asentamiento larvario en el polígono de Arnela (Ría de Ares-Betanzos) mediante muestreos semanales en los periodos 2009-2010 y 2012-2015. La intensidad de asentamiento ,N, se mide en individuos por metro de cuerda (ind/m). Consideramos que se ha producido asentamiento, $Y = 1$, cuando $N \geq 500$.

Como variables explicativas hemos utilizado datos diarios de radiación solar $(R, MJ/(m^2 dia))$, recogidos en la estación meteorológica CIS-Ferrol (http://www2.meteogalicia.gal), datos diarios de velocidad $(W, m/s)$ y dirección (θ) del viento recogidos en la boya de Vilano (http:/www.puertos.es), y la suma de los aportes diarios $(Q, m^3/s)$ de los ríos Eume (ENDESA S.A) y Mandeo (Augas de Galicia). Estas variables han sido seleccionadas teniendo en cuenta la influencia de la radiación solar en el ciclo reproductivo de los invertebrados marinos [4], [5] y la influencia del aporte de los ríos y el viento en los patrones de asentamiento larvario en la Ría de Ares-Betanzos [6].

La probabilidad de asentamiento larvario se ha estimado mediante ajustes GAM con familia binomial y link logit [1], [2]. El análisis de bondad de ajuste de distintos modelos nos ha llevado a seleccionar el siguiente ajuste

$$P\left(Y = 1|R\right) = H\left(\alpha + f(R)\right) = \frac{\exp\left(\alpha + f(R)\right)}{1 + \exp\left(\alpha + f(R)\right)} \tag{1}$$

donde R es la radiación solar media de los 30-45 días previos al muestreo de asentamiento larvario, y la función f se representa mediante splines planos penalizados (TPRS).

Para estimar la intensidad del asentamiento hemos utilizado ajustes GAM en dos etapas, con medias quincenales de las variables explicativas, y ajustes GAM para datos funcionales [3], con curvas que recogen el comportamiento de las condiciones ambientales durante un mes (15-45 días anteriores al muestreo). El análisis de bondad de ajuste de los candidatos evaluados nos ha llevado a seleccionar el siguiente ajuste:

$$E\left[N|fR, fQ\right] = H\left(\beta + g_1(fR) + g_2(fQ)\right) \tag{2}$$

donde fR y fQ son las curvas mensuales de radiación solar y caudal de ríos, y $g_j, j = 1, 2$ se obtienen mediante estimadores núcleo. En este caso hemos usado la familia binomial negativa con link logarítmico, para reducir el efecto de la sobredispersión.

3. RESULTADOS

En primer lugar hemos ajustado el modelo (1) en el periodo 2009-2014 y utilizado este ajuste para predecir la probabilidad de asentamiento larvario durante 2015: La figura 1 muestra que el ajuste adelanta 10 días el inicio del periodo de asentamiento larvario, y predice el fin para una semana después del último muestreo. Además de predecir el inicio y fin del periodo de asentamiento, este ajuste identifica un umbral de radiación solar, $R \approx 11 MJ/(m^2 dia)$, por debajo del que no se produce asentamiento larvario.

La Figura 2 muestra el patrón de asentamiento larvario observado durante 2015 y la predicción a 15 días a partir de las curvas mensuales de radiación solar y aporte de ríos (2). El modelo identifica correctamente el periodo de asentamiento y proporciona un ajuste razonable de la intensidad, aunque tiende a infraestimar los picos más altos.

3. CONCLUSIONES

En este trabajo se han aplicado ajustes GAM para datos reales y funcionales para desarrollar un sistema de predicción de asentamiento larvario. Hemos visto que la radiación solar, que apenas había sido considerada entre los factores que determinan la producción de larvas de mejillón, permite predecir con un mes de antelación el inicio y fin del periodo de captación de semilla.

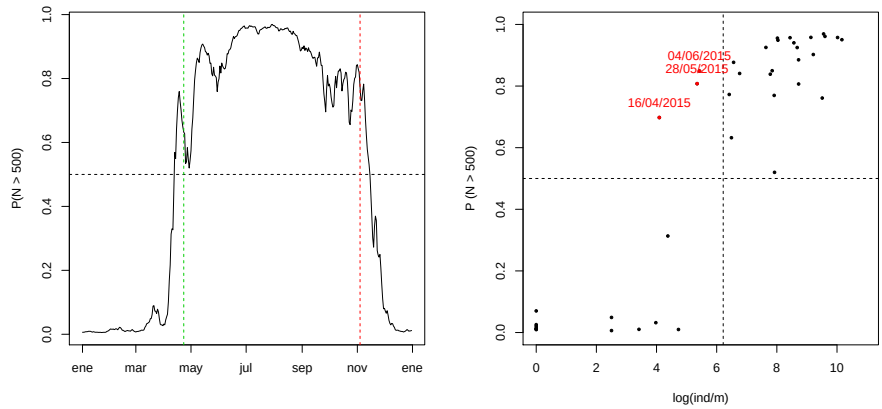


Figura 1: Izquierda: Predicción de la probabilidad de asentamiento larvario para el año 2015. Línea verde: inicio del periodo de asentamiento (23/04/2015); línea roja: fecha del último muestreo (04/11/2015). Derecha: comparación entre densidad observada y clasificación proporcionada por el modelo. Rojo: muestreos con N $<$ 500 ind/m clasificados incorrectamente.

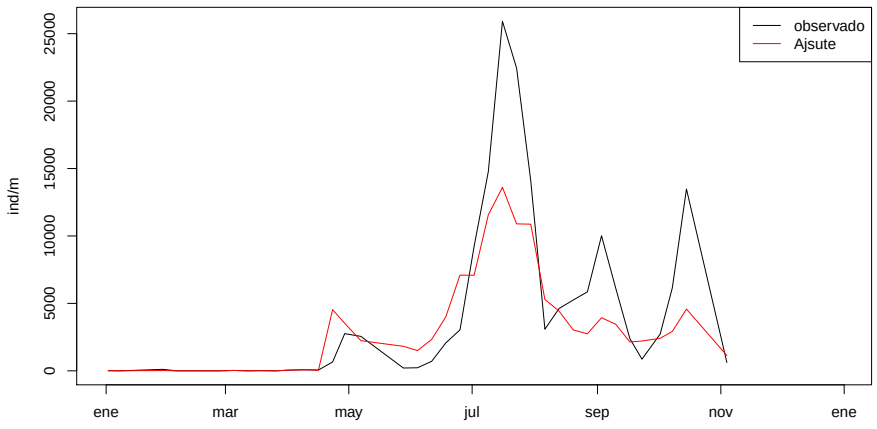


Figura 2: Comparación entre densidad de asentamiento observada en los muestreos (negro) y predicción (rojo) proporcionada por el modelo (2)

Los datos funcionales permiten incorporar en los modelos biológicos toda la información derivada de la monitorización continua de las condiciones ambientales. Esto resulta de gran utilidad en nuestro caso dado que el asentamiento es el final de la etapa larvaria del mejillón, que puede durar un mes.

En conclusión. la flexibilidad de los ajustes GAM y los datos funcionales nos han permitido desarrollar un sistema de predicción de asentamiento larvario de bajo coste, que puede resultar de gran utilidad en el diseño de estrategias de captación de semilla.

AGRADECIMIENTOS

Queremos agradecer la colaboración de la empresa PROINSA S.A., en particular a H. Regueiro y M. García por su labor en los muestreos. Este trabajo ha sido financiado por el convenio entre PROINSA y CSIC CSIC 0704101100001 y los proyectos CSIC 201540E107 y ClimeFish (EU 677039).

Referencias

[1] Hastie, T. J., Tibshirani, R. J. (1990). G*eneralized Additive Models*. Chapman and Hall/CRC, London.

[2] Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, London.

[3] Febrero-Bande, M., González-Manteiga, W. (2013). Generalized additive models for functional data. *Test* 22, 278–292

[4] Fiorillo I, Rossi S, Alva V, Gili JM, López-González PJ (2013) Seasonal cycle of sexual reproduction of the Mediterranean soft coral Alcyonium acaule (Anthozoa, Octocorallia). *Marine Biology* 160:719–728

[5] Van Woesik R, Lacharmoise F, Köksal S (2006) Annual cycles of solar insolation predict spawning times of Caribbean corals. *Ecology Letters* 9:390–398

[6] Peteiro LG, Labarta U, Fernández-Reiriz MJ, Álvarez-Salgado XA, Filgueira R, Piedracoba S (2011). Influence of intermittent-upwelling on Mytilus galloprovincialis settlement patterns in the Ría de Ares-Betanzos. *Marine Ecology Progress Series* 443:111–127

Comunicacións orais

Sesión X: Modelización conxunta

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

MODELAGEM CONJUNTA DE DADOS LONGITUDINAIS E SOBREVIVÊNCIA EM PACIENTES COM SEPSE

Omar Pereira[1], Marcos Peres[1], Cristina Kuroda[2], Elza Kimura[2], Isolde Previdelli[1]

[1]Programa de Pós-Graduação em Bioestatística
[2]Programa de Pós-Graduação em Biociências e Fisiopatologia

RESUMO

O objetivo deste trabalho foi estudar a concentração de creatinina sérica em pacientes com sepse. Os dados são referentes a 155 pacientes da Unidade de Terapia Intensiva (UTI) da Universidade Estadual de Maringá, Brasil. Estes pacientes foram divididos em dois grupos, com e sem insuficiência renal, e a creatinina sérica e o balanço hídrido foram registrados ao longo dos sete primeiros dias de internamento. Um modelo misto estimou a creatinina sérica em função do tempo, grupo e balanço hídrico acumulado e uma regressão de Cox foi ajustada considerando o tempo de internamento na UTI em função das covariáveis sexo e idade dos pacientes. Verificou-se que a creatinina estimada pelo modelo misto é maior para os pacientes com insuficiência renal, porém, esta pode ser diluída quando o volume de líquido nos pacientes aumenta. Também constatou, a partir da regressão de Cox, que a idade e o sexo são fatores de risco relacionados a morte de pacientes com sepse, indicando que mulheres e pacientes acima de 60 anos tem maior risco de morte. Além disso, uma análise conjunta dos dados longitudinais e de sobrevivência indicou que a concentração de creatinina em um tempo particular está fortemente associada ao risco de morte em pacientes com sepse.

Palavras e frases chave: Creatinina, fatores de risco, insuficiência renal, modelos mistos, regressão de Cox.

1. INTRODUÇÃO

A sepse é um conjunto de manifestações graves em todo o organismo, mais conhecida como infecção generalizada. Sepse grave e choque séptico são grandes problemas de saúde, afetando a cada ano, milhões de pessoas ao redor de todo o mundo, matando um a cada quatro paciente e sua incidência vem aumentando (Dellinger et al., 2008; Dombrovskiy et al., 2007). Pacientes com estes problemas apresentam grandes perdas hídricas para o espaço intersticial. Se não for revertida, a isquemia tecidual promove a disfunção de múltiplos órgãos, podendo levar a óbito (Vincent and Gerlach, 2004).

A disfunção renal de pacientes com sepse tem sido identificada por meio da creatinina sérica. No entanto, esta substância pode permanecer na faixa da normalidade mesmo quando cerca da metade das funções renais tenham sido perdidas (Soni et al., 2009). Além disso, uma possível diluição da creatinina sérica pode ocorrer devido ao aumento da carga hídrica acumulada no organismo e mascarar a disfunção renal (Macedo et al., 2010).

Durante muitos anos, foi comum a análise de dados longitudinais e de sobrevivência separadamente, sem considerar as possíveis associações entre elas. A metodologia mais utilizada para acomodar as correlações dos dados longitudinais é a adição de efeitos aleatórios nos modelos de regressão (Laird and Ware, 1982). Em estudos clínicos, em que a resposta é o tempo até a ocorrência de um evento de interesse, o modelo de regressão de Cox é o mais utilizado (Klein and Moeschberger, 2005). No entanto, a variável resposta longitudinal em um tempo específico, pode estar fortemente associada com o risco de acontecer um evento neste mesmo tempo (Rizopoulos, 2012).

A análise conjunta de dados longitudinais e de sobrevivência busca por quaisquer associações que possam existir entre as medidas em um mesmo indivíduo e o tempo até observar-se o evento de interesse. Dessa forma, o modelo conjunto permite ligar os modelos longitudinal e sobrevivência construídos separadamente.

O objetivo deste trabalho foi estudar a concentração de creatinina sérica em pacientes com sepse em função de grupos com e sem insuficiência renal, do tempo, mais especificamente durante os primeiros sete dias de internamento, e do balanço hídrico acumulado; relacionar o risco de morte de pacientes com sepse às covariáveis idade e sexo; e, além disso, medir a associação entre a concentração de creatinina e o risco de morte.

2. COLETA DOS DADOS

Os dados foram coletados na Unidade de Terapia Intensiva (UTI) do Hospital da Universidade Estadual de Maringá, Brasil, no período entre janeiro de 2011 a dezembro de 2012. O total de 155 indivíduos foi dividido em dois grupos, com e sem insuficiência renal (87 e 68 indivíduos). A creatinina sérica e o balanço hídrico dos pacientes foram registrados a cada 24 horas ao longos dos sete primeiros dias de internamento. O balanço hídrico acumulado ao longo dos dias foi categorizado em três grupos caracterizados por ter volume abaixo de cinco litros, entre cinco e dez, e acima de dez. A idade foi categorizada em dois grupos, abaixo e acima de 60 anos.

As análises deste trabalho foram realizadas com o *software* estatístico `R` (R Core Team, 2015) com o uso dos pacotes `nlme` (Pinheiro et al., 2015), `survival`, (Therneau, 2013) e `JM` (Rizopoulos, 2010).

3. MODELO MISTO

A ideia principal deste modelo foi estudar a tendência da variável resposta, concentração de creatinina, ao longo dos primeiros dias de internamento de pacientes com sepse. Para isto, o tempo foi considerado uma variável preditora contínua com efeito linear sobre a creatinina. Assim, o modelo tornou-se mais flexível, permitindo o ajuste de uma reta de regressão, com intercepto e inclinação diferentes, devido à inclusão de efeitos aleatórios nestes parâmetros, para cada paciente.

O modelo linear misto proposto para o ajuste dos dados de creatinina, Cr_{ijk}, do paciente i com $i = 1, \ldots, 155$, pertencente ao j-ésimo grupo com $j = 1, 2$, no k-ésimo balanço hídrico acumulado com $k = 1, \ldots, 3$, ao longo do tempo T, pode ser escrito como

$$Cr_{ijk} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})T + \gamma_j + \tau_k + (\gamma T)_j + (\tau T)_k + (\gamma\tau)_{jk} + \epsilon_{ijk} \tag{1}$$

em que β_0 é uma média global, β_1 o efeito fixo do tempo, γ o efeito dos grupos, τ o efeito do balanço hídrico acumulado, b_{0i} e b_{1i} são os efeitos aleatórios do paciente i associado com o intercepto e inclinação com $\boldsymbol{b}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$ em que $\boldsymbol{\Psi}$ é a matriz de covariância, os quais são considerados independentes entre os pacientes, $\epsilon_{ijkt} \sim N(0, \sigma^2)$ representa o resíduo e são independentes dos efeitos aleatórios, e os demais são as interações entre os efeitos fixos. O grupo sem insuficiência renal e o balanço hídrico acumulado menor que cinco litros foram considerados as *baselines* no modelo.

4. MODELO DE REGRESSÃO DE COX

Ao estimar os parâmetros do modelo de regressão de Cox, considerou-se a seguinte formulação:

$$\lambda(t) = \lambda_0(t) \exp\left\{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2\right\}, \tag{2}$$

onde $\lambda(t)$ é a função de risco estimada, sendo t o período de internamento na UTI, $\lambda(t)_0$ a função base, β_1 o coeficiente ajustado da variável idade (x_1), β_2 o coeficiente ajustado da variável sexo (x_2) e β_3 o coeficiente estimado da interação entre a variável idade e sexo. Considerou-se o grupo com idade menor que 60 anos e o sexo masculino como *baselines*.

5. RESULTADOS E DISCUSSÃO

A Tabela 1 mostra as estimativas para os parâmetros dos efeitos fixos do modelo (1) com seus intervalos de confiança, erros padrão (SE) e p-valores. Na análise de resíduos do modelo verificou de forma satisfatória as suposições de não correlação e de normalidade.

Parâmetros	Estimativas	LI	LS	SE
β_0	0.8403	0.7366	0.9441	0.0529
β_1	-0.0408	-0.0638	-0.0177	0.0117
γ_2	0.7908	0.6337	0.9479	0.0795
τ_2	-0.1431	-0.2273	-0.0589	0.0429
τ_3	-0.3589	-0.5106	-0.2071	0.0773
$(\gamma T)_2$	-0.0757	-0.1105	-0.0410	0.0177
$(\tau T)_2$	0.0358	0.0177	0.0540	0.0092
$(\tau T)_3$	0.0779	0.0499	0.1058	0.0142
$(\gamma\tau)_{22}$	0.1798	0.1041	0.2555	0.0386
$(\gamma\tau)_{23}$	0.3155	0.1756	0.4557	0.0713

Tabela 1: Estimativas, limite inferior e superior (LI and LS) e erro padrão (SE) para o modelo misto. Todas as estimativas foram significativas com p-valor $< 0{,}001$.

O estudo mostrou que a concentração de creatinina pode estar sendo diminuída conforme aumenta-se o volume de líquido administrado ao paciente. Quando este volume ultrapassa os 10 litros, a diminuição estimada é de 0.35 mg/dL. A interação entre os grupos e o balanço hídrico acumulado mostrou que, para os indivíduos com insuficiência renal, houve um aumento nos valores da creatinina, especialmente para aqueles pacientes com maior volume de líquido.

A Tabela 2 apresenta as estimativas e a exponencial dos parâmetros para o modelo (2), intervalos de confiança, erros padrão (SE) e p-valores. Nesta análise, os valores dos coeficientes de Pearson foram todos próximos de zero e, tanto o teste global quanto os testes para cada covariável, não rejeitaram a hipótese nula de riscos proporcionais.

Parâmetros	Estimativas	exp(est)	LI	LS	SE	p-valor
β_1	1.8926	6.6368	2.46124	17.896	0.5061	$< 0,0001$
β_2	1.7355	5.6717	1.78589	18.012	0.5896	0.0032
β_3	-1.3047	0.2712	0.07338	1.0030	0.6670	0.0505

Tabela 2: Estimativas, exponencial das estimativas (exp(est)), limite inferior e superior do exp(est) (LI e LS), erro padrão (SE) e p-valores para o modelo de regressão de Cox.

Observa-se, na Tabela 2, que tanto os fatores idade e sexo quanto sua interação foram significativos. Isto indica que estes fatores estão relacionados com o risco de morte de pacientes com sepse. O modelo de regressão de Cox apontou que pacientes com mais de 60 anos de idade possuem risco de morte aumentado em mais de seis vezes. Observou-se que, dos 88 indivíduos com menos de 60 anos, apenas 13 morreram (15 %), enquanto que, dos 67 com mais de 60 anos, 41 morreram (61 %). Este modelo também indicou que mulheres com sepse tem risco de morte cinco vezes maior que os homens. Dos 103 homens da amostra, 23 morreram (22 %), enquanto que, das 52 mulheres, 31 morreram (60 %). A interação dada pelo modelo de Cox, entre sexo e idade, mostrou que mulheres mais velhas tem um risco de morte menor que os homens mais velhos.

A análise conjunta dos dados longitudinais e de sobrevivência mostrou que a concentração de creatinina em um tempo específico está fortemente associada ao risco de morte em pacientes com sepse. Cada unidade aumentada na concentração de creatinina está associada com 1.37 vezes (95 % IC: 1.12; 1.68) ao aumento no risco de morte do paciente.

6. CONSIDERAÇÕES FINAIS

O modelo misto evidenciou que a creatinina sérica é maior para o grupo com insuficiência renal e que o balanço hídrico acumulado pode diluir-la. A regressão de Cox indicou que há um

risco de morte maior para as mulheres e pacientes com idade acima de 60 anos. A análise conjunta dos dados longitudinais e de sobrevivência indicou que, para um tempo específico, o risco de morte de pacientes com sepse está fortemente associada a concentração de creatinina sérica.

Tradicionalmente, a disfunção renal é indicada pela creatinina sérica, que pode permanecer na faixa de normalidade, mesmo os rins tendo perdido metade de suas funções. Além disso, uma possível diluição da creatinina devido ao aumento da carga hídrica acumulada no organismo pode mascara uma disfunção renal, resultado em um diagnóstico tardio e perigoso.

Referencias

Dellinger, R. P., Levy, M. M., Carlet, J. M., Bion, J., Parker, M. M., Jaeschke, R., Reinhart, K., Angus, D. C., Brun-Buisson, C., Beale, R., et al. (2008). Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2008. *Intensive care medicine*, 34(1):17–60.

Dombrovskiy, V. Y., Martin, A. A., Sunderram, J., and Paz, H. L. (2007). Rapid increase in hospitalization and mortality rates for severe sepsis in the united states: a trend analysis from 1993 to 2003. *Critical care medicine*, 35(5):1244–1250.

Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Macedo, E., Bouchard, J., Soroko, S. H., Chertow, G. M., Himmelfarb, J., and Ikizler, T. A. (2010). Research fluid accumulation, recognition and staging of acute kidney injury in critically-ill patients.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-117.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizopoulos, D. (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, 35(9):1–33.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

Soni, S. S., Ronco, C., Katz, N., and Cruz, D. N. (2009). Early diagnosis of acute kidney injury: the promise of novel biomarkers. *Blood purification*, 28(3):165–174.

Therneau, T. (2013). A package for survival analysis in s. r package version 2.37-4. *http://CRAN.R-project.org/package=survival*.

Vincent, J.-L. and Gerlach, H. (2004). Fluid resuscitation in severe sepsis and septic shock: an evidence-based review. *Critical care medicine*, 32(11):S451–S454.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Modelação conjunta de dados longitudinais e dados de sobrevivência na presença de riscos competitivos: Avaliação de um programa de diálise peritoneal

Laetitia Teixeira[1], Ipek Guler[2], Carmen Cadarso-Suarez[3], Inês Sousa [4], Anabela Rodrigues[5] e Denisa Mendonça[6]

[1]Instituto de Ciências Biomédicas Abel Salazar & CINTESIS, Universidade do Porto
[2]Center for Research in Molecular Medicine and Chronic Diseases, University of Santiago de Compostela, A Coruna, Spain
[3]Center for Research in Molecular Medicine and Chronic Diseases, University of Santiago de Compostela, A Coruna, Spain
[4]Departamento de Matemática e Aplicações, Centro de Matemática, Universidade do Minho, Braga, Portugal
[5]Departamento de Nefrologia, Centro Hospitalar do Porto - Hospital Geral de Santo António & Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Porto, Portugal
[6]Instituto de Ciências Biomédicas Abel Salazar & Instituto de Saúde Pública, Universidade do Porto, Porto, Portugal

RESUMO

A doença renal crónica é um problema global de saúde pública e tem vindo a verificar uma prevalência crescente em todo o mundo [7]. Esta patologia está associada ao aumento de risco de doenças cardiovasculares, anemia e doença mineral óssea, contribuindo para a diminuição da qualidade de vida do doente [10,11,14,15]. Adicionalmente, estão associados a esta patologia custos elevados inerentes aos cuidados de saúde necessários (nomeadamente tratamentos dialíticos) bem como perda de produtividade dos doentes [10, 14, 15].

Apesar de a hemodiálise continuar a ser o tratamento de substituição da função renal mais utilizado em Portugal, a utilização da diálise peritoneal tem vindo a verificar um aumento nos últimos anos. A preservação da função renal residual e dos acessos vasculares, o custo reduzido associado à terapia e a conveniência da terapia ser realizada em casa são algumas das vantagens associadas a esta técnica dialítica [2]. Além disso, relativamene ao bem-estar do doente, a percepção da qualidade de vida dos doentes em diálise peritoneal é usualmente melhor [9]. Assim, motivado pelo aumento de doentes em diálise peritoneal bem como pelas vantagens associadas a esta terapia, torna-se necessário a avaliação deste programa de tratamento, contribuindo para uma melhor gestão e qualidade do mesmo.

Num programa de diálise peritoneal, a monitorização de um doente é imprescindível e é realizada ao longo do tratamento (tempo de seguimento) através de visitas regulares de controlo onde são avaliados e registados indicadores associados à função renal bem como outros indicadores de saúde. De entre os indicadores registados, destacam-se i) sobrevivência do doente, da técnica ou combinada (dados de sobrevivência) e ii)

avaliação da evolução de parâmetros clínicos como, por exemplo, albumina, cálcio, fósforo e creatinina (dados longitudinais).

O primeiro indicador é obtido com recurso à análise de sobrevivência, considerando um evento de interesse específico de um conjunto de possíveis eventos (morte, transferência para hemodiálise e transplante renal). Para a análise de sobrevivência do doente, o evento de interesse é a morte do doente. A sobrevivência da técnica é avaliada com recurso ao evento "transferência para hemodiálise". Por fim a sobrevivência combinada analisa a permanência do doente no programa de diálise peritoneal até à saída do doente do programa por morte ou por transferência para hemodiálise. Dado que as razões de saída do doente do programa de tratamento são mutuamente exclusivas, para cada doente apenas é possível observar o primeiro evento a ocorrer. Assim, quando o interesse é avaliar um dos indicadores de análise de sobrevivência é necessário ter em conta a presença de eventos de riscos competitivos, isto é, eventos cuja ocorrência impedem a observação do evento de interesse.

O segundo indicador considerado, a avaliação da evolução de parâmetros clínicos, fornece informação relevante para avaliação da progressão da doença e/ou do tratamento em estudo que não pode ser obtida através da informação recolhida apenas na *baseline* [1]. Dado que a solução utilizada na diálise peritoneal apresentada concentração relativamente elevada de cálcio, os doentes em tratamento apresentam um maior risco de hipercalcemia [12] que, a longo prazo, poderá resultar num impacto negativo na função renal residual. Assim, o cálcio é um dos parâmetros clínicos monitorizados ao longo do tratamento em diálise peritoneal, sendo que o número e tempo decorrido entre medições diferem para cada doente.

Dada a relevância destes dois indicadores (dados longitudinais e dados de sobrevivência na presença de riscos competitivos) na avaliação de um programa de diálise peritoneal, torna-se essencial a avaliação de ambos. Para isso, é possível recorrer a modelos clássicos propostos para avaliar estes dois indicadores separadamente. No entanto, quando o objetivo é avaliar conjuntamente dados longitudinais e dados de sobrevivência na presença de riscos competitivos, estes métodos apresentam-se ineficazes, produzindo resultados enviesados, sendo por isso necessário recorrer a metodologias de modelação conjunta [13].

A utilização de uma abordagem de modelação conjunta de dados longitudinais e de dados de sobrevivência na presença de riscos competitivos é uma abordagem recomendada quando os dados longitudinais estão associadas aos dados de sobrevivência (isto é, na presença de censura informativa), quando os dados longitudinais são medidos com erro ou ainda na presença de valores omissos. Estes aspectos, que são características de dados reais clínicos, resultam em inferências enviesadas se métodos tradicionais são aplicados, sendo por isso necessário recorrer a abordagens adequadas como é exemplo a modelação conjunta [3, 4, 6, 8, 13].

A análise conjunta de dados longitudinais e de dados de sobrevivência na presença de riscos competitivos tem vindo a verificar uma atenção especial, resultando no aumento de abordagens propostas para este tipo de dados [5]. No entanto, esta abordagem ainda não se encontra amplamente disseminada em contexto clínico.

O presente trabalho tem como principal objetivo avaliar a associação entre o cálcio (dados longitudinais) e o tempo até à observação de um dos três possíveis eventos, morte, transferência para hemodiálise ou transplante renal (dados de sobrevivência), recorrendo a uma abordagem de modelação conjunta de dados longitudinais e de dados de sobrevivência na presença de riscos competitivos.

A amostra deste estudo é composta por doentes que integraram o programa de diálise peritoneal da Unidade de Diálise Peritoneal do Departamento de Nefrologia, Centro Hospitalar do Porto - Hospital Geral de Santo António (Porto, Portugal). O parâmetro

cálcio foi considerado como indicador longitudinal e a sobrevivência do doente e a sobrevivência da técnica como indicadores de sobrevivência. As covariáveis sexo e idade foram consideradas em ambos os processos longitudinal e de sobrevivência.

O processo longitudinal pode descrito pela seguinte expressão:

$$y(t) = \beta_0 + \beta_1 sexo + \beta_2 idade + \beta_3 tempo + b_0 + b_1 tempo + \varepsilon(t) = m(t) + \varepsilon(t)$$

onde β_0, β_1, β_2 e β_3 são coeficientes de efeitos fixos e b_0 e b_1 representam efeitos aleatórios ao nível do indivíduo, e $m(t)$ representa o verdadeiro e não observado valor do processo longitudinal.

O processo de sobrevivência, e dada a existência de três eventos de riscos competitivos, pode ser descrito pela seguinte expressão:

$$\begin{cases} \lambda_1(t) = \lambda_{01}(t)\exp\{\beta_{11}^* sexo + \beta_{21}^* idade + \alpha_1 m(t)\} \\ \lambda_2(t) = \lambda_{02}(t)\exp\{\beta_{12}^* sexo + \beta_{22}^* idade + \alpha_2 m(t)\} \\ \lambda_3(t) = \lambda_{03}(t)\exp\{\beta_{13}^* sexo + \beta_{23}^* idade + \alpha_3 m(t)\} \end{cases}$$

onde $\lambda_1(t)$, $\lambda_2(t)$ e $\lambda_3(t)$ representam os riscos para os eventos morte, transferência para hemodiálise e transplante renal, respetivamente. O parâmetro $\alpha_k(t), k = 1, 2, 3$ representa o coeficiente de associação entre o processo longitudinal e o processo de sobrevivência.

Considerando os resultados obtidos para o processo longitudinal, podemos concluir que as covariáveis sexo e idade apresentam um efeito significativo nos níveis de cálcio, apesar de este parâmetro não variar significativamente ao longo do tempo. Através da análise do submodelo de sobrevivência, apenas a idade apresenta-se como um fator de risco significativo para o transplante renal (idade mais avançada apresenta menor risco de transplante renal). Analisando os coeficientes de associação, verificamos uma associação entre o cálcio e o risco de transferência para hemodiálise, indicando que doente com maior aumento nos níveis de cálcio apresentam menor risco de transferência para hemodiálise.

A análise conjunta de dados longitudinais e dados de sobrevivência na presença de riscos competitivos apresenta sérias vantagens relativamente aos modelos de regressão standard que avaliam cada indicador individualmente. No contexto da diálise peritoneal, esta metodologia fornece informação adicional e fiável, permitindo obter uma avaliação mais clara e completa do programa de tratamento.

Palavras e frases chave: Modelação conjunta, submodelo longitudinal, submodelo de sobrevivência, riscos competitivos, diálise peritoneal.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado conjuntamente pelo Ministério da Economia e Competitividade (Espanha) e pelos Fundo Europeu de Desenvolvimento Regional (FEDER).

REFERÊNCIAS

Referencias

[1] Eleni-Rosalina Andrinopoulou, Dimitris Rizopoulos, Ruyun Jin, Ad JJC Bogers, Emmanuel Lesaffre, and Johanna JM Takkenberg. An introduction to mixed models and joint modeling: analysis of valve function over time. *The Annals of thoracic surgery*, 93(6):1765–1772, 2012.

[2] Pranav Dalal, Harbaksh Sangha, and Kunal Chaudhary. In peritoneal dialysis, is there sufficient evidence to make "pd first" therapy? *International journal of nephrology*, 2011, 2011.

[3] Peter J Diggle, Inês Sousa, and Amanda G Chetwynd. Joint modelling of repeated measurements and time-to-event outcomes: the fourth armitage lecture. *Statistics in medicine*, 27(16):2981–2998, 2008.

[4] Robert M Elashoff, Gang Li, and Ning Li. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. *Biometrics*, 64(3):762–771, 2008.

[5] A Lawrence Gould, Mark Ernest Boye, Michael J Crowther, Joseph G Ibrahim, George Quartey, Sandrine Micallef, and Frederic Y Bois. Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group. *Statistics in medicine*, 34(14):2181, 2015.

[6] Xu Guo and Bradley P Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16–24, 2004.

[7] Jonathan Himmelfarb and Mohamed H Sayegh. *Chronic Kidney Disease, Dialysis, and Transplantation: A Companion to Brenner and Rector's The Kidney*. Elsevier Health Sciences, 2010.

[8] Joseph W Hogan and Nan M Laird. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in medicine*, 16(3):259–272, 1997.

[9] Erika Juergensen, Diane Wuerth, Susan H Finkelstein, Peter H Juergensen, Ambek Bekui, and Fredric O Finkelstein. Hemodialysis and peritoneal dialysis: patients' assessment of their satisfaction with therapy and the impact of the therapy on their lives. *Clinical Journal of the American Society of Nephrology*, 1(6):1191–1196, 2006.

[10] Stefano Maffei, Silvana Savoldi, Giorgio Triolo, et al. When should commence dialysis: Focusing on the predialysis condition. *Nephro-urology monthly*, 5(2):723–7, 2013.

[11] Rachel L Perlman, Fredric O Finkelstein, Lei Liu, Erik Roys, Margaret Kiser, George Eisele, Sally Burrows-Hudson, Joseph M Messana, Nathan Levin, Sanjay Rajagopalan, et al. Quality of life in chronic kidney disease (ckd): a cross-sectional analysis in the renal research institute-ckd study. *American journal of kidney diseases*, 45(4):658–666, 2005.

[12] B Piraino, J Bernardini, J Holley, JR Johnston, JA Perlmutter, and L Martis. Calcium mass transfer in peritoneal dialysis patients using 2.5 meq/l calcium dialysate. *Clinical nephrology*, 37(1):48–51, 1992.

[13] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[14] Daniel E Weiner. Causes and consequences of chronic kidney disease: implications for managed health care. *Journal Of Managed Care Pharmacy*, 13(3 Supp A):1–9, 2007.

[15] Daniel E Weiner. Public health consequences of chronic kidney disease. *Clinical Pharmacology & Therapeutics*, 86(5):566–569, 2009.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Joint Modelling of Longitudinal and Survival Data on Breast Cancer

Ana Borges[1], Inês Sousa[2] e Luís Castro[3]

[1]CIICESI, ESTGF — Porto Politécnico
[2]CMAT, DMA-ECUM, Universidade do Minho
[3] Hospital de Braga

ABSTRACT

In order to contribute to the understanding of the progression of breast cancer, within the Portuguese population, we propose a statistical model with more complex assumptions than the traditional analysis, to study breast cancer within the senology unit of Braga's Hospital, located in north of Portugal. We aim to infer which risk factors affect the survival of Braga's Hospital patients, diagnosed with breast tumour. Whilst analysing risk factors that affect a tumour markers used on the surveillance of disease progression – the Carcinoembryonic antigen (CEA). As survival and longitudinal processes may be associated, it is important to model these two processes together. Hence, a joint modelling of these two processes to infer on the association of these was conducted.

A data set of 540 patients, along with 50 variables, was collected from medical records of the Hospital. A joint model approach was used to analyse these data. Two different joint models were applied to the same data set, with different parameterizations which give different interpretations to model parameters. These were used by convenience as the ones implemented in R software. Results from the two models were compared.

Results from joint models, showed that the longitudinal CEA values were significantly associated with the survival probability of these patients. A comparison between parameter estimates obtained in this analysis and previous independent survival [1] and longitudinal analysis [2], lead us to conclude that independent analysis brings up bias parameter estimates. Hence, an assumption of association between the two processes in a joint model of breast cancer data is necessary.

Palabras e frases chave: Breast Cancer, Survival analysis, Longitudinal analysis, Joint modelling.

1. Methodology

In previous independent longitudinal analysis [2], we verified that, observing the subset of patients that died from breast cancer, and using as reference time the one since death until date of blood tests, the results of the time effect, as so of the explanatory variables, were different from those when considering all patients in our data base. Showing that there is a relation between survival and longitudinal measures of tumour marker CEA, as so a joint modelling of these two processes seems reasonable.

The present joint analysis adopts of random effect methodology developed by Wulfsohn and Tsiatis (1997)[6] with the extension of Henderson (2000)[3], making use of two packages of R software: *JM* [5], and *joineR*[4].
The joint model implemented in the *JM* package focuses on the survival process and how it is influenced by a longitudinal time-dependent covariate that is measured with error. The *joineR* package implements a joint model where the focus is on both processes with the purpose of inferring about the strength of the link between both processes. Since both objectives are of interest for the present analysis, we made use of both packages.
Both model implementations use a linear mixed effects model, to represent the longitudinal process given by:

$$\text{Longitudinal Process: } Y_{ij} = \mu_{ij} + U_i + Z_{ij} = X_1\beta + U_{0i} + U_{1i}t_{ij} + Z_{ij}, \tag{1}$$

where X_1 is the design matrix for the fixed covariates, with corresponding regression parameters β. The latent variable given by (U_{0i}, U_{1i}) is a realization of a $MVN(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \nu_1^2 & \nu_{12} \\ \nu_{12} & \nu_2^2 \end{pmatrix}$ is the variance/covariance matrix of the random effects. Z_{ij} are Normal i.i.d. realizations of $N(0, \tau^2)$, representing the measurement error (variability non specified).
However, the main difference between *joineR* and *JM* packages relies on the way the incorporation of the latent random variable in the survival process is made. *JM* incorporates the precise estimate of the longitudinal unobserved process, $m_i(t) = \hat{\mu}_{ij} + \hat{U}_i$, into the survival process as it follows:

$$\text{JM Survival Process: } h_i(t) = h_0(t)\exp\{X_{2i}\beta_2 + \alpha m_i(t)\}, \tag{2}$$

where X_{2i} represents the baseline covariates, and β_2 the vector of corresponding regression parameters. The parameter α represents the effect of the true longitudinal response on the survival process, in this study the longitudinal responses are the log value of tumour marker CEA.
On the other hand, *joineR* incorporates the longitudinal random effects into the survival model through the random effects, as it follows:

$$\text{joineR Survival Process: } h_i(t) = h_0(t)\exp\{X_{2i}\beta_2 + \gamma_0 U_{0i} + \gamma_1 U_{1i}t\}, \tag{3}$$

where γ_0 and γ_1 represent the effect of the longitudinal random intercept and slope on the survival process, that is, the association parameters.
Actually both approaches differ only on the survival sub-model, since the longitudinal sub-model used in both software is the same. In fact, the difference between the two survival sub-models may lead to different estimate values of β_2 and even to differences on the effects of the covariates in the individuals'survival (i.e., covariates may have a statistical significant effect on subject's survival within the joint model fitted with *joineR* but no significant effect within the model fitted with *JM*). This can be easily perceived by the main difference between equations (2) and (3) that is the incorporation of the effect of $m_i(t) = \hat{\mu}_{ij} + \hat{U}_i$, which will lead to an effect of the longitudinal covariates in the survival process, i.e., in the estimates of the survival covariates, within the joint model fitted with *JM* package. On the other hand, estimates of β_2 within the joint model fitted by *joineR* package will only incorporate the effect of the covariates of the survival process.
To synthesize, using the model implemented by *JM* package we will be able to determinate the factors that influence the change in CEA values, on the patient's and what effect that change has on their survival. *joineR* package is appropriate to use in our analysis since we are interested in determining the effect of the initial response, γ_0, and change in the response over time, γ_1, on a patient's survival. As there seems to exist a link between the longitudinal process CEA and patient's survival.
For model diagnose we use graphical representation of the subject specific residuals versus the fitted responses for the validity of the mixed effects model. This plot is used to assess the assumption of constant variance of Z_{ij}. Also, a Q-Q plot of the subject-specific residuals is represented to verify the assumption of normality of Z_{ij}.To validate the survival process fitted we made use of the graphical representation of Cox-Snell residuals.

2. Main Results

Only patients with information of CEA values available were included. Which means that only 532 patients were eligible for the present analysis. The normality assumption of the response variable CEA was first tested preforming a Shapiro-Wilk normality test, with returned p-value $< 0{,}0001$ and therefore was rejected. It was used a log-transformation of tumor markers values. Table 1 presents the results from joint modelling the longitudinal process of CEA tumour marker values and the survival process, where the event is death from breast cancer and the reference time is time (in months) since tumour diagnose until death from tumour, using both packages. It is shown the estimates obtained for the covariates considered in the longitudinal and survival process, as well its standard errors (SE) and respective p-values.

Tabela 1: Estimated parameters values for Joint Model – CEA tumour marker

	JoineR Package			**JM Package**		
Covariates	**Estimates**	**SE**	**p-value**	**Estimates**	**SE**	**p-value**
LONGITUDINAL PROCESS						
Intercept	-0.297	0.118	0.012	-0.517	0.090	$< 0{,}0001$
Time (month)	0.004	0.001	$< 0{,}0001$	0.006	0.0004	$< 0{,}0001$
Age at diagnosis	0.011	0.001	$< 0{,}0001$	0.013	0.001	$< 0{,}001$
Tumour Stage (III or IV)	0.203	0.066	0.002	0.242	0.045	$< 0{,}001$
Degree (Gx or G1 or G2)	-0.012	0.062	0.846	-0.008	0.038	0.827
Estrogene Receptor (Positive)	0.151	0.064	0.019	0.190	0.059	0.001
Tumour Size						
(T4)	1.052	0.366	0.004	0.979	0.130	$< 0{,}001$
(Tis)	0.0120	0.224	0.930	0.139	0.171	0.418
SURVIVAL PROCESS						
Triple Negative (with)	3.163	0.576	$< 0{,}0001$	3.243	0.523	$< 0{,}0001$
Age at diagnosis						
(≥ 44)	-0.416	0.382	0.113	-0.503	0.347	0.147
Tumour stage						
(III or IV)	1.797	0.499	$< 0{,}0001$	1.259	0.367	0.0006
Tumour Degree						
(G3)	-0.464	0.456	0.317	-0.386	0.415	0.353
(Gx or G1)	-0.734	0.566	0.353	-0.986	0.625	0.115
Lymph Node						
(Nx or N0 or N1)	0.529	0.566	0.351	0.654	0.378	0.083
LATENT ASSOCIATION						
Association γ_0	0.959	0.175	$< 0{,}0001$			
parameter(s) γ_1	0.929	0.172	$< 0{,}0001$	α 1.048	0.115	$< 0{,}0001$
ν_1^2	0.2235			0.4757		
ν_2^2	$1{,}94 \times 10^{-7}$			0.0003		
τ^2	0.0117			0.1320		
Loglikelihood	-2254.97			-2308.11		

For both joint models adjusted, covariates such as triple negative (yes versus no), Tumour stage (0 or I or II versus III or IV) have a significant effect on patients'survival. However lymph node involvement (Nx or N0 or N1 versus N2 or N3), tumour grade (Gx or G1 versus G2 versus G3) and age at diagnosis (categorized as < 44 years old versus ≥ 44 years old) do not show a significant effect in patients survival within both joint models adjusted, contrariwise to the results from the independent survival analysis reported by Borges et al (2014). Covariates such as time, age at diagnosis, tumour stage, tumour degree, estrogene receptor expression and tumour size have significant effect on initial values of the linear mean progression of CEA log value, For the joint model obtained with *JM* package, the significance of the association parameter $\hat{\alpha}$ (p-value$< 0{,}0001$) in the survival process highlights the importance for a joint model for the present data. This significant association verifies the relationship between the longitudinal and survival processes. Likewise, within the joint model adjusted with *joineR* package, both association parameters, $\hat{\gamma_0}$ and $\hat{\gamma_1}$, are significant (p-value $< 0{,}0001$) in the survival process. Which confirms, in fact, the relevance for a joint model of the present data.

With regard to results obtain for *JM*'s joint model, through determination of hazard ratio (HR), it is apparent that individuals with higher CEA log value tend to have a worst survival (HR $= \exp(1{,}048) \approx 2{,}85$).

Within estimates obtained for the joint model adjusted with *joineR* package, through calculation of hazard ratios, $\hat{\gamma_0}$ indicates that individuals with initial values of CEA log values that are higher

than the population average tend to have worst survival (HR = $exp(0{,}959) \approx 2{,}61$). Likewise, $\hat{\gamma_1}$ indicates that individuals who have a greater increase in CEA log values linear mean progression tend to have worst survival (HR=$exp(0{,}929) \approx 2{,}53$). This was expected as high values of CEA corresponds to a deterioration of disease.

3. Discussion

Modelling together the survival and longitudinal processes, adopting the methodology of random effects [6], shows that the association between the longitudinal processes and survival is significant and it is essential its recognition. The importance of adopting the joint modelling methodology to deal with longitudinal and survival data arises from the fact that separate analysis based on the longitudinal model and the survival model may lead to biased estimates or even to inefficient results.
From the exposed it is clear that independent analysis brings up bias parameter estimates and an assumption of association between the two processes in a joint model is necessary.
At the moment there are only two packages that can easily fit these models, only available in R. However, these can only fit restricted models with random effects. For example, longitudinal models with stochastic correlation structure is not possible. This is mainly due to computation complexity.

Referencias

[1] Borges, A., Sousa, I., Castro, L. (2014). Breast Cancer Survival at Braga's Hospital – Portugal. In *Theoretical and Applied Issues in Statistics and Demography* C H Skiadas. ISBN 978-618-81257-7-3

[2] Borges, A., Sousa, I., Castro, L.(2015). Longitudinal Analysis of Tumor Marker CEA of Breast Cancer Patients from Braga's Hospital. *REVSTAT – Statistical Journal* 13(1), 63-–78.

[3] Henderson, R., Diggle, P.J., Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1(4), 465-–480.

[4] Philipson, P., Sousa, I., Diggle, P., Williamson, P., KolamunnageDona, R., Henderson, R. and R Core Team (2012). JoineR: Joint modelling of repeated measurements and time-to-event data. R package version: 1.0-3, http://CRAN.R-project.org/package=joineR.

[5] Rizopoulos, D. (2010). JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *Journal of Statistical Software* 35(9), 1–33.

[6] Wulfsohn, M.S., Tsiatis, A.A. (1997). A Joint model for survival and longitudinal data measured with error. *Biometrics* 53, 330-–339.

Comunicacións en póster

IIGalician-Portuguese Meeting of Biometry,
Santiago de Compostela, 30 June – 2 July 2016

APLICACIÓN DE REDES BAYESIANAS ÁS RELACIÓNS ENTRE ESPECIES VEXETAIS DESPOIS DE INCENDIO FORESTAL E A SÚA SENSIBILIDADE Ó TAMAÑO DAS UNIDADES DE MOSTRAXE.

García-Duro J.[1],Álvarez R[2],Basanta M[1]Casal M.[1],Reyes O.[1]

[1]Depto. de Bioloxía Celular e Ecoloxía. Facultade de Bioloxía. Campus Vida. Universidade de Santiago de Compostela 15783. Santiago de Compostela. e-mail: juan.garcia.duro@usc.es
[2]IES Legio VII. Plaza de Santo Martino nº6. 24003. León.

1. INTRODUCTION

Dous dos conceptos máis importantes en Ecoloxía, nicho e hábitat, xurdiron dos patróns de distribución das especies, que están determinados polas relacións existentes entre as plantas, o medio físico e os demáis compoñentes do ecosistema, que se expresan en distintas escalas espaciais.

O concepto funcional de nicho ecolóxico refírese a posición ou o papel que ten unha especie dentro dunha comunidade dada, mentres que o concepto de hábitat fai relación á distribución dunha especie nun rango de ambientes e comunidades [1]. De acordo con Hutchinson, o nicho ecolóxico das especies é o conxunto de características ambientais onde unha especie é capaz de sobrevivir e reproducirse; é o hipervolume n-dimensional en n variables ambientais [2]. Polo tanto, existe un gran volume de relacións que hai que descubrir á hora de determinar a posición dunha especie nas comunidades. Ademáis danse relacións moi complexas entre especies na conformación das comunidades, polo que as técnicas de redes bayesianas poden resultar moi axeitadas para identificar esas relacións [3] [4].

Ademáis, existen problemas na identificación correcta dos nichos ecolóxicos das especies debido á alta variabilidade dos límites dos rangos de distribución das especies e das derivas metodolóxicas, o que conduce a infraestimación dos rangos potenciais das especies [2]. As publicacións de disciplinas como a Ecoloxía Cuantitativa [5] e Ecoloxía espacial [6] teñen abordado moitos dos problemas derivados da distribución das plantas nas distintas escalas espaciais, e puxeron de relevancia a influencia que teñen o número e o tamaño da unidades de motraxe no estudo das características estruturais da vexetación.

A distribución das plantas nas distintas escalas espaciais varía en resposta a factores ambientais. Un destes factores ós que se enfrontan a maior parte dos ecosistemas terrestres do mundo é o lume [7].

Para dar luz ós problemas metodolóxicos habituais no estudo da estrutura das comunidades vexetais derivados da escala da mostra plantéxase este traballo co obxectivo de a) determinar as redes ecolóxicas máis probables que se producen tras incendio nunha comunidade de *P. tridentatum* mediante o emprego de redes bayesianas, e b) determinar o efecto do tamaño da unidade de mostraxe na valoración das interaccións interespecíficas.

2. MATERIAL E MÉTODOS

Para estudar as redes ecolóxicas, reflexadas na estructura dunha comunidade, a distintas escalas espaciais, seleccionouse unha área queimada en Oímbra na que se instalaron 10 rumbos fixos de 5 m, paralelos e separados entre si 1 m. A medida estructural medida foi a cobertura tralo incendio. Nos rumbos mediuse a cobertura lineal de cada unha das especies leñosas en tramos de 0.5 m de lonxitude. Excluíronse das análises posteriores as especies con menos dun 0,5% de cobertura total. As mostraxes realizáronse semestralmente, do 2º ó 4º ano tras incendio.

Empregando redes bayesianas analizouse a relación entre as especies con unidades de mostraxe de tamaño crecente, a partir dos datos de cobertura medidos nos 100 tramos de 0.5 m, en tramos continuos de 1m de lonxitude, en 20 tramos de 2.5 m, e nos 10 rumbos de 5 m. A unha escala espacial maior, agrupáronse os datos de cobertura de distintos rumbos, incluíndo dentro de cada grupo os 2 máis próximos e os 5 máis próximos. Por último calculouse o valor medio para o conxunto dos rumbos.

A partir deses datos, construíuse a rede ecolóxica de especies vexetais empregando redes bayesianas, construídas mediante o algoritmo de búsqueda Tabu [8] [9]. Na construcción destas redes especificáronse simultáneamente a relación do tempo transcorrido tras incendio cara tódalas especies e do cadrado do tempo, para permitir unha regresión cuadrática, dado que a relación entre o tempo transcorrido tras incendio e a cobertura das especies non adoita ser lineal. Non se controlou de ningún xeito a correlación espacial entre os datos. Estas análises realizáronse co paquete bnlearn de R [9].

3. RESULTADOS

Atopáronse diferencias importantes na cobertura das distintas especies leñosas co paso do tempo (Figura 1A). Do 2º ó 4ª ano tras incendio producíuse un incremento xeneralizado da cobertura na maioría das especies leñosas. As especies que aportaron maior cobertura foron *Halimium lasianthum* (Lam.) Spach e *Pterospartum tridentatum* (L.) Willk., seguidos por *Cytisus multiflorus* (L'Her) Sweet e *Erica australis* L.. Por último, *Glandora prostrata* (Loisel.) D.C.Thomas e *Polygala microphylla* L. tiveron pouca cobertura durante o período de estudo, que incluso diminuíu despois do segundo ano tras incendio.

Coa aplicación de redes bayesianas ós datos de cobertura (Figura 1B-H) atopáronse relacións significativas entre especies en tódalas lonxitudes e tamaños de mostra empregadas. Detectáronse cambios na intensidade das relacións, no seu signo e incluso na xerarquía das especies. Nas redes construidas coas mostras de pequeno tamaño (datos de grao fino) tódalas relacións entre especies foron negativas mentres que nas redes extraídas das mostras de maior tamaño (datos de grao groso) predominaron as relacións positivas entre especies.

Nas redes de datos de grao fino atopouse unha relación negativa entre a maioría das especies leñosas, destacando as relacións negativas entre as especies *C. multiflorus*, *P. tridentatum* e *E. australis*. Tamén se encontrou unha relación negativa entre *H. lasianthum* e as especies *P. tridentatum* e *E. australis*.

Nas redes de datos de grao groso hai unha relación positiva entre a maioría das especies, sendo consistente a varias escalas a relación entre *E. australis* e *H. lassianthum*, *P. tridentatum* e *C. multiflorus*, *P. tridentatum* e *G. prostrata*. Non se encontrou ningunha relación entre H. lasianthum e *G. prostrata* a esta escala.

4. DISCUSIÓN

Atopáronse cambios importantes na intensidade e no signo das relacións, de modo que nas redes construidas a partir de datos finos existen un gran número de relacións negativas e nas redes de datos grosos destacan relacións positivas entre especies. Tamén se atoparon cambios na xerarquía das especies. Polo tanto, as relacións detectadas entre as distintas especies dunha comunidade dependen do tamaño da unidade de mostraxe empregado. Os resultados obtidos mediante o emprego de redes bayesianas son sensibles ó tamaño da unidade de mostraxe, revelando as interaccións co medio e entre especies que ocorren a distintas escalas.

Nas redes de datos de grao fino detectáronse principalmente as relacións de competencia entre individuos de distintas especies de plantas. Cando unha especie alzcanza unha elevada cobertura nun determinada mostra a probabilidade de atopar outras especies nese mesmo lugar diminúe, aínda que é posible unha certa superposición entre especies. Nalgunhas comunidades de mato de Galicia a vexetación leñosa alcanza valores medios de superposición superiores ó 140% [10]. De todas formas, as redes de escalas espaciais pequenas tamén poderían atopar relacións positivas entre especies, que reflexan mecanismos de facilitación. Así, en escenarios de forte sequía a presenza dunha cuberta vexetal favorece a emerxencia de plantas en comparación cos lugares onde o solo non ten a protección desta cuberta [11]. A presenza dunha cuberta vexetal tamén adoita proporcionar unha certa protección fronte a predación [12]

Pola contra, nas redes de datos de grao groso recóllese dunha maneira máis efectiva a heteroxeneidade espacial da vexetación e as relacións das especies co medio, aparecendo asociadas aquelas especies que se sitúan próximas espacialmente porque ocupan ambientes similares, aínda que exista forte competencia entre elas e tendan a excluírse mutuamente na competencia polos recursos a pequena escala. A unha escala maior que a estudada, estas interaccións levarán á formación de distintos tipos de comunidades [13] [14].

En conxunto, á hora de explorar as interaccións entre especies que se producen durante a rexeneración da vexetación tras incendio [4] as redes Bayesianas amosaron ser sensibles ó tamaño da unidade de mostraxe e permitiron a detección desas relacións existentes a distintas escalas espaciais.

AGRADECEMENTOS

Programa de axudas predoutoraisMaria Barbeito e ós proxectos de investigación GESFIRE (AGL2013-48189-C2-2-R), financiado polo Ministerio de Economía y Competitividad y polos fondos FEDER e ó proxecto PGIDIT06RFO20001PR, financiado pola Xunta de Galicia.

BIBLIOGRAFÍA

[1] Whittaker R.H., Levin S.A., Root R.B. (1973). *American Naturalist* 107 (955), 321-338.

[2] Stockwell D. (2007). *Niche modeling. Predictions from Statistical distributions*. Chapman & Hall/CRC, London.

[3] Pozsgai G., Baird J., Littlewood NA., Pakeman R.J., Young M. (2015). Interspecific networks in ground beetle (Coleoptera: Carabidae) assemblages. *Ecological Indicators*. . URL http://www.sciencedirect.com/science/article/pii/S1470160X15006639

[4] Newton A.C., Stewartb G.B., Diaza A., Golicherc D., Pullinb A.S. (2007). Bayesian Belief Networks as a tool for evidence-based conservation management. *Journal for Nature Conservation* 15, 144—160.

[5] Greig Smith, P. (1983). *Quantitative Plant Ecology*. 3ª ed. Blackwell. Oxford. UK.

[6] Dale M., Fortin M.J. (2014). *Spatial Analysis. A Guide For Ecologists*. 2nd Edition.

[7] Bond y Keeley. 2005. Fire as a global "herbivore": the ecology and evolution of flammable ecosystems. Trends in Ecology and Evolution. 20 (7): 387-394

[8] Radhakrishnan Nagarajan, Marco Scutari, Sophie Lebre. (2013) Bayesian Networks in R with Applications in Systems Biology. Springer, New York. ISBN 978-1461464457.

[9] Marco Scutari (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software* 35(3), 1-22. URL http://www.jstatsoft.org/v35/i03/.

[10] Basanta M., Díaz Vizcaíno E.A., Casal M. (1988). Structure of Shrubland communities in Galicia (NW Spain). In During H.J., Werger M.J.A., Willems J.H.: *Diversity and pattern in plant communities.* SPB Academic Publishing. pp. 25-36. The Hague, The Netherlands.

[11] Lloret F., Peñuelas J., Estiarte M. (2005). Effects of vegetation canopy and climate on seedling establishment in Mediterranean shrubland. *Journal of Vegetation Science* 16, 67-76.

[12] Torroba-Balmori P., Zaldívar P., Alday J.G., Fernández-Santos B., Martínez-Ruiz C. (2015) Recovering Quercus species on reclaimed coal wastes using native shrubs as restoration nurse plants *Ecological Engineering* 77, 146–153.

[13] Díaz-Vizcaíno E, Basanta M, y Morey M. (1988). Comportamiento ecológico de algunas especies arbustivas de la cuenca del Tambre (La Coruña) frente a diversos factores topográficos y edáficos. Bol. R. *Soc. Esp. Hist. Nat. (Biol.)* 84 (1-2), 53-72.

[14] Díaz-Vizcaíno E., Basanta M., Morey M. (1989). Preferencias ecológicas y óptimos ecológicos de diferentes comunidades de matorral y de sus especies en Galicia. *Studia Oecológica* VI, 41-67.

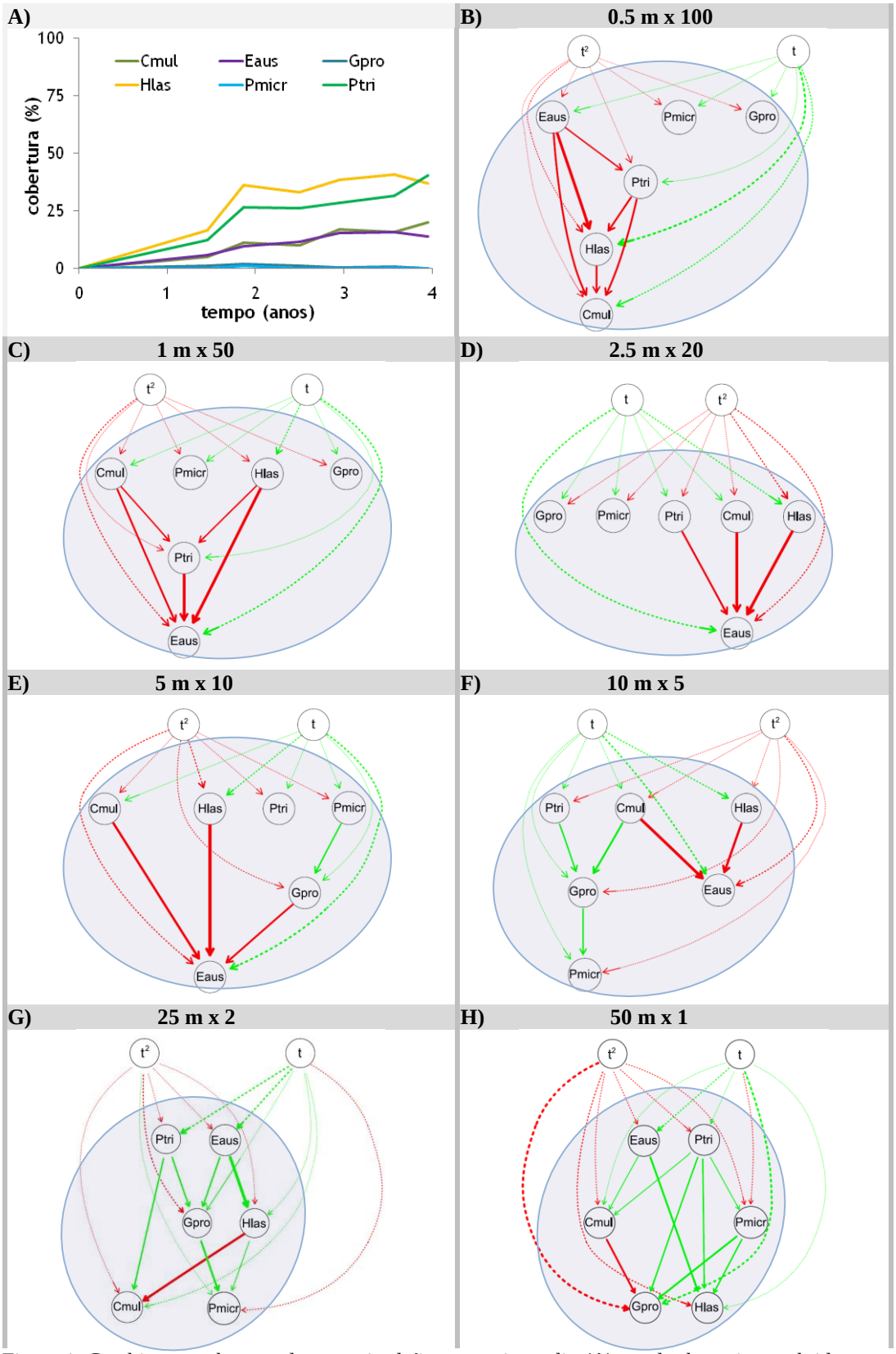


Figura 1: Cambios na cobertura das especies leñosas tras incendio (A) e redes bayesianas obtidas para cada escala espacial (B-H). Indícase o tamaño e o número de unidades de mostraxe empregadas. t: tempo(anos); t^2: tempo ó cadrado; Cmul: *C. multiflorus*; Eaus: *E.australis*; Gpro: *G. prostrata*; Hlas: *H. lasianthum*; Pmicr: *P. microphylla*; Ptri: *P. tridentatum*.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

USO DE GRANDES MASAS DE DATOS PARA ESTUDOS EN ECOLOXÍA

[1] Antón Blanco Vilariño. Biólogo estudante do máster interuniversitario de técnicas estadísticas e investigación operativa.

[2] Óscar Lado Baleato. Biólogo estudante do máster interuniversitario de técnicas estadísticas e investigación operativa.

RESUMO

Tal e como recolle o artigo "Big data and the future of ecology" de Hampton S. E. et al, (2013)[1] os ecólogos como colectivo producen unha grande cantidade de datos a través de diversos proxectos individuais que non chegan a compartirse na súa totalidade, para sacar suficiente información deses datos estes deben ser tratados como un producto de calidade e non como precursores dunha publicación, esta nova metodoloxía xa adoptada por outros colectivos dentro da bioloxía como o caso da xenómica e biomedicina, permitirá á ecoloxía enfrontar retos mundiais como o cambio climático, a crise alimentaria e a transmisión de enfermidades a nivel global.

Neste contexto propoñemos un exemplo de aplicación tomando os datos da EPA, axencia de proteción ambiental dos EUA , referentes á contaminación, cruzando os mesmos cos da BBS unha axencia para o estudo das aves nidificantes dos EUA. Ambas bases de datos recollen unha amplitude temporal similar; ademais de ser as estacións de motraxe de ambos estudos próximas no espazo.

Tomamos este estudo sobre unha rexión dos EUA onde ambos estudos que se levaron a cabo de forma indepedente presenten unha maior coincidencia de estacións.

Palabras e frases chave: Open data sources, Breeding birds population, Air pollution, Big data.

Referencias

[1] Stephanie E Hampton, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter (2013). Frontiers in Ecology and the environment; doi:10.1890/120103

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Using Ripley's K function to analyse spatial termite mounds patterns: a case study

Ana J. Righetto[1], Djair D.R. Frade[1], Luiz R. Nakamura[1],
Thiago G. Ramires[1], Ezequiel A.L. Bautista[2] and Paulo J. Ribeiro Junior[3]

[1]Departamento de Ciências Exatas, Universidade de São Paulo, Piracicaba, São Paulo, Brazil
[2]Facultad de Agronomía, Universidad de San Carlos de Guatemala, Guatemala, Guatemala
[3]Departamento de Estatística, Universidade Federal do Paraná, Curitiba, Paraná, Brazil

ABSTRACT

In this work we studied spatial patterns of termites in a teak plantation in the northern of Republic of Guatemala, splitting the data set according to termite mounds sizes (small, medium and large) and analysing their individual and cross patterns. For individual patterns we noticed that when only small termite mounds or when all termite mounds were considered an aggregation distribution throughout the studied region was detected, while large termite mounds presented this pattern only in a specific area and medium ones showed a regularity pattern in the whole region. For cross patterns no attraction or inhibition relationships were observed.

Key words: *Nasutitermes nigriceps*, Ripley's K function, Spatial point processes, Teak tree, *Tectona grandis* L.f.

1. INTRODUCTION

Teak (*Tectona grandis* L.f.) is a tree native to the rainforests of Southeast Asia and belongs to the Verbenaceae family. Its natural range is wide, extending between parallels 09° N and 25° N, comprising regions located from sea level up to a thousand meters, subject to annual rainfall between 500 and 5,000 mm and absolute temperatures between 2° C and 48° C. It is a large tree, reaching up to 25 and 35 m high and 1 m in diameter at breast height (Lima et al., 2009) with a straight trunk in its natural environment, with rough, thin and cracked bark and grey or greyish brown branches. Although its bark is not thick, it is a thermal insulation, resulting in a high resistance to fire.

One of the main problemas on teak plantations are the termites. Different studies regarding the termite mounds spatial distribution are presented in a wide literature. However, those studies cannot be extrapolated to all regions, since different regions have distinct properties and require different policies in order to control these insects in the most efficient and sustainable possible way. Hence, the main objective of this work is to provide subsidies in plantations of Tectona Grandis L.f. in Guatemala through spatial analysis, identifying whether there is any spatial pattern near the edges or in the centre of the plantation, using uni and bivariate Ripley's K (and L) functions.

2. MATERIAL AND METHODS

Study area

The data set analysed in this paper regards to a study conducted on 15 July 2013 at GM-99 farm, located in La Libertard, Departamento de Petén, Guatemala. The samples were randomly collected over an area of 46.69 ha, georeferencing each point in the branches of teak trees. The size of the termite mounds were classified according to their diameters: i) small (< 20 cm); ii) medium (20 – 30 cm); and iii) large (> 30 cm).

Methodology

The Ripley's K function (Ripley, 1977) has the main advantage to assist in the detection of spatial patterns on different scales of distances simultaneously. Moreover, it helps in the investigation of spatial independence level between any two groups, e.g. termite mounds of different sizes. However, in practice, the Rippley's K function does not present a simple interpretation since it produces a parabola. In order to simplify its visualisation and interpretation, Besag and Diggle (1997) proposed a slightly correction in it, that resulted in a new function called the L function, which produces a linear plot with a similar interpretation. Thus, in order to verify whether a process presents aggregation, the observations just need to be positive. Therefore, we used the L function in this work, which is defined as

$$\hat{L}(r) = \sqrt{\frac{|S|}{N^2\pi} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{I_r(d_{i,j} \leq r)}{w_{s_i,s_j}}} \tag{1}$$

where $I_r(\cdot)$ is an indicator function, N corresponds to the number of individuals in region S with area $|S|$, $d_{i,j}$ is the distance between the i^{th} and j^{th} individuals and w_{s_i,s_j} is a correction based on the circunference ratio within region S.
Later, in order to verify a possible relationship between termite mounds with different sizes, we used the bivariate estimator, which is derived from the Ripley's K function, to test the complete spatial randomness and independence (CSRI). Analogously to its univariate version, we shall transform the bivariate Ripley's K function in order to simplify its visualisation and interpretation. The resulting function, called the bivariate L function, is given by

$$\hat{L}_{ij}(r) = \sqrt{\frac{|S|}{N_1 N_2 \pi} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \frac{I_r(d_{i,j} \leq r)}{w_{s_i,s_j}}}$$

where N_1 and N_2 correspond to the number of individuals from the first and second group, respectively, in region S and the other terms are as defined in (1).
In the uni and bivariate cases, we use Monte Carlo envelopes are used in order to perform the hypotheses tests (Diggle, 2003). In the univariate case we test whether there is complete spatial randomness (CSR) and in the bivariate case whether the groups are independents, present attraction or inhibition. Basically, we reject those hypotheses if the observed values are within the confidence envelope.

3. RESULTS

We sampled 178 termite mounds, of which 67 were classified as small (diameter smaller than 20 cm), 31 as medium (diameter between 20–30 cm) and 80 as large (diameter greater than 30 cm), and they are displayed in Figure 1 on four different panels, respectively.

Univariate analysis

Figure 2 displays the returned L functions. Panel (a) provides the L function considering all termite mounds and Panels (b), (c) and (d) plot the L function for small, medium and large termite

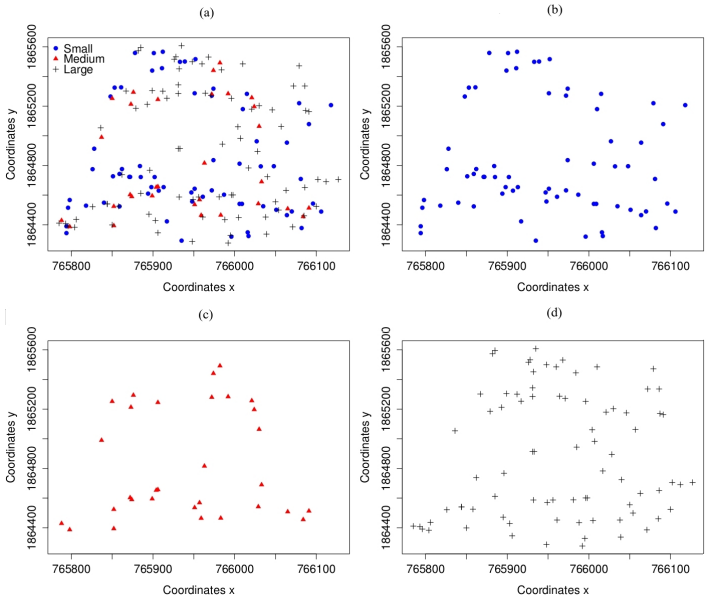


Figura 1: Termite mounds spatial distribution map: (a) all; (b) diameter smaller than 20 cm; (c) diameter between 20-30 cm; and (d) diameter greater than 30 cm

mounds, respectively. Dashed lines indicate the 99 % confidence intervals obtained through Monte Carlo simulations for CSR, while the continuous line indicates the fitted L function to the data. We can see that the spatial pattern from all the data set, does not satisfy the CSR hypothesis in distances until r=385 m, since the function stays outside the upper bound of the envelope, i.e., the termite mounds are not randomly distributed in the study area, but present an aggregation spatial pattern in this scale. For distances greater than r=420 m, the termite mounds present a regularity pattern. Moreover, Figure 2(b) shows that small termite mounds (diameter smaller than 20 cm)

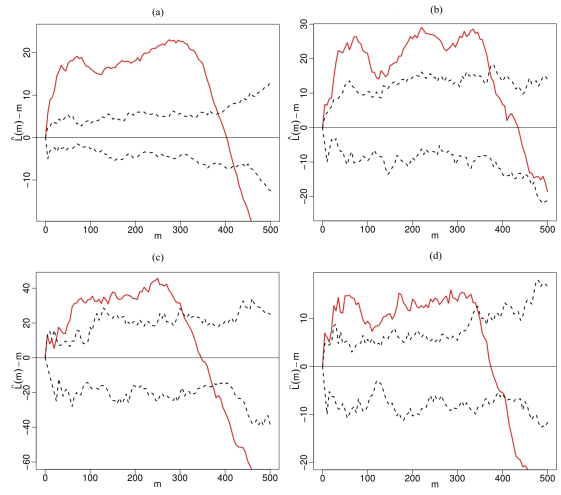


Figura 2: Spatial dependence analysis of the termite mounds: (a) all; (b) diameter smaller than 20 cm; (c) diameter between 20–30 cm; and (d) diameter greater than 30 cm

present an aggregation spatial pattern and after r=375 m the individuals'distribution pattern is random. Medium (diameter between 20–30 cm) and large (diameter greater than 30 cm) termite mounds, displayed in panels (c) and (d) respectively, have a similar behaviour than when all termite mounds are considered. There is an aggregation spatial pattern until r=325 m and regularity for distances greater than r=390 m in these sizes.

Hence, the termite mounds spatial distribution map on this study, may indicate that there are no spots throughout the region in study presenting better conditions for the establishment of termite mounds, or the absence of intra-specific competition of space and food resources (Begon and Harper, 1996). Finally, the low density of termite mounds throughout the region in study might be another factor to contribute with this pattern.

Bivariate analysis

Figure 3 displays the confidence envelopes for the bivariate L function in order to perform the CRSI analyses between termite mounds of different sizes: i) Panel (a), (b) and (c) display the analysis between small versus medium, small versus large and medium versus large termite mounds, respectively. The analyses between small versus medium and small versus large termite mounds presented an attraction relationship with scale up to 100 m and 90 m, respectively. For scales greater than these values, the CSRI hypothesis was not violated. Further, medium and large termite mounds presented an attraction relationship in scales up to 320 m and, for greater values than 400 m it can be seen an inhibition relationship. Grohmann et al. (2010) highlight that if we consider two termite mounds founded on the same period, in which one has better growing conditions, the fastest growing colony can suppress the growth of the other colony, or that the smaller termite mounds will be a branch of the larger one, i.e. polydomous colonies.

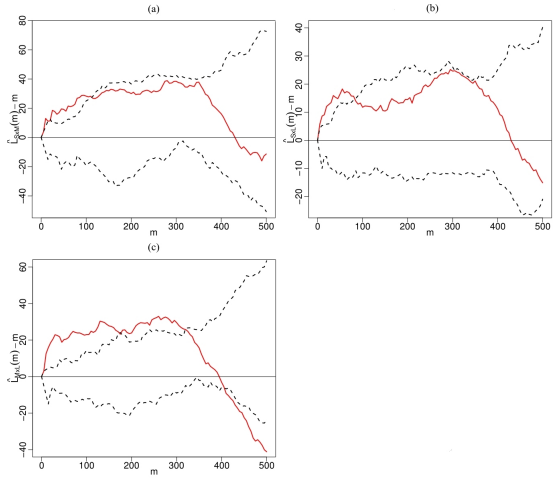


Figura 3: Analysis of the spatial interaction between the different sizes of termite mounts: (a) small versus medium, (b) small versus large, (c) medium versus large

REFERENCES

Referencias

[1] Begon, M.J.L. and Harper, C.R. (1996). *Ecology: Individuals, Populations and Communities*. Blackwell Sciencie.

[2] Besag, J. and Diggle, P.J. (1977). Simple monte carlo tests for spatial pattern. *Journal of the Royal Statistical Society – Series C (Applied Statistics)* 26, 327–333.

[3] Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns*. Academic Press.

[4] Grohmann, C., Oldeland, J., Stoyan, D. and Linsenmair, K. E. (2010). Multi-scale pattern analysis of a mound-building termite species. *Insectes Sociaux* 57, 367–494.

[5] Lima, I.L., Florsheim, S.M.B. and Longui, E.L. (2009). Influência do espaçamento em algumas propriedades físicas da madeira de Tectona grandis Linn, *Cerne* 15, 244–250.

[6] Ripley, B.D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society – Series B (Methodological)* 39, 172–212.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

A INFLUÊNCIA DO ÍNDICE DE OSCILAÇÃO DO ATLÂNTICO NORTE SOBRE AS ADMISSÕES HOSPITALARES POR MEIO DOENÇAS DO APARELHO CIRCULATÓRIO EM LISBOA

Giovani L. Silva[1], Ricardo Almendra[2], Paula Santana[2], João Vasconcelos[3], Fábio Gonçalves[4] e Tércio Ambrizzi[4]

[1]CEAUL e Departamento de Matemática-IST, Universidade de Lisboa, Portugal
[2]Centro de Estudos de Geografia e Ordenamento do Território, Universidade de Coimbra, Portugal
[3]Escola Superior de Turismo e Tecnologia do Mar, Instituto Politécnico de Leiria, Portugal
[4]Instituto de Astronomia, Geofísica e Ciências, Universidade de São Paulo, Brasil

RESUMO

O objetivo deste trabalho é analisar a relação entre a Oscilação do Atlântico Norte (NAO), variáveis meteorológicas, poluentes atmosféricos e internações por doenças do aparelho circulatório em Lisboa (Portugal) durante os meses de inverno (2003-2012). Nesse sentido, este trabalho visa analisar o impacto da NAO sobre a saúde através de sua influência sobre o *stress* térmico e poluição do ar em Lisboa.

Este estudo utiliza dados meteorológicos (sintetizados em um índice de conforto térmico), medidas de poluentes atmosféricos e o índice NAO (todos agrupados em ciclos de 10 dias para lidar com a variabilidade diária do índice NAO). A relação entre morbidade, índice de conforto térmico, índice NAO e os poluentes do ar foi explorado através de vários modelos lineares ajustados à sazonalidade através de uma função periódica.

O efeito indireto possível entre o índice NAO e internações hospitalares foi testado, assumindo que a variável independente NAO afeta a variável resposta internações hospitalares através das variáveis desconforto térmico e/ou níveis de poluição (testadas como mediadoras individuais). Este teste foi realizado por meio da análise de mediação causal e ajustado para variações sazonais.

Em muitas áreas científicas, o objetivo dos investigadores não só é estimar os efeitos causais de um tratamento, mas também a compreensão do processo em que o tratamento afeta causalmente o resultado. A análise de mediação causal é frequentemente usada para avaliar potenciais mecanismos causais (vide *e.g.* Imai *et al.*, 2010). O pacote R *mediation* implementa um conjunto abrangente de ferramentas estatísticas para a realização de tal análise (Tingley *et al.*, 2014).

Os resultados deste estudo destacam os impactos que os padrões de circulação atmosférica podem ter na saúde. Além disso, a compreensão da influência dos padrões de circulação atmosférica pode apoiar a melhoria dos planos de emergência existentes. A eficácia dos sistemas de alerta de poluição do ar relacionados com a saúde poderia ser melhorada, tendo em conta os resultados desta pesquisa e, assim, ajudar-se-ia a diminuir os impactos na saúde de condições ambientais nocivas.

Palavras e frases chave: Oscilação do Atlântico Norte; doenças do aparelho circulatório; poluição do ar; admissões hospitalares.

AGRADECIMENTOS

Este trabalho foi realizado no âmbito da cooperação internacional do INCLINE (INter-disciplinary CLimate Investigation cEnter) e parcialmente apoiado pelo Fundo Europeu de Desenvolvimento Regional, através do Programa de Competitividade Operacional (doutoramento SFRH/BD/92568/-2013 - Ricardo Almendra), pela FCT - Fundação para a Ciência e Tecnologia, Portugal, através do Projeto UID/MAT/00006/2013 - Giovani Silva, e pelo CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico - e FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo - Tércio Ambrizzi.

REFERÊNCIAS

[1] Imai K., Keele L., Yamamoto T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* 25(1), 51–71.

[2] Tingley D., Yamamoto T., Hirose K., Keele L., Imai K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software* 59(5), 1–38. Available from: http://www.jstatsoft.org/v59/i05/

II Encontro Galaico-Português de Biometría
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

ANÁLISE COMPARATIVA DE RESULTADOS DE ESPETROMETRIA GAMA PORTÁTIL EM AFLORAMENTOS ROCHOSOS

Flávio Dias[1], Marco Lima[2], Jorge Sanjurjo-Sánchez[3] e Carlos Alves[4]

[1] LandS/Lab2PT - Laboratório de Paisagens, Património e Território (FCT UID/AUR/04509/2013; FEDER COMPETE POCI-01-0145-FEDER-007528)
[2] LandS/Lab2PT - Laboratório de Paisagens, Património e Território (FCT UID/AUR/04509/2013; FEDER COMPETE POCI-01-0145-FEDER-007528)
[3] Instituto Universitario de Xeoloxía "Isidro Parga Pondal", Universidade da Coruña, A Coruña, Espanha
[4] LandS/Lab2PT - Laboratório de Paisagens, Património e Território (FCT UID/AUR/04509/2013; FEDER COMPETE POCI-01-0145-FEDER-007528) e Escola de Ciências, Universidade do Minho, Braga, Portugal

RESUMO

São comparadas estimativas de elementos obtidas por espetrometria gama portátil em afloramentos de rochas metamórficas e ígneas, assim como medições de campo e análises laboratoriais para uma das rochas ígneas consideradas, tendo em vista contribuir para a discussão da utilidade da espetrometria portátil técnica na avaliação dos materiais.

Palavras e frases chave: materiais geológicos, granitos, rochas metassedimentares, radiação gama externa.

1. INTRODUÇÃO

Os materiais geológicos utilizados como material de construção (como pedra natural ou agregados) podem contribuir para a dose gama externa a que os seres humanos estão expostos no ambiente construído (nomeadamente nos espaços interiores), devido, em geral, aos seus teores em potássio, urânio e tório, os quais podem ser extremamente variáveis em função do tipo de rocha e, adicionalmente, ser afetados por processos de alteração (posteriores à formação da rocha). A avaliação dessa contribuição é, habitualmente, baseada em índices (ver, e.g., [1]), calculados a partir das atividades específicas do ^{40}K (potássio) e de isótopos representativos das séries de decaimento do urânio (U) e do tório (Th), determinadas por dispendiosas e morosas análises laboratoriais. Adicionalmente, ainda que o procedimento de análises laboratoriais de amostras de rocha possa ser utilizado na avaliação dos materiais que vão ser aplicados, dificilmente será viável desenvolver o mesmo para o estudo generalizado dos materiais em habitações existentes. Existe, consequentemente, interesse em poder realizar uma avaliação de campo, mesmo que essencialmente comparativa, dos teores dos elementos radioativos nos materiais (o que permitiria, se bem sucedida, uma comparação da perigosidade dos materiais em termos de radiação gama externa). Dias et al. [2] compararam medições num tipo de granito e em rochas metassedimentares do Silúrico, tendo em vista o estudo de um perfil num local sem afloramentos. A análise grupal separou claramente as medições no granito das medições nas rochas metassedimentares, indicando ainda uma agregação dos dados do perfil compatível com a cartografia geológica destas rochas na escala 1:50000.

Neste trabalho o estudo comparativo dos resultados de campo de espetrometria gama é estendido a um outro tipo de granito da mesma área, sendo feita, também, a comparação de dados de campo com análises de laboratório para um dos granitos considerados.

2. MATERIAIS E MÉTODOS

Os espetros de radiação gama foram adquiridos com um espectrómetro portátil GF Instruments GRS-2000, transportado por um operador e equipado com uma sonda BGO ($Bi_4Ge_3O_{12}$) e um fotomultiplicador que regista fotões em 512 canais (medindo energias até 3 MeV –Megaelétron Voltios). Cada espetro corresponde a um período de aquisição igual a 180 segundos, segundo a calibração e otimização das medições feitas para o equipamento utilizado. A partir destes espetros são obtidas estimativas de teores de K, de equivalente de urânio (eU) e equivalente de tório (eTh). O teor em potássio (K, %) é estimado a partir do pico do ^{40}K. As séries de decaimento do urânio e do tório têm vários isótopos radioativos e, assumindo equilíbrio secular, são obtidas estimativas de eU (em ppm ou partes por millao) e eTh (em ppm ou partes por milhão) a partir dos picos do do ^{214}Bi (bismuto) e ^{208}Tl (tálio), respetivamente. As estimativas apresentadas são expressas em diferentes escalas uma vez que o potássio é um elemento maior na crusta e apresenta nos materiais geológicos teores muito superiores aos teores de urânio ou tório (que são elementos traço). Com base nestas estimativas calcularam-se valores de taxa de dose absorvida de radiação gama externa utilizando os fatores indicados em [3] e em [4].

Estas medições foram realizadas por contacto direto em superfícies de afloramentos cartografados [5] como rochas metassedimentares do Silúrico, assim como granitos do Sameiro e de Braga, ambos granitos biotíticos porfiroides mas, respetivamente, de grão grosseiro e de grão fino a médio. As medições foram realizadas em locais que apresentavam diferentes feições em termos de processos de alteração presumivelmente superficial (meteorização). Foram ainda colhidas amostras com diferentes graus de meteorização do granito de Braga para análises laboratoriais de espetrometria gama num detetor de germânio tipo coaxial marca CANBERRA modelo GR6022 (amostras seladas durante 25 dias para equilíbrio do Rn), cujos resultados, publicados em [6], mostraram variações em amostras com diferentes graus de meteorização. Para comparação com as medições de campo, foram calculadas estimativas de teores dos elementos (utilizando os fatores indicados em [3]) e valores de taxa de dose absorvida (pelo procedimento anteriormente indicado).

Considerando que os parâmetros são expressos em escalas diferentes, e para evitar o efeito de variações de valores mais elevadas, foram utilizados, nas análises multivariadas, os valores normalizados (média igual a zero e desvio-padrão igual a 1) das estimativas de K, eU e eTh. As análises multivariadas comparativas, análise de semelhança (ANOSIM) e MANOVA não paramétrica (NPMANOVA ou PERMANOVA), ambas com 99999 permutações, foram realizadas com o *PAST-PAleontological STatistics* [7]. Todas as outras análises e todos os diagramas apresentados foram preparados com o *Statistica* 11 (Statsoft).

4. RESULTADOS E DISCUSSÃO

Na Figura 1a-d, utilizando diagramas de pontos, são comparadas as estimativas (valores normalizados) obtidas a partir dos espetros de campo em afloramentos das diferentes litologias consideradas. Parece existir uma tendência para valores mais baixos nas medições realizadas nas rochas metassedimentares silúricas (em relação às rochas graníticas) mas os tipos de rochas graníticas consideradas mostram alguma sobreposição de valores. A análise de componentes principais (Figura 1e) mostra uma clara separação entre rochas metassedimentares e rochas graníticas na componente principal 1 (CP1) sendo que a CP2 parece sugerir alguma tendência para separação do granito do Braga em relação ao granito de Sameiro. O diagrama das variáveis (Figura 1f) sugere que esta CP2 será afetada pela relação entre o urânio e o tório. Os resultados da análise grupal (Figura 1g) também mostram a clara separação das rochas metassedimentares em relação às rochas graníticas e uma certa tendência para agrupar as medições do granito do Sameiro. Os resultados de ANOSIM para a comparação de medições no granito de Braga e no granito do Sameiro fornecem valores de p baixos, e.g. $p<0,10$ para a distância Chord e $p<0,050$ utilizando as distâncias euclidiana, Manhatan, Mahalanobis e Gower. A NPMANOVA fornece resultados semelhantes, com exceção da distância de Mahalanobis para a qual são obtidos valores de $p<0,10$.

Na Figura 2 são comparados os valores normalizados das medições de campo no granito de Braga e dos resultados das análises laboratoriais, parecendo existir alguma tendência para valores mais baixos nos resultados obtidos em laboratório. O teste ANOSIM forneceu alguns valores muito baixos de p (por exemplo $p<0,010$ para as distâncias Chord ou Mahalanobis e $p<0,0050$ utilizando as distâncias euclidiana, Manhatan ou Gower). A NPMANOVA forneceu resultados semelhantes com exceção da distância de Mahalanobis para a qual são obtidos valores de $p<0,05$.

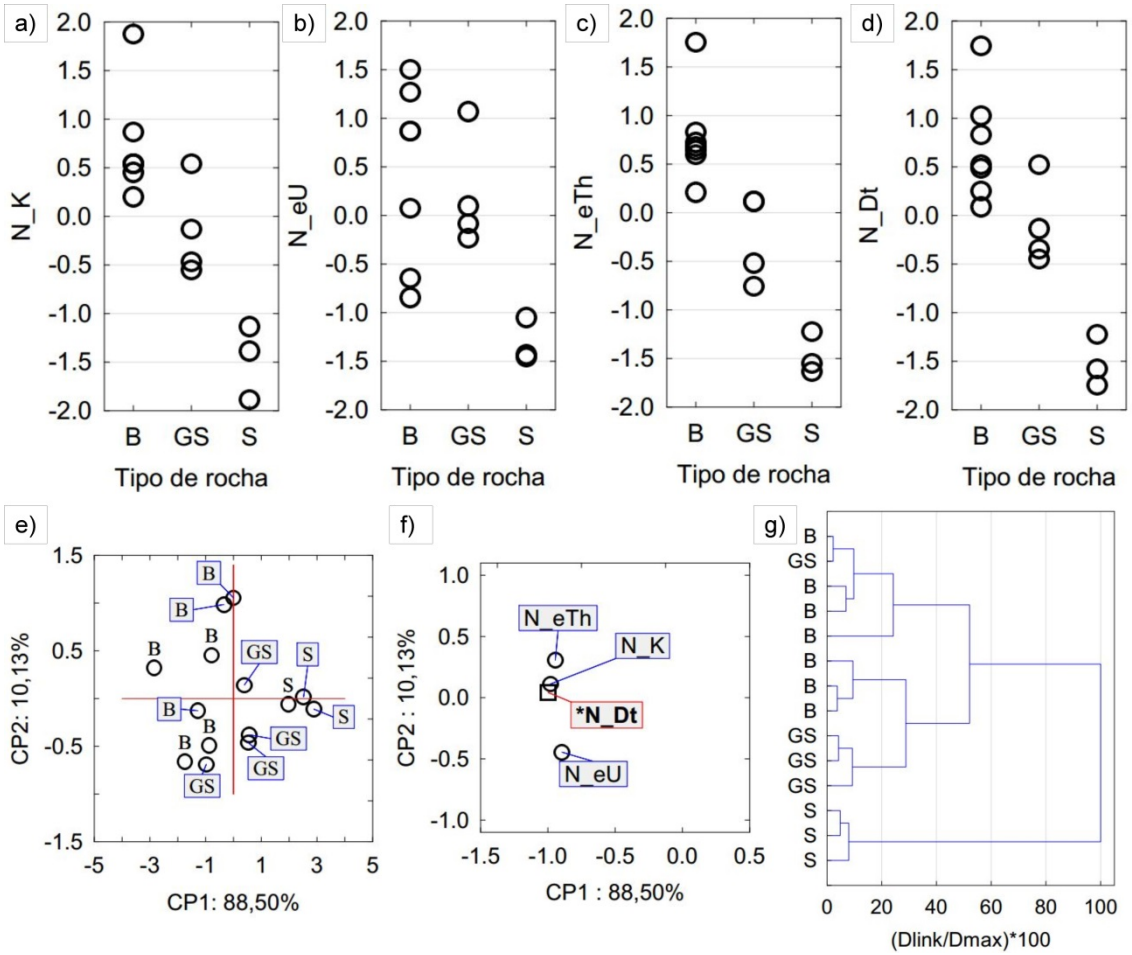


Figura 1: Análise das estimativas (normalizadas) obtidas por espetrometria gama portátil (assim como taxa de dose absorvida calculada a partir destas) em afloramentos de granitos de Braga (B) e do Sameiro (GS) e rochas metassedimentares do Silúrico (S) utilizando: diagramas de pontos para (a) K (N_K), (b) eU (N_eU), (c) eTh (N_eTh) e (d) taxa de dose absorvida (N_Dt); (e) Análise de Componentes Principais (com base em N_K, N_eU e N_eTh) das medições; (f) Análise de Componentes Principais das variáveis (N_Dt como variável suplementar); (g) dendrograma com base em N_K, N_eU e N_eTh (método de Ward).

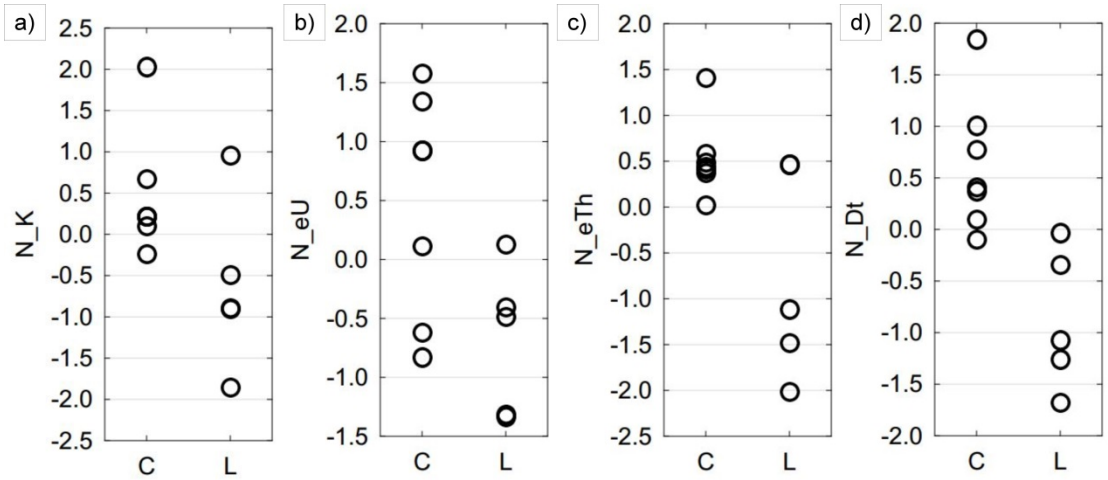


Figura 2: Diagramas de pontos para comparação das estimativas (normalizadas) de campo (C) por espetrometria gama portátil em afloramento de granito de Braga com as estimativas calculadas a partir dos resultados da espetrometria gama laboratorial (L) em amostras do mesmo granito: (a) K (N_K), (b) eU (N_eU), (c) eTh (N_eTh) e (d) taxa de dose absorvida calculada a partir destas estimativas (N_Dt).

3. CONSIDERAÇÕES FINAIS

A espetrometria gama portátil mostrou (como anteriormente) capacidade para distinguir as rochas metassedimentares silúricas das rochas graníticas, indicando que as primeiras apresentarão uma menor perigosidade em termos de radiação gama externa. Todavia, a distinção entre rochas graníticas petrologicamente semelhantes é mais complexa (possivelmente indicando níveis semelhantes de perigosidade). Será interessante estender este programa comparativo a outras variações petrológicas de rochas graníticas para avaliar a capacidade diferencial deste procedimento de campo. Na comparação das diferentes litologias, a utilização da espetrometria gama portátil enfrenta, para além das diferenças primárias (no momento de formação das rochas), os efeitos dos processos de alteração (hidrotermal/meteórica) que podem afetar as relações entre os diferentes elementos e as dificuldades na avaliação do grau de alteração no campo, problema que se estenderá aos materiais no ambiente construído (onde essa avaliação poderá ser ainda mais complexa dadas as limitações inerentes aos objetos construídos). A comparação de medições de campo no ambiente construído enfrentará, adicionalmente, dificuldades relacionadas com a existência de vários materiais no mesmo local e com as proporções desses materiais, problema que poderá requerer a definição de tipologias de locais.

Os resultados obtidos parecem sugerir que há uma tendência para valores mais elevados nas medições de campo com a espetrometria gama portátil. Embora existam vários fatores a considerar (geometria, humidade, etc.), se esta tendência for confirmada, as medições de campo serão a favor da segurança, no sentido em que os valores de campo fornecem valores mais elevados.

AGRADECIMENTOS

O Lab2PT - Laboratório de Paisagens, Património e Território (UID/AUR/04509/2013) é financiado pela Fundação para a Ciência e Tecnologia (Portugal), com verbas da República Portuguesa e com verbas da União Europeia (FEDER, Programa Operacional Factores de Competitividade – COMPETE 2020 – Programa Operacional Competitividade e Internacionalização, POCI, POCI-01-0145-FEDER-007528). Gostaríamos, também, de agradecer à Mara Dória e ao Tomaz Assunção pela colaboração no trabalho de Campo.

REFERÊNCIAS

[1] Markkanen, M. (1995). *Radiation Dose Assessments for Materials with Elevated Natural Radioactivity*. Report STUK-B-STO 32 Radiation and Nuclear Safety Authority – STU. Finnish Centre for Radiation and Nuclear Safety, Helsínquia.

[2] Dias, F., Lima, M., Sanjurjo-Sánchez, J., Alves, C. (2016). Analysis of spectra from portable handheld gamma-ray spectrometry for terrain comparative assessment. *Journal of Environmental Radioactivity*, 154, 93–100.

[3] IAEA - International Atomic Energy Agency (2003). *Guidelines for radioelement mapping using gamma ray spectrometry data. IAEA-TECDOC-1363*. Viena.

[4] UNSCEAR - United Nations Scientific Committee on the Effects of Atomic Radiation: Sources and effects of ionizing radiation (2010). *UNSCEAR 2008 report to the General Assembly, with scientific annexes*. Organização das Nações Unidas, Nova Iorque.

[5] Ferreira, N., Dias, G., Meireles, C. & Sequeira Braga, M. A. (2000). *Carta Geológica de Portugal, na escala de 1:50 000. Folha 05-D (Braga). Notícia explicativa da folha 5-D (Braga)*. Lisboa, IGM.

[6] Lima, M., Alves, C., Sanjurjo-Sánchez , J. (2015). Radiación gamma en rocas usadas como material de construcción el granito de Braga (NO de Portugal). *Cadernos do Laboratorio Xeolóxico de Laxe*, 38, 79-92

[7] Hammer, Ø., Harper, D.A.T., Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* 4(1): 9pp.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

DETECCIÓN DE ATÍPICOS ESPACIALES: UNA APLICACIÓN A DATOS DE BIOMONITORIZACIÓN

Andrea Meilán-Vila[1] y Rosa M. Crujeiras[1]

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

Resumen

En la observación de procesos espaciales, al igual que en variables escalares, es posible encontrar observaciones cuyo valor difiera significativamente de los valores de su entorno. Cuando se da esta situación, aunque la observación en cuestión no sea muy diferente de otros valores observados, se dirá que estamos ante un atípico espacial, para cuya identificación es necesario tener en cuenta tanto el valor de la observación como características espaciales de la misma. La detección de este tipo de observaciones es importante a la hora de implementar procedimientos tanto de estimación como de predicción. En este trabajo se presentan algunas de las herramientas exploratorias para la detección de atípicos espaciales, se estudia su funcionamiento mediante un estudio de simulación y se aplican a un conjunto de datos reales.

Palabras y frases clave: Atípicos espaciales, análisis exploratorio de datos espaciales, nube del variograma, diagrama de Moran.

1. Introducción

Es habitual en ciencias aplicadas como la geología, ecología o ciencias medioambientales, disponer de observaciones en las que se conoce las coordenadas geográficas en las que fueron tomadas. El mecanismo generador de este tipo de datos son los procesos estocásticos espaciales, $\{Z(\mathbf{s})|\mathbf{s} \in D\}$, constituidos por colecciones de variables aleatorias indexadas en un dominio determinado $D \subset \mathbb{R}^d$ (región de observación) con una distribución conjunta bien definida. La naturaleza del índice (localización espacial) proporciona una clasificación de los procesos espaciales (véase Cressie (1993)), siendo la geoestadística una colección de herramientas y métodos desarrollados bajo la premisa de variación continua del índice. Las observaciones derivadas de procesos espaciales, que denotaremos por $\{Z(\mathbf{s}_i),\ i = 1, \ldots, n\}$, tienden a exhibir una característica importante: observaciones cercanas tienden a ser más parecidas que las que están lejos. Por lo tanto, tales observaciones no pueden ser tratadas como independientes y la estructura de dependencia subyacente debe tenerse en cuenta tanto para la estimación de las características del proceso (en términos de tendencia o de estructura

de segundo orden) como para la predicción de valores en localizaciones no observadas. En estas situaciones, la aparición de observaciones atípicas en la muestra (en el caso de datos de control medioambiental, la presencia de datos atípicos puede venir motivada por la presencia de fuentes de contaminación) puede distorsionar los resultados obtenidos al aplicar procedimientos de estimación y/o predicción, por lo que resulta de especial relevancia su detección.

En este trabajo presentaremos algunas herramientas exploratorias que permiten la detección de atípicos espaciales (Sección 2) y analizaremos su funcionamiento a partir de un estudio de simulación (Sección 3). También se mostrará un ejemplo de aplicación a datos reales (Sección 4).

2. Técnicas exploratorias para detectar atípicos espaciales

2.1. Atípicos espaciales

Los datos espaciales son una colección de objetos referenciados espacialmente, los cuales tienen dos categorías de dimensiones de interés: espacial y no espacial. Dentro de los atributos espaciales, tenemos la ubicación (coordenadas geográficas, por ejemplo) a partir de la cual se pueden definir entornos espaciales (véase Shekhar *et al.* (2003)), mientras que como atributo no espacial se incluye la observación tomada. Consecuentemente, un valor atípico espacial se define como un objeto referenciado espacialmente cuyos valores de los atributos no espaciales son significativamente diferentes con respecto a otros objetos referenciados espacialmente de su entorno espacial. Es decir, un valor atípico espacial puede verse como una inestabilidad local (en valores de atributos no espaciales), o bien, un objeto con referencias espaciales cuyos atributos no espaciales son extremos en relación con sus vecinos, aunque pueden no ser significativamente diferentes de los restantes datos.

En el análisis exploratorio de datos, para llevar a cabo la detección de atípicos se puede emplear el diagrama de caja o boxplot, entre otros. Sin embargo, esta herramientas gráfica no es eficaz para observaciones generadas a partir de un proceso espacial ya que no tiene en cuenta la dimensión espacial. Por tanto, se requerirá de otro tipo de técnicas, como las que se describen a continuación.

2.2. Herramientas exploratorias

El estudio de atípicos espaciales puede abordarse mediante técnicas exploratorias, como el diagrama de dispersión de Moran, o bien a través de la observación de la nube del variograma. Existen otras herramientas exploratorias para su detección, entre ellas, el mapa de caja, el mapa de percentiles y el cartograma (véase Anselin (2005)).

La nube del variograma, propuesta por Cressie (1993), muestra pares de puntos vinculados por relaciones de vecindad, proporcionando una estimación de la estructura de dependencia del proceso espacial. Es decir, sería una estimación de la herramienta geoestadística básica, el variograma, el cual se denota por $2\gamma(\cdot)$ y se define como sigue

$$2\gamma(\mathbf{u}) = \mathrm{Var}(Z(\mathbf{s}+\mathbf{u}) - Z(\mathbf{s})), \quad \forall \mathbf{s}, \mathbf{s}+\mathbf{u} \in D. \tag{1}$$

Dado que la estructura de dependencia suele ser desconocida, la expresión (1) debe ser estimada. Cressie y Hawkins (1980) propusieron un estimador robusto, el cual se define por

$$2\bar{\gamma}(\mathbf{u}) = \frac{1}{(0{,}457 + 0{,}494/|N(\mathbf{u})|)} \left\{ \frac{1}{|N(\mathbf{u})|} \sum_{(i,j)\in N(\mathbf{u})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} \right\}^4, \tag{2}$$

donde $|N(\mathbf{u})|$ es el número de pares existentes en $N(\mathbf{u}) \equiv \{(i,j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{u}\}$. Si el proceso es isotrópico (la dependencia entre dos observaciones depende únicamente de la distancia entre las localizaciones donde se han tomado), entonces se puede escribir el estimador del variograma como función de $\|\mathbf{u}\| = u$. Por tanto, en la nube del variograma se representa para cada distancia u el

valor de la estimación del (semi)variograma dado en (2) (véase Figura 1, izquierda). Distancias pequeñas pero con grandes valores de la semivarianza pueden indicar un atípico espacial, a pesar de que los valores en ambas localizaciones no representen atípicos si se omite la dimensión espacial.

Otro método que también utiliza la información geográfica es el diagrama de dispersión de Moran, introducido por Anselin (1995), que proporciona una indicación del grado de asociación lineal entre los valores de los atributos normalizados, $(Z(\mathbf{s}_i) - \bar{z})/s_z$, $i = 1, \ldots, n$, ($\bar{z}$ y s_z denotan la media y desviación típica muestral de los valores de los atributos $Z(\mathbf{s}_i)$, $i = 1, \ldots, n$, respectivamente), y el promedio ponderado de los valores de los atributos vecinos normalizados, $\tilde{Z}(\mathbf{s}_i) = \sum_j w_{ij} Z(\mathbf{s}_j)$, $i = 1, \ldots, n$, donde $w_{ij} > 0$ y $\sum_j w_{ij} = 1$. El algoritmo de k-vecinos más próximos se utilizará para calcular la matriz de ponderación de los w_{ij}, y consecuentemente, los valores atípicos identificados dependerán del valor de k. Dado que la representación se realiza en base a los valores de los atributos estandarizados (estarían centrados en el origen de coordenadas; véase Figura 1, derecha), los atípicos pueden visualizarse como observaciones que están a más de dos unidades del origen.

3. Estudio de simulación

Se realiza un estudio de simulación para ilustrar el funcionamiento de las técnicas exploratorias introducidas en la sección anterior. Se considera una muestra de tamaño 121 simulada regularmente (11×11) de un proceso espacial Gaussiano (su distribución finito-dimensional es normal o Gaussiana) considerando como soporte el cuadrado unitario. Se elige el modelo exponencial como función covarianza, definida por

$$C(\mathbf{u}) = \sigma^2 \exp(\mathbf{u}/\phi), \tag{3}$$

donde σ^2 y ϕ denotan la varianza y el rango, respectivamente ($\sigma^2 = 1$ y $\phi = 2$). Destacar que, el valor del rango, 2, es mayor que cualquiera de las distancias existentes en el cuadrado unidad, lo que conlleva una dependencia fuerte. Se modifica la observación 61 (central del cuadrado) otorgándole el valor 0.6748 (mediana $+ 3RIC$, RIC: rango intercuartílico): un atributo no espacial excesivamente alto para la muestra simulada, por lo que este dato seria un atípico.

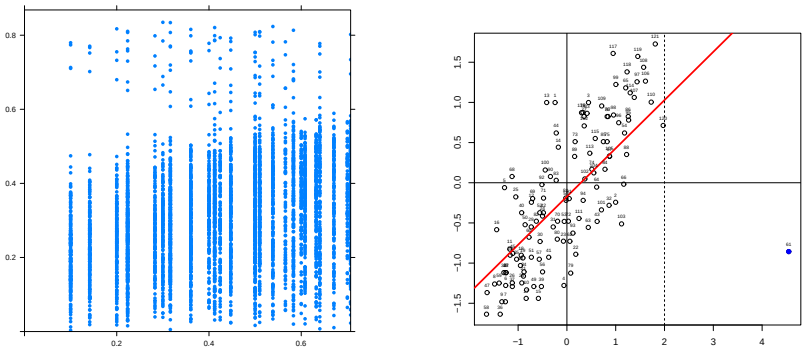


Figura 1: Herramientas exploratorias para la detección de atípicos: nube de variograma (izquierda) y diagrama de Moran con $k = 1$ (derecha, atípicos en azul). Muestra de tamaño 121 simulada regularmente de un proceso espacial gaussiano con covariograma exponencial (3) de parámetros $\sigma^2 = 1$ y $\phi = 2$.

En la Figura 1 (izquierda) se muestra la nube de variograma para los datos simulados. Se puede ver que existen varios pares de datos para los que se tienen distancias pequeñas y un valor del semivariograma grande. La observación central puede ser identificada como un atípico espacial pues aparece en los pares de datos mencionados. En la Figura 1 (derecha), se expone el diagrama de dispersión de Moran con $k = 1$ para los datos considerados. Dado que la observación central se encuentra a más de dos desviaciones estándar de la media (en el eje horizontal) podría ser candidato a ser atípico espacial. Por tanto, a partir de ambas representaciones se puede concluir que la observación central puede ser un atípico espacial del conjunto considerado. Destacar que, a

partir de un boxplot se podría identificar esta observación como un atípico global, omitiendo así la dimensión espacial. Si se considerara un valor más pequeño para ϕ en la expresión (3), ambas herramientas detectarían como atípica espacial la observación central, siempre y cuando este valor de ϕ sea superior a 0.3. Para valores inferiores, podrían identificarse como atípicos también otras observaciones. La variación de k, en este caso no influye en el resultado .

4. Aplicación a datos reales

Una vez visto el correcto funcionamiento para datos simulados, se ilustran esas técnicas para datos reales. Se consideran las concentraciones de Aluminio en musgo en Galicia en Julio de 2000. En la Figura 2 (izquierda) se muestra la nube de variograma para los datos considerados, las observaciones 56 y 80 son candidatos a ser atípicos espaciales, pues aparecen en los pares de puntos correspondientes a distancias pequeñas y valores del semivariograma grandes. En la Figura 2 (central) se puede ver el diagrama de Moran considerando $k = 1$: las observaciones $44, 56$ y 80 son posibles atípicos espaciales (se encuentran a más de dos desviaciones de la tendencia central). La observación 44 no se considera atípica espacial cuando $k = 2$, véase Figura 2 (derecha).

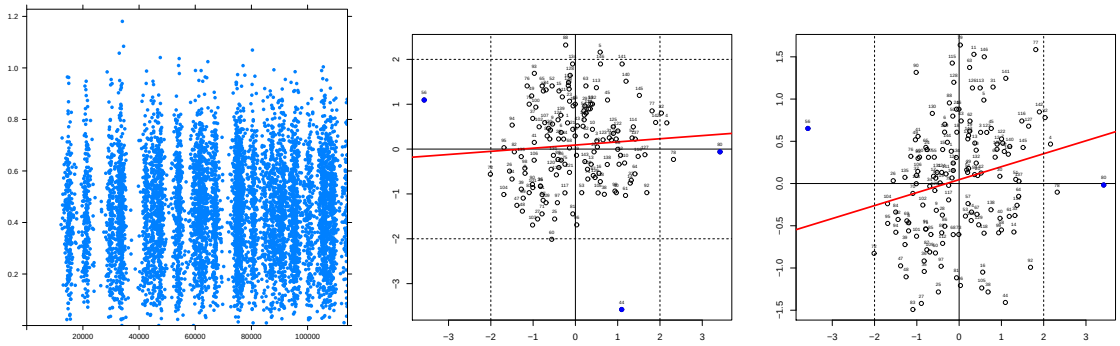


Figura 2: Herramientas exploratorias para la detección de atípicos: nube de variograma (izquierda), diagrama de Moran eligiendo $k = 1$ (central, atípicos en azul) y diagrama de Moran con $k = 2$ (derecha, atípicos en azul). Concentraciones de Aluminio en musgo en Galicia en Julio de 2000.

Agradecimientos

Las autoras agradecen la financiación del proyecto MTM2013-041383P del Ministerio de Economía y Competitividad y la red Interuniversity Attraction Poles Programme (IAP-network P7/06), Belgian Science Policy Office. También se agradece al grupo de "Ecotoxicoloxía e Ecofisioloxía Vexetal" de la Universidad de Santiago de Compostela por la cesión de los datos.

Referencias

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical analysis,*, **27**(2), 93-115.

Anselin, L. (2005). *Exploring spatial data with GeoDaTM: a workbook.* Urbana-Champaign, IL: University of Illinois, Center for Spatially Integrated Social Science, Spatial Analysis Laboratory.

Cressie, N. (1993). *Statistics for spatial data.* John Wiley & Sons, New York.

Cressie, N. y Hawkins, D.M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, **12**(2), 115-125.

Shekhar, S., Lu, C.T. y Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica*, **7**(2), 139-166.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Influencia de las variables dasométricas en la probabilidad de incendios para la especie *Pinus pinaster* en Galicia

Ginzo-Villamayor, M.J.[1], Iglesias-Allones, J.L.[2] y Marey-Perez, M.F[3]

[1] Grupo de Investigación MODESTYA GI-1914, Universidad de Santiago de Compostela
[2] ITMATI, Universidad de Santiago de Compostela
[3] Grupo de Investigación PROePLA GI-1716, Universidad de Santiago de Compostela

RESUMEN

Con el presente estudio se pretende analizar si las variables dasométricas ayudan a explicar el riesgo de incendio en parcelas dominadas por la especie *Pinus pinaster* en la comunidad de Galicia. Mediante consultas espaciales y utilizando la ubicación de los incendios, ocurridos en Galicia entre los años 2000 y 2005, se particionaron las parcelas del Tercer Inventario Forestal Nacional (IFN3) en dos grupos: parcelas afectadas o próximas a incendios y parcelas no afectadas por incendios. Se aplicó un modelo de regresión logística para estudiar las variables dasométricas en los dos grupos de parcelas apreciándose que algunas variables pueden ayudar a explicar un mayor riesgo de incendio.

Palabras y frases clave: Incendio forestal, Pinus pinaster, dasometría, regresión logística.

1. INTRODUCCIÓN

En el sur de Europa (Francia, Grecia, Italia, Portugal y España) arden cada año por incendios forestales [1], aproximadamente, medio millón de hectáreas. En los últimos años hubo un incremento en la intencionalidad como causa de estos siniestros (incendios provocados) [2]. Son diferentes las causas que explican este comportamiento, como el incremento de superficies forestales; nuevos patrones en el paisaje agrario; usos tradicionales de zonas agrícolas; la expansión de superficie matorralizada; cambios socioeconómicos y cambios en la distribución de la renta y aumento de la tasa de desempleo.
Un enfoque habitual en el estudio de los incendios forestales es la utilización de variables dasométricas y el empleo de la regresión logística como método de análisis, como se puede ver en estudios de Cataluña [3] y Portugal [4-7].

2. OBJETIVO Y ÁREA DE ESTUDIO

El **objetivo** del presente estudio es analizar las variables dasométricas, obtenidas del IFN3, como posibles variables regresoras del riesgo de incendios en parcelas de *Pinus pinaster* de Galicia.

En el caso de Galicia la superficie forestal, 2.060.453 ha, ocupa el 69% de la región. En Galicia se produjeron 245.593 incendios forestales desde el año en que existen las estadísticas en esta materia, 1961, y hasta octubre de 2011. Estos incendios abarcan una superficie de 1.794.578 ha, equivalente al 62% de la superficie total de Galicia y, al menos, un 87,5% de la superficie total forestal.

3. MATERIAL Y MÉTODOS

Para el estudio se empleó el siguiente material:

- Tercer Inventario Forestal Nacional (IFN3): se seleccionaron las parcelas ubicadas en Galicia y sus correspondientes variables dasométricas (variables explicativas de la regresión).
- El Mapa Forestal Español (MFE2): se recopilaron las teselas (masas homogéneas de arbolado) de Galicia, con sus localizaciones y características.

– Una base de datos de incendios extraída de informes del Ministerio de Agricultura, Alimentación y Medio Ambiente. Para el estudio se seleccionaron todos los incendios de Galicia en el periodo 2000-2005, lo que supone un total de 64.149 incendios.

Con la ayuda del sistema de información geográfico QGIS[1] se realizó una selección y un particionado de las parcelas del IFN3 en dos grupos: parcelas afectadas o próximas a incendios y parcelas no afectadas por incendios. Posteriormente, se aplicó un modelo de regresión logística para estudiar el efecto de las variables dasométricas en dichos grupos.

3.1 Selección de parcelas dominadas por la especie *Pinus Pinaster*:

En primer lugar, se filtraron las teselas del MFE2 cuya especie dominante fuese *Pinus Pinaster* (ver polígonos naranjas en la Figura 1). Posteriormente, mediante una consulta geográfica en QGIS se obtuvieron 1.604 parcelas del IFN3 ubicadas dentro de dichas teselas.

3.2 Particionado de parcelas.

Partiendo de las 1.604 parcelas se realiza un particionado en dos grupos: parcelas afectadas o próximas a incendios y no afectadas por incendios.

Partiendo de la idea de que las parcelas ubicadas dentro de la misma tesela tienden a tener dasometrías similares, se aplicó la siguiente metodología de particionado: en primer lugar, partiendo de las ubicaciones de los incendios del periodo 2000-2005 y mediante una consulta geográfica con QGIS, se seleccionaron las teselas en las que se haya producido algún incendio en dicho periodo. Posteriormente, mediante una nueva consulta de QGIS se asignan las parcelas a alguno de los dos grupos comentados, comprobando si están o no ubicadas en las teselas con incendios seleccionadas en la etapa previa.

Se obtienen 548 parcelas afectadas por incendio (ver puntos amarillos en la Figura 2) y 1.056 parcelas no afectadas por incendio (ver puntos azules en la Figura 2).

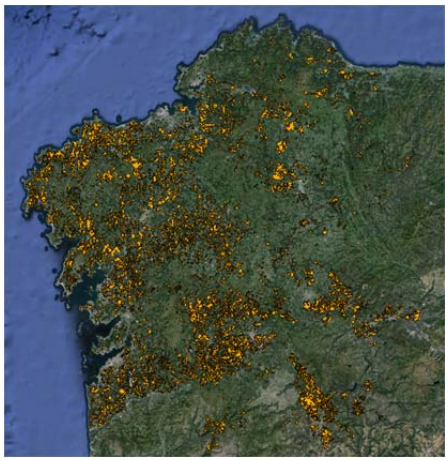


Figura 1: Teselas del MFE2 con especie dominante *Pinus Pinaster*.

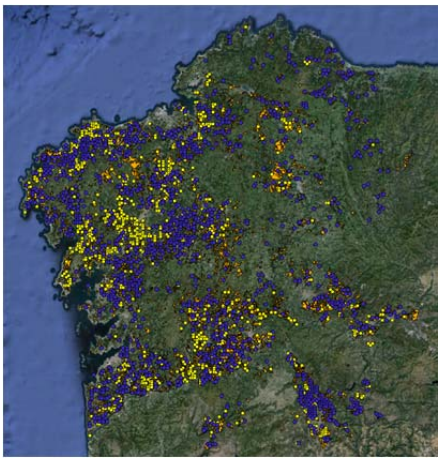


Figura 2: Localización de las parcelas del IFN3 con especie dominante *Pinus Pinaster* utilizadas en el análisis. En amarillo se representan parcelas afectadas por incendio, en azul las no afectadas.

3.3 Modelo de regresión logística

Como variable dependiente del modelo está la presencia o ausencia de algún incendio en las parcelas seleccionadas. Como variables explicativas se tomaron características dasométricas de las parcelas:

–	PROVINCIA	provincia donde está ubicada la parcela
–	n	número de pies totales en la parcela
–	prop_pinus_pinas	proporción de pies *Pinus Pinaster* (*Pinus P*) sobre total de pies en la parcela
–	Dm, Dm_26	diámetro medio de todos los pies de la parcela, y de la especie *Pinus P*
–	Dvar, Dvar_26	varianza del diámetro de todos los pies de la parcela y de la especie *Pinus P*
–	Htm, Htm_26,	altura media de todos los pies de la parcela y los de la especie *Pinus P*
–	Htvar, Htvar_26	varianza de la altura de todos los pies de la parcela y los de la especie *Pinus P*
–	Gm, Gm_26	área basimétrica media (superficie del tronco cortado a 30 cm del suelo) y de la especie *Pinus P* en

[1] Proyecto QGIS: http://www.qgis.org/es/site/

la parcela.

- Gvar, Gvar_26 — varianza del área basimétrica de todos los pies de la parcela y los de la especie *Pinus P*
- VCCm, VCCm_26 — volumen con corteza media de los pies de la parcela y de la especie *Pinus P*
- VCCvar, VCCvar_26 — varianza del volumen con corteza de todos los pies de la parcela y de la especie *Pinus P*
- VSCm, VSCm_26 — volumen sin corteza media de los pies de la parcela y de la especie *Pinus P*
- VSCvar, VSCvar_26 — varianza del volumen sin corteza de todos los pies de la parcela y de la especie *Pinus P*

Se ha empleado el lenguaje R[2] y la técnica *Backward Stepwise Regression* para seleccionar las variables del modelo de regresión logística y *ANOVA* para comparar los diferentes modelos.

4. RESULTADOS

El Modelo logístico para Galicia resultante sería:

$$P = \frac{1}{1+e^{-(1.164-1.308\,prov_{27}-0.094\,prov_{32}-0.638\,prov_{36}+0.443\text{prop_pinus_pinas}-0.009\text{Dm}+0.1\text{Htm}+0.011\text{Htvar_26}+32{,}230\text{Gm}-0.002\text{VCCm})}}$$

Donde $prov_{27}$, $prov_{32}$ y $prov_{36}$, representan las provincias de Lugo, Ourense y Pontevedra, respectivamente. La parte de Coruña va incluida en la constante (intercepto). El resto de variables han quedado explicadas en la Sección anterior. Los coeficientes del modelo se pueden ver en la Tabla 1.

	Coef. estimado	Error estándar	z valor	Pr(>|z|)	
(Intercepto)	-1,164	0,533	-2,182	0,029	*
PROVINCIA27	-1,308	0,234	-5,589	0,000	***
PROVINCIA32	-0,094	0,142	-0,659	0,510	
PROVINCIA36	-0,635	0,143	-4,455	0,000	***
prop_pinus_pinas	0,443	0,331	1,337	0,181	
Dm	-0,009	0,006	-1,578	0,115	
Htm	0,100	0,041	2,456	0,014	*
Htvar_26	0,011	0,005	2,195	0,028	*
Gm	32,230	16,320	1,974	0,048	*
VCCm	-0,002	0,001	-2,062	0,039	*
VCCvar	0,000	0,000	2,023	0,043	*
VSCvar	0,000	0,000	-1,974	0,048	*
VSCvar_26	0,000	0,000	-2,641	0,008	**

Tabla 1: Coeficientes de las variables utilizadas en el modelo.

Los resultados de la Tabla 1 muestran que la probabilidad de ocurrencia de incendio aumenta al aumentar las dasométricas Htm y Gm y por el contrario disminuye al aumentar VCCm.
En cuanto a la estimación del modelo obtenemos una tasa de acierto de un 65.71% de las veces.

También se realizó un contraste de igualdad entre las medias de todas las variables explicativas de los dos grupos de parcelas (quemadas y no quemadas). En la Tabla 2 se muestran los valores medios de las variables en las que se rechaza la igualdad de medias (alpha=0,05) junto con el p-valor obtenido.

5. CONCLUSIONES

La metodología utilizada ha sido suficientemente contrastada por otros autores para determinar que variables pueden explicar el mayor o menor riesgo de incendios en las masas arboladas.

[2] R Development Core Team. 2009. 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.

	Mínimo			Máximo			Media			
	Tot.	**No Q**	**Q**	**Tot.**	**No Q**	**Q**	**Tot.**	**No Q**	**Q**	**p-valor**
Dm	75,75	76,00	75,75	528,82	528,82	492,77	272,07	267,06	281,72	0,001
Dm_26	75,75	76,00	75,75	637,25	593,60	637,25	277,39	272,75	286,34	0,004
Dvar	0,00	0,00	0,50	68820,50	68820,50	61581,06	10501,10	9877,49	11702,79	0,000
Dvar_26	0,00	0,00	0,50	68820,50	68820,50	66795,13	8329,03	7945,51	9068,09	0,007
Htm	3,50	3,50	3,75	29,50	29,50	29,32	15,16	14,82	15,82	0,000
Htm_26	3,50	3,50	3,75	31,77	31,77	27,30	15,24	14,92	15,86	0,000
Htvar	0,00	0,00	0,00	219,17	146,64	219,17	18,21	17,07	20,40	0,003
Htvar_26	0,00	0,00	0,00	210,13	91,00	210,13	11,92	11,15	13,40	0,003
Gm	0,01	0,01	0,01	0,23	0,23	0,21	0,07	0,07	0,08	0,001
Gm_26	0,01	0,01	0,01	0,33	0,28	0,33	0,07	0,07	0,08	0,008
Gvar	0,00	0,00	0,00	0,04	0,03	0,04	0,00	0,00	0,00	0,005
VCCm	10,29	10,29	10,57	2629,09	2629,09	2230,59	534,51	510,99	579,85	0,001
VCCvar	0,02	0,03	0,02	6507375,79	6507375,79	6212527,34	296060,46	273192,83	340126,55	0,035
VSCm	6,66	6,66	6,81	3872172,86	3872172,86	3434283,05	398,69	380,32	434,09	0,001
VSCm_26	6,66	6,66	6,81	2099,84	2099,84	1844,76	402,35	387,09	431,74	0,008

Tabla 2: Valores mínimos, máximos y medios de las variables en las que se rechaza la igualdad de medias y p-valor obtenido. ***Tot.*** Representa el total para Galicia, ***No Q*** teselas no quemadas y ***Q*** teselas quemadas.

Los resultados en Galicia muestran que no existe gran diferencia, en términos dasométricos, entre las parcelas que han experimentado y no incendios. La explicación a estos resultados podría estar en que se trata de incendios provocados que cuando se producen en zonas forestales arboladas no parecen seguir un patrón de variables relativas al tipo de arbolado diferenciado.
Este trabajo forma parte de la línea de investigación que se está llevando desde este grupo y en el futuro está previsto profundizar en el detalle del análisis especialmente en diferenciar el efecto de la reincidencia de incendios en una zona y su relación con el tipo de arbolado.

AGRADECIMIENTOS

Los investigadores Mª José Ginzo y José Luis Iglesias agradecen el apoyo del proyecto ENJAMBRE del CDTI, además Mª José a los proyectos 2013-PG064 de la Consellería de Educación e Ordenación Universitaria, al MTM2013-41383-P del Ministerio de Economía y Competitividad y a la Red IAP P7/06 de Belgian Science Policy.

Referencias

[1] Moreira, F., Viedma, O., Arianoutsou, M., Curt, T., Koutsias, N., Rigolot, E., Barbati, A., Corona, P., Vaz, P., Xanthopoulos, G., Mouillot, F. & Bilgili, E. 2011. Landscape wildfire interactions in southern Europe: implications for landscape management. Journal of Environmental Management, 92, 2389–2402.
[2] Juan, P., Mateu, J., Saenz, M. 2012. Pinpointing spatio-temporal interactions in wildfire patterns. Stoch Environ Res Risk Assess 26, 1131-1150.
[3] Gonzalez-Olabarria, J.R., Brotons, L., Gritten, D., Tudela, A. and Teres, J.A. 2012. Identifying location and causality of fire ignition hotspots in a Mediterranean region. International Journal of Wildland Fire 21, 905-914.
[4] García-Gonzalo, J., Pukkala, T., Borges, J.G. (2011). Integrating fire risk in stand management scheduling. An application to Maritime pine stands in Portugal. Annals of Operation Research 219, 379-395.
[5] Marques, S., García-Gonzalo, J., Botequim, B., Ricardo, A., Borges, J.G., Tome, M., Oliveira, M.M. (2012). Assessing wildfire occurrence probability in *Pinus pinaster* Ait. stands in Portugal. Forest Systems 21, 111-120.
[6] García-Gonzalo, J., Marques, S., Borges, J.G., Botequim, B., Oliveira, M.M., Tomé, J., Tomé, M. (2011). A three-step approach to post-fire mortality modelling in maritime pine (*Pinus pinaster* Ait) stands for enhanced forest planning in Portugal. Forestry 84, 197-206.
[7] Botequim, B., García-Gonzalo, J., Marques, S., Ricardo, A., Borges, J.G., Tomé, M., Oliveira, M.M. (2013). Developing wildfire risk probability models for *Eucalyptus globulus* stands in Portugal. Iforest 6, 217-227.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

MODELAÇÃO ESPACIAL APLICADA A DADOS DE PRECIPITAÇÃO EXTREMA NA ILHA DA MADEIRA

Prata Gomes, D.[1], Gouveia-Reis, D.[2] e Neves, M.M.[3]

[1]CMA e Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Costa da Caparica, 2829-516 Caparica, Portugal
[2]CEAUL e Faculdade de Ciências Exactas e da Engenharia, Universidade da Madeira, Campus Universitário da Penteada, 9020-105 Funchal, Portugal
[3]CEAUL e Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisboa, Portugal

RESUMO

O estudo de dados de valores extremos, quando as observações são localizadas no espaço e/ou tempo, tem vindo a revelar cada vez maior interesse pela necessidade de estimar e prever o comportamento extremal de fenómenos ambientais localizados espacialmente. Neste trabalho apresenta-se uma revisão de procedimentos que constituem investigação recente na modelação de extremos espaciais. De entre várias abordagens referidas na literatura os processos max-estáveis são os mais usados. São esses que serão aqui revistos e utilizados num conjunto de dados geo-referenciados, referentes a valores máximos mensais, de precipitação diária registada no período de 1970 a 2016, em estações na ilha da Madeira, identificadas por estudos prévios como localizadas em zonas de ocorrência valores extremos de precipitação. Serão usados *packages* recentes do *software* R que contém alguns dos modelos max-estáveis mais usados.

Palavras e frases chave: Extremos espaciais; Madeira; precipitação extrema; processos max-estáveis

1. INTRODUÇÃO e NOÇÕES PRELIMINARES

A necessidade de utilizar modelos adequados de teoria de valores extremos para explicar, modelar e prever o comportamento extremal de processos ambientais é bem reconhecida há muito tempo. Encontram-se inúmeras referências à aplicação da teoria de valores extremos no estudo da precipitação, rajadas de vento, temperaturas, níveis de mar, etc. Modelos e procedimentos inferenciais foram desenvolvidos na primeira metade do séc. XX para estudar observações extremas. Estas observações são, por definição, raras e inesperadas. Métodos univariados e multivariados para estudar o comportamento extremal numa localização fixa e também para processos estacionários têm sido intensamente estudados e desenvolvidos. Porém, recentemente a investigação nesta área tem-se dirigido para o desenvolvimento de modelos e métodos de estudo dos valores extremos quando as observações apresentam uma estrutura mais complexa – uma estrutura de dependência espacial, que terá que ser tida em conta nas análises a realizar. Efectivamente as variações na intensidade, persistência e frequência de fenómenos extremos poderá ter impacte local e regional, nomeadamente no risco associado a situações de cheia.

Faz-se uma breve revisão dos principais aspectos da teoria de extremos univariados e multivariados mostrando a necessidade de incluir a geoestatística no tratamento de situações de dimensionalidade infinita. É feita uma breve revisão aos modelos max-estáveis habitualmente usados e ilustra-se a sua aplicação a dados de máximos mensais de registos diários de precipitação na ilha da Madeira, no período de 1970 a 2016, recorrendo a *packages* do *software* R (Ribatet et al., 2013).

2. EXTREMOS ESPACIAIS E OS MODELOS MAX-ESTÁVEIS

Muitos problemas de extremos de interesse nas áreas do ambiente, climatologia, hidrologia, usam observações registadas em estações de monitorização especificamente localizadas, pelo que é necessário o recurso a metodologias de modelação espacial. A resposta a diferentes aplicações que foram surgindo motivou e incentivou o desenvolvimento de métodos e ferramentas para análise e caracterização espacial de dados de extremos. Compreender a variabilidade espaço-temporal de acontecimentos extremos é crucial para prever a ocorrência desses extremos. Mas a geoestatística convencional não pode ser aqui aplicada uma vez que é fortemente baseada em processos gaussianos e o comportamento dos extremos está longe de ser normal. Uma das ferramentas crucias em geoestatística, o variograma, pode não existir porque pode acontecer que o valor esperado e portanto a variância do processo espacial sejam infinitos. Métodos para modelar extremos espaciais, baseados na generalização de procedimentos de extremos multivariadas foram propostos por Tawn (1988, 1990) e Smith et al. (1990). Smith (1990) sugeriu procedimentos da teoria de processos max-estáveis para modelar observações registadas numa grelha de pontos no espaço. Os métodos estatísticos para estes processos e a sua aplicação a dados reais podem ser encontrados em Smith (1990), Coles (1993), Coles e Walshaw (1994), Coles e Tawn (1990, 1996) e Joe et al. (1991). Aplicações mais recentes para dados de precipitação podem ver-se em Buishand et al. (2008), Smith e Stephenson (2009), Padoan et al. (2010), Davison et al. (2012), Olinda et al. (2014) e Neves et al. (2011) e Prata Gomes et. al (2015); na modelação dos extremos de neve Blanchet e Lehning (2010), Blanchet e Davison (2011), dados de temperatura máxima Davison e Gholamrezaee (2011). No tratamento de uma aplicação deverá verificar-se se os dados exibem algum tipo de dependência espacial antes de modelar os extremos espaciais. A forma mais conveniente de estudar a dependência é através do F-madograma, $\upsilon_F(h)$, e da sua relação com o coeficiente extremal, $\theta(h)$ (Cooley et al., 2006). Se $\{Z(x)\}_{x\in K}$ é um processo max-estável Fréchet unitário, então

$$\upsilon_F(h) = \frac{1}{2}E[|F\{Z(0)\} - F\{Z(h)\}|], \qquad \theta(h) = \frac{1+2\upsilon_F(h)}{1-2\upsilon_F(h)}, \tag{1}$$

em que $1 \leq \theta(h) \leq 2$ com limite inferior e o limite superior correspondendo à dependência completa do máximo em duas localizações e à independência assintótica, respectivamente.

Os processos max-estáveis são os modelos naturais para extremos espaciais. Na última década estes processos têm estado a ocupar um papel de relevo na modelação estatística de processos extremais que incorporam uma dependência temporal ou espacial. Os processos max-estáveis foram desenvolvidos, inicialmente, por de Haan em 1984. Desde então vários autores, dos quais citamos Smith (1990), Schlather (2002), Kabluchko, Schlather e de Haan (2009), Ribatet (2013), Davison e Gholamrezaee (2011), Blanchet e Davison (2011) e Cooley et al. (2012) desenvolveram o processo max-estável no sentido de torná-lo mais flexível na sua aplicação. Todos os dados são de dimensão finita, no entanto este enquadramento é inadequado para lidar com localizações não observadas e por esta razão assume-se que estamos perante um processo estocástico. Seja S a região em estudo e $\mathbf{s}$ uma localização dessa região. Note-se que nas aplicações espaciais quase sempre tem-se que $\mathbf{s} \in \mathbf{R}^2$. Assumiremos que $X_i(\mathbf{s})$ é um processo estocástico e neste caso pode pensar-se em i como indexando o dia da observação. A motivação assintótica para estes processos é semelhante ao caso da distribuição generalizada de valores extremos univariada. A definição do processo max-estável como uma generalização infinita da distribuição max-estável multivariada dá-nos uma classe bem definida de modelos mas não sugere como construir estes processos. Uma construção conceiptual destes processos foi dada primeiramente com uma representação espectral do processo estremal (de Haan, 1984; de Haan e Ferreira, 2006). Seja $\{X_i(\mathbf{s})\}_{i\in\mathbf{N}}$ realizações independentes do processo estocástico $\{X(\mathbf{s})\}$ e ζ_i pontos de um processo de Poisson. Então

$$Z(x) = max_{i\geq 1}\zeta_i Y_i(x), \quad x \in X \subset \mathbf{R}^d \tag{2}$$

é um processo max-estável com distribuições marginias Fréchet unitária. Diferentes escolhas de $X_i(\mathbf{s})$ conduzem a diferentes classes de processos max-estáveis. Assim, se considerarmos $X_i(\mathbf{s})$ com densidade gaussiana com matriz de covariância Σ, obtém-se o processo de Smith (1990). Schlather (2002) sugeriu uma classe de processos mais flexiveis tomando para $X_i(\mathbf{s})$ um processo gaussiano isotrópico estacionário com variância unitária, correlação $\rho(\cdot)$ e $\mu^{-1} = \sqrt{2\pi}$. Kabluchko et al. (2009) propuseram como alternativa para $X_i(\mathbf{s})$ a exponencial de um processo gaussiano com incrementos estacionarios substraindo $1/2$ do semivariograma, γ. Este processo é designado por processo Brown-Resnick. Considerando $X_i(\mathbf{s})$ um processo gaussiano standard com um expoente $\nu \geq 1$ obtêm-se o processo Extremal-t. Para o modelo de Smith os parâmetros de dependência são os elementos da matrix de covariância que aparecem nas densidades gaussianas, no modelo de Schlather são os parâmetros da função de correlação, para o modelo Brown-Resnick são os parâmetros do semivariograma e para o modelo Extremal-t são os parâmetros da função de correlação e os graus de liberdade. Sendo então fixa a distribuição das marginais, é apenas necessário estimar os parâmetros de dependência e, ajustado o modelo aos dados, simulamos a partir deste.

3. APLICAÇÃO A DADOS DE PRECIPITAÇÃO EXTREMA NA MADEIRA

Os dados que vão ser usados neste estudo referem-se aos máximos mensais de precipitações diárias registadas em 5 estações meteorológicas na ilha da Madeira no período de 1970 a 2016. As estações consideradas foram Bica da Cana, Funchal, Lugar de Baixo, Santa Catarina e Santo da Serra. A sua localização, aproximada, está indicada na Figura 1 e é representada pelas letras A a E, respectivamente. Para outras estações estão disponíveis dados, no entanto os períodos de registo são mais curtos e/ou apresentam muitos valores omissos. Por essa razão neste estudo consideramos apenas estas estações.

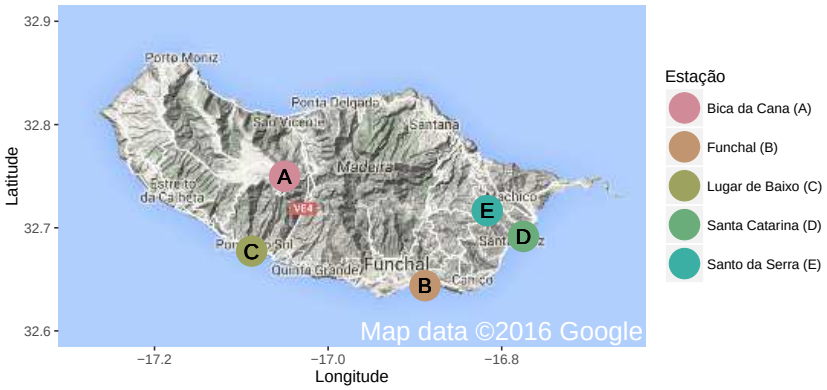


Figura 1: Localização das estações em estudo.

A topografia da região influencia bastante os valores da precipitação pelo que, nas estações em estudo, ter-se-á ainda em consideração a sua altitude e serão exploradas outras possíveis covariáveis. Todo o estudo é desenvolvido em ambiente R, sendo usado principalmente o *package* `SpatialExtremes`, na pesquisa da dependência espacial, estimação dos parâmetros de dependência, elaboração de gráficos e modelação e previsão de superfícies de resposta.

AGRADECIMENTOS

Este estudo foi financiado pela FCT (Fundação para Ciência e Tecnologia) através do projeto UID/MAT/00297/2013 (Centro de Matemática e Aplicações) e do projecto UID/MAT/00006/2013 (Centro de Estatística e Aplicações). Gostaríamos de agradecer ao Instituto Português do Mar e da Atmosfera pelos dados disponibilizados e, em particular ao Doutor Victor Prior.

Referencias

[1] Blanchet, J., Davison, A. C. (2011). Spatial modeling of extreme snow depth. *The Ann. App. Stat.* 5 (3), 1699-1725.

[2] Blanchet, J.; Lehning, M. (2010). Mapping snow depth return levels: smooth spatial modeling versus station interpolation. *Hydrol. Earth Syst. Sci., Gottingen*, v.14, n.12, 2527–2544.

[3] Buishand, T. A., de Haan, L., Zhou, C. (2008). On spatial extremes: With application to a rainfall problem. *Ann. Appl. Stat.* 2, 624-642.

[4] Coles, S. G., (1993). Regional modelling of extreme storms via max-stable processes. *J. R. Stat. Soc. B* 55, 797–816.

[5] Coles, S.G., Tawn, J.A. (1990). Statistics of coastal flood prevention. *Phil. Trans. R. Soc. Lond.* A 332, 457-476.

[6] Coles, S. G., Tawn, J. A. (1996). Modelling extremes of the areal rainfall process. *J. R. Stat. Soc.* B 58, 329–347.

[7] Coles, S. G., Walshaw, D. (1994). Directional modelling of extreme wind speeds. *Appl. Stat.* 43, 139–157.

[8] Cooley, D., Naveau, P., Poncet P., (2006). Variograms for spatial max-stable random fields. *Lect. Notes Statist.* 187, 373-390, Springer, New York.

[9] Cooley, D., Cisewski, J., Erhardt, R.J., Jeon, S., Mannshardt, E., Omolo, B.O., Sun, Y. (2012). A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *REVSTAT* 10, 135–165.

[10] Davison, A., Gholamrezaee, M. (2011). Geostatistics of extremes. *J. R. Soc. Series A*, London, 581-608.

[11] Davison, A., Padoan, S., Ribatet, M. (2012). Statistical modeling of spatial extremes (with discussion). *Stat. Sci.* 7, 161–186.

[12] Haan L de., (1984). A spectral representation for max-stable processes. *Ann. Prob.* 12: 1194-1204.

[13] Haan, L de., Ferreira, A. (2006). *Extreme Value Theory*, Springer, New York.

[14] Joe, H., Smith, R.L., Weissman, I. (1991). Bivariate threshold models for extremes. *J. Roy. Statist. Soc.* B 54, 171–183

[15] Kabluchko, Z., Schlather, M., de Haan, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Prob.* 37, 2042-2065.

[16] Neves, M. M., Prata Gomes, D., (2011). Geostatistics for spatial extremes. A case study of maximum annual rainfall in Portugal. *Procedia Environmental Sciences,* Spatial Statistics, Elsevier 7, 246-251.

[17] Olinda, R. A., Blanchet, J., Ozaki, V. A., Ribeiro Jr, P. J. (2014). Spatial modeling applied to extreme precipitation data in the State of Paraná. Hydrol. Earth Syst. Sci. Discuss., 11, 12731-12764.

[18] Padoan, S., Ribatet, M., Sisson, S., (2010). Likelihood-based inference for max-stable processes. *J. Am. Stat. Assoc.* 105, 263–277.

[19] Prata Gomes, D., Neves, M. M., Moreira, E. (2015). An exploratory study of spatial annual maximum of monthly precipitation in the northern region of Portugal. *Physics and Chemistry of the Earth Sci. Direct*. doi 10.1016/j.pce.2015.12.001.

[20] Ribatet, M., (2013). Spatial extremes: Max-stable processes at work. *Journal de la Société Française de Statistique*, 154(2), 156-177.

[21] Ribatet, M. R. Singleton, R Core Team, (2013). *SpatialExtremes: Modelling Spatial Extremes 2013*. R package version 2.0-1.

[22] Schlather, M. (2002). Models for stationary max-stable random fields. *Extremes* 5, 33-44.

[23] Smith, R., (1990). Max-stable processes and spatial extremes. *Unpublished manuscript.*

[24] Smith, R.L., Tawn, J.A., Yuen, H.K. (1990). Statistics of multivariate extremes. *Internat. Statist. Rev.* 58, 47–58.

[25] Smith, E. L., Stephenson, A. G. (2009). An extended Gaussian max-stable process model for spatial extremes. *J. Stat. Plan. Inf.* 139, 1266– 1275.

[26] Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika* 75, 397–415.

[27] Tawn, J.A. (1990). Modelling multivariate extreme value distributions. *Biometrika* 77, 245-253.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

MODELLING ENVIRONMENTAL MONITORING DATA COMING FROM DIFFERENT SURVEYS

L. Margalho [1], R. Menezes [2] and I. Sousa[2]

[1] Dep. of Physics and Mathematics-Coimbra Institute of Engineering & CMAT-Minho University
[2] Dep. of Mathematics and Applications-Minho University & CMAT-Minho University

ABSTRACT

Environmental monitoring networks are providing large amounts of spatio-temporal data. Air pollution data, as other environmental data, exhibit a spatial and a temporal correlated nature. To improve the accuracy of predictions at unmonitored locations, there is a growing need for models capturing those spatio-temporal correlations.
With this work, we propose a spatio-temporal model for gaussian data collected in a few number of surveys. We assume the spatial correlation structure to be the same in all surveys. In an application of this model to real data, concerning heavy metal concentrations in mosses collected from three surveys occurring between 1992 and 2002 in mainland Portugal, the data set is dense in the spatial dimension but sparse in the temporal one, thus our model-based approach corresponds to a saturated correlation model in the time dimension. A novel interpretation for the space-time covariance function is introduced. A simulation study, aiming to validate the model, provided better results in terms of accuracy with the novel covariance function.
Prediction maps of the observed variable for the most recent survey, and of the interpolation error as a measure of accuracy, are presented.

Keywords: Environmental pollution monitoring; Space-time modelling; Sparse time dimension; Separable covariance structure.

1. INTRODUCTION

Nowadays, due to technology developments and worldwide policies, environmental monitoring networks are providing large amounts of data exhibiting a spatial and a temporal correlated nature, and as a consequence a large number of models and techniques to analyze this sort of data has emerged.
Although in environmental sciences data are typically collected through monitoring stations, it may also be collected through biomonitoring surveys covering extensive areas. Some examples of studies involving moss samples as biomonitors of atmospheric heavy metal deposition are Aboal *et al.* (2006) and Diggle *et al.* (2010) with data from Galicia, northern Spain, Harmens *et al.* (2010) considering data from several countries across Europe, Steinnes *et al.* (2003) and Steinnes *et al.* (2011) concerning Norway data, Zechmeister *et al.* (2008) with data from Austria.
It is common to have studies, such as the ones mentioned before, involving environmental spatio-temporal data containing a dense time dimension but only a sparse spatial one, as a result of the

easiness of gathering data enabled by modern technologies. That is not the case of the biomonitoring data being used in this work, which are related to measurements of heavy metal concentrations made at 146 spatial locations in only 3 surveys.
Our aim is to propose a naive spatio-temporal framework which incorporates into the model both time and space correlations, capable to fit spatio-temporal data containing a reduced number of time observations. Due to this particular characteristic of having few temporal records, and under the hypothesis of separability of the correlation structure, it may be the case that the number of parameters to estimate in the temporal correlation function equals the number of temporal observations, which corresponds to have a saturated correlation model in the time dimension, *i.e.*, a model perfectly reproducing the data.

2. MODEL PROPOSAL

We propose a spatio-temporal model for Gaussian data, collected at location $\mathbf{s}$ and time t,

$$Y(\mathbf{s},t) = \mu(\mathbf{s},t) + \boldsymbol{Z}(\mathbf{s},t) + \boldsymbol{\varepsilon}(\mathbf{s},t) \tag{1}$$

The mean component $\mu(\mathbf{s},t)$, depending on possibly observed covariates $f_i(\mathbf{s},t)$, will be considered as

$$\mu(\mathbf{s},t) = \sum_{i=1}^{p} \beta_i f_i(\mathbf{s},t) \tag{2}$$

where $E[Y(\mathbf{s},t)] = \mu(\mathbf{s},t)$. The non-observed spatio-temporal process $\boldsymbol{Z}(\mathbf{s},t)$ is defined by

$$\boldsymbol{Z} \sim MVN\left(\mathbf{0}, \Sigma\right) \tag{3}$$

and $\boldsymbol{\varepsilon}$ represents Gaussian space-time measurement errors,

$$\boldsymbol{\varepsilon} \sim MVN\big(\mathbf{0}, \tau^2 I_{NT}\big) \tag{4}$$

We assume a separable covariance structure,

$$\begin{aligned} \Sigma^{i,j,k,l} &= \mathrm{Cov}_{ST}\big[Z(\mathbf{s}_i,t_k), Z(\mathbf{s}_j,t_l)\big] \\ &= \mathrm{Cov}_S\left(\|\mathbf{s}_i - \mathbf{s}_j\|\right) \times \mathrm{Cov}_T\left(|t_k - t_l|\right) \\ &= \mathrm{Cov}_S\left(h_S\right) \times \mathrm{Cov}_T\left(h_T\right) \end{aligned}$$

and propose two interpretations for Σ

- $\Sigma(h_S,h_T) = \sigma^2_{total}\, R_S(h_S) \otimes R_T(h_T)$ (Rodriguez-Iturbe & Mejia (1974))
- $\Sigma(h_S,h_T) = \sigma^2_S R_S(h_S) \otimes \sigma^2_T R_T(h_T)$

($\otimes$ represents the Kronecker product of matrices). In the application of the model, the covariates included in the model are indexed in space (the sampling locations intensity) and indexed in time (the specific contribution of a given survey).

3. SIMULATION STUDY

For a model validation purpose, a simulation study was conducted. A set of 50 randomly chosen space locations was considered in the square $[0,1]^2$. In order to have a region with more intensified sampling density, mimicking the behavior of the real data set used in the application, 15 of those locations belong to the square $[0.45, 0.55]^2$. Observations are assumed to be collected at 3 different moments, according to an AR(2) model. Using the absolute value of the coefficient of variation as an accuracy measure, simulations with $\Sigma(h_S,h_T) = \sigma^2_S R_S(h_S) \otimes \sigma^2_T R_T(h_T)$ provided better results.

4. CASE STUDY. HEAVY METAL CONCENTRATION IN MOSSES

Mosses are widely used as biomonitors of atmospheric heavy metal deposition. In Europe, they have been used since 1990, with the aim of map spatial and temporal patterns of accumulation in ecosystems, after the establishment of the international mapping project *Atmospheric Heavy Metal Deposition in Europe*, which is surveying the atmospheric deposition of heavy metals using moss species as biomonitors, with the aim of investigate the existence of correlations between heavy metal concentrations in mosses. Portugal was one of the participating countries in the mentioned project, performing surveys every 5 years since its beginning in 1990. Moss samples of species *Hypnum cupressiforme* and *Scleropodium touretti* were collected in three nationwide surveys across mainland Portugal, referred to as the 1992, 1996 and 2002 surveys.
The model described previously assumes that the hidden process $\boldsymbol{Z}(\mathbf{s},t)$ and the measurement error $\boldsymbol{\varepsilon}(\mathbf{s},t)$ are Gaussian. It is well known that for a non-observed location $\mathbf{s}_0$ and a time t_0, the joint distribution of $Y(\mathbf{s}_0,t_0)$ and $Y(\mathbf{s},t)$ is

$$\begin{bmatrix} Y(\mathbf{s}_0,t_0) \\ \mathbf{Y}(\mathbf{s},t) \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \mu(\mathbf{s}_0,t_0) \\ \mu(\mathbf{s},t) \end{bmatrix}, \begin{bmatrix} C_{0,0} & \mathbf{c}_0^T \\ \mathbf{c}_0 & \mathbf{C}_Y \end{bmatrix} \right) \tag{5}$$

where $\mu(\mathbf{s}_0,t_0)$ and $\mu(\mathbf{s},t)$ are defined by (2), $C_{0,0} = \text{Var}(Y(\mathbf{s}_0,t_0))$, $\mathbf{c}_0 = \text{Cov}(Y(\mathbf{s}_0,t_0), Y(\mathbf{s},t))$, and $\mathbf{C}_Y = \Sigma + \tau^2 I$.
Under the assumption (5), the predicted value at an unsampled location $Y^*(\mathbf{s}_0,t_0)$ is given (Cressie and Wikle (2011)) by

$$Y^*(\mathbf{s}_0,t_0) = \text{E}\left[Y^*(\mathbf{s}_0,t_0)|\mathbf{Y}(\mathbf{s},t)\right] = \mu(\mathbf{s}_0,t_0) + \mathbf{c}_0^T\mathbf{C}_Y^{-1}\left(\mathbf{Y}(\mathbf{s},t) - \mu(\mathbf{s},t)\right) \tag{6}$$

and the variance of the prediction is

$$\sigma^2(\mathbf{s}_0,t_0) = \text{E}\left(Y(\mathbf{s}_0,t_0) - Y^*(\mathbf{s}_0,t_0)\right)^2 = C_{0,0} - \mathbf{c}_0^T\mathbf{C}_Y^{-1}\mathbf{c}_0 \tag{7}$$

The estimates of the model parameters are in Table 1. The computation of the standard errors was made via Monte-Carlo simulation.

Param.	β_0	β_1	β_2	β_3	ρ_{12}	ρ_{13}	ρ_{23}	σ_S^2	σ_T^2	τ^2	ϕ
Estim.	7.45	0.01	0.15	-0.24	0.97	0.91	0.96	0.98	1.48	1.02	58624.08
St. Error	0.04	<0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	1470.20

Table 1: Model parameter estimates with standard errors

Figure 1 shows the predicted Mn transformed concentration map for the most recent survey and the corresponding interpolation error map, over mainland Portugal.

5. CONCLUSIONS

The new model proposal provides better results in terms of interpolation prediction error, than the ones obtained in Margalho *et al.* (2014) with the same data set using a different geostatistical model.

References

[1] Aboal, J.; Real, C.; Fernández, J. and Carballeira, A. (2006). Mapping the results of extensive surveys: The case of atmospheric biomonitoring and terrestrial mosses, *Science of The Total Environment*, **356**, 256–274.

[2] Cressie, N. and Wikle, C. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons Inc, New Jersey.

[3] Diggle, P.; Menezes, R. and Su, T. (2010). Geostatistical inference under preferential sampling, *Applied Statistics*, **59**, 2, 191–232.

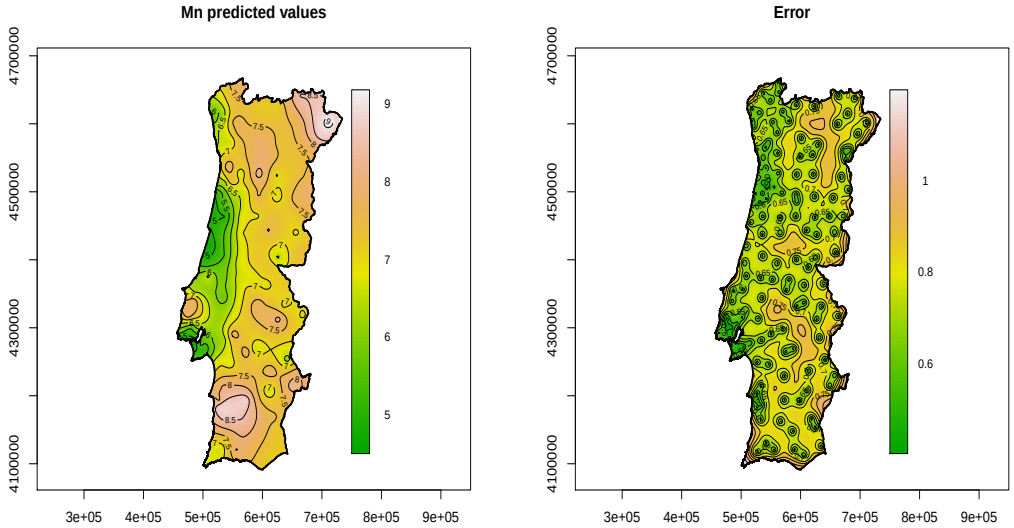


Figure 1: Prediction map of the Box-Cox transformed Mn data (left) and interpolation error map (right).

[4] Harmensm H.; Norris, D.; Steinnes, E.; Kubin, E.; Piispanen, J.; Alber, R.; Aleksiayenak, Y.; Blum, O.; Coskun, M.; Dam, M.; De Temmerman, L.; Fernández, J.; Frolova, M.; Frontasyeva, M.; González-Miqueo, L.; Grodzińska, K.; Jeran, Z.; Korzekwa, S.; Krmar, M.; Kubin, E.; Kvietkus, K.; Leblond, S.; Liiv, S.; Magnússon, S.. Maňkovská, B.; Pesch, R.; Rüling, A.; Santamaria, J.; Schöder, W.; Spiric, Z.; Suchara, I.; Thöni, L.; Urumov, L.; Yurukova, L. and Zechmeister, H. (2010). Mosses as biomonitors of atmospheric heavy metal deposition: Spatial patterns and temporal trends in Europe, *Environmental Pollution*, **158**, 3144–3156.

[5] Margalho, L.; Menezes, R. and Sousa, I. (2014). Assessing interpolation error for space-time monitoring data, *Journal of Stochastic Environmental Research and Risk Assessment*, **28**, 5, 1307–1321.

[6] Rodriguez-Iturbe, I. and Mejia, J. (1974). The design of networks in time and space, *Water Resources Research*, **10**, 4, 713–728.

[7] Steinnes, E.; Berg, T. and Sjøbakk, T. (2003). Temporal and spatial trends in Hg deposition monitored by moss analysis, *The Science of the Total Environment*, **304**, 215–219.

[8] Steinnes, E.; Berg, T. and Uggerud, H. (2011). Three decades of atmospheric metal deposition in Norway as evident from analysis of moss samples, *The Science of the Total Environment*, **412-413**, 351–358.

[9] Zechmeister, H.; Hohenwallner, D.; Hanus-Illnar, A.; Hagendorfer, H.; Roder, I. and Riss, A. (2008). Temporal patterns of metal deposition at various scales during the last two decades, *Atmospheric Environment*, **42**, 1301–1309.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Searching for soil pollution sources by means of geographically weighted principal components analysis (GWPCA)

Javier Roca-Pardiñas[1], Celestino Ordóñez [2], Tomás R. Cotos-Yáñez[1] and Rubén Pérez-Álvarez[3]

[1] Department of Statistics and Operations Research, University of Vigo, 36208 Vigo, Spain
[2] Department of Mining Explotation and Prospecting, University of Oviedo, Spain
[3] Department of Transports and Technology of Projects and Processes, University of Cantabria, Spain

RESUMEN

GWPCA is applied to determine patterns of spatial distribution as well as possible pollution sources in an industralized area. First we analyze the existence of spatial heterogeneity in the data using a statistical contrast test . Then, we apply GWPCA to the data and evaluate the spatial distribution of the winning variables for the three first principal components. The results obtained indicate that there is spatial heterogeneity in the spatial distribution of the pollutans and that this distribution is mainly related to industrial and urban activities.

Keywords:: Spatial data; principal components, covariance structure, pollution

1. INTRODUCTION

Geographically weighted principal component analysis (GWPCA) is preferred to standard PCA when there is spatial heterogeneity in the data. In summary, GWPCA performs a local PCA assuming that the covariance matrix changes along the space ([1], [2], [3]). Consequently, differents results of the standard PCA, such as loadings or component scores, are obtained for each observation . In order to justify the use of GWPCA instead of PCA, a previous analysis should be performed to check the existence of spatial heterogeneity ([3]). Normally this analysis is based on determining a statistics that depends on the eigenvalues of the covariance matrix ([4]), although we propose a different test that uses the eigenvectors of new variables obtained after applying a standard PCA to the raw data. Similarly to PCA, the analysis of the WGPC loadings in each component can help to find relationships among data that could not be found otherwise. However, as PCA does not consider the possibility of changes in the covariance structure along the study area, some of that relantionships can be obscured.

Let us consider that $\mathbf{x}$ is a vector of p random variables with a matrix of covariances Σ. By definition Σ is a positive semi-definite matrix, and therefore it is possible to perform an eigen decomposition according to

$$\Sigma = \mathbf{P}\Lambda\mathbf{P}^t \tag{1}$$

where Λ is the diagonal matrix of eigenvalues of Σ and $\mathbf{P}$ is an orthogonal projection called loading matrix, whose k_{th} column is the k_{th} eigenvector of Σ. We have directly that $\mathbf{P}^t\Sigma\mathbf{P} = \mathbf{\Lambda}$ (for a

comprehensive text regarding principal component analysis see Jolliffe [5]. Usually the elements of $\mathbf{\Lambda}$, the eigenvalues, are in descending order, so that

$$\Lambda = diag(\lambda_1, \ldots, \lambda_p) \quad with \quad \lambda_1 \geq \lambda_2 \ldots \geq \lambda_p \geq 0 \tag{2}$$

We are concerned with a spatial study where the standard PCA is replaced with a spatial variant approach. In these situations, the vector of variables $\mathbf{x}$ has a spatial location, given by $\mathbf{s} = (s_1, s_2)$ (two-dimensional locations, for example) [6, 3]. In this case, the spatial data may not be well described by a global model of PCs but there are localized regions in the attribute data space where a suitably localized set of PCs provide a better description (local models). That is, in different parts of the data space, a different set of PCs is needed. This heterogeneity can be modelled by a covariance matrix depending on the spatial positions. More explicitly, the covariance matrix at $\mathbf{s}$ can be expressed as

$$\Sigma(\mathbf{s}) = \mathbf{P}(\mathbf{s})\Lambda(\mathbf{s})\mathbf{P}(\mathbf{s})^t \tag{3}$$

where $\Lambda(\mathbf{s})$ is the diagonal matrix of eigenvalues of $\Sigma(\mathbf{s})$, and $\mathbf{P}(\mathbf{s})$ is an orthogonal projection matrix verifying $\mathbf{P}(\mathbf{s})^t\Sigma(\mathbf{s})\mathbf{P}(\mathbf{s}) = \Lambda(\mathbf{s})$.

2. NON PARAMETRIC ESTIMATION ALGORITHM

Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of $\mathbf{x}$ with associate spatial positions $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the estimated variance matrix $\hat{\Sigma}$ in the spatial position (s_1, s_2) is obtained as follows:

$$\hat{\Sigma}(s_1, s_2) = \mathbf{X^t W}(s_1, s_2)\mathbf{X} \tag{4}$$

where $\mathbf{X}$ is the data matrix with n rows representing the observation and p columns representing the variables. We assumed that the columns of $\mathbf{X}$ have been standardised with zero mean and unit variance, i.e, PCA is based on correlation matrix. Moreover, $\mathbf{W}(s_1, s_2) = diag\{W_1(s_1, s_2), \ldots, W_n(s_1, s_2)\}$ is a diagonal matrix of geographic weights. Finally, the estimated matrices $\hat{\mathbf{P}}(s_1, s_2)$ and $\hat{\Lambda}(\mathbf{s})$ originating the eigen decomposition

$$\hat{\Sigma}(\mathbf{s}) = \hat{\mathbf{P}}(s_1, s_2)\hat{\Lambda}(s_1, s_2)\hat{\mathbf{P}}^t(s_1, s_2) \tag{5}$$

are obtained as in the standard PCA. Note that $W_i(s_1, s_2)$ is a weighted function that contains a smoothing positive parameter h. An interesting discussion on density and regression kernel estimation can be found in Wand [9].

2.1 BANDWIDTH SELECTION

Following Harris [3], we used the cross-validation automatic selection bandwidth based on the approximations obtained with the first q principal components. Denoting $\mathbf{P}_q$ as the matrix containing the first q columns of $\mathbf{P}$ then the corresponding scores are given by $\mathbf{Z}_q = \mathbf{XP_q}$ and the proportion of the total variance explained is $\sum_{i=1}^{q} \lambda_i / trace(\Lambda)$. Therefore an approximation of $\mathbf{X}$ is given by $\hat{\mathbf{X}} = \mathbf{X}\mathbf{P}_q\mathbf{P}_q^t$. In a context of local PCA, the bandwidth h can be selected by means of cross-validation, that is, minimizing

$$CV(h) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i^{(-i)}|| \tag{6}$$

with $\hat{\mathbf{x}}_i^{(-i)} = \hat{\mathbf{P}}_{q,(-i)}(\mathbf{s}_i)\hat{\mathbf{P}}^t_{q,(-i)}(\mathbf{s}_i)\mathbf{x}_i$ and $\hat{\mathbf{P}}_{q,(-i)}(\mathbf{s}_i)$ the matrix containing the first q columns of $\hat{\mathbf{P}}(\mathbf{s}_i)$ leaving out the i-sample data $(\mathbf{s}_i, \mathbf{x}_i)$.

3. TESTING SPATIAL STRUCTURE

In order to justify the use of GWPCA, we shall consider the development of a hypothesis test for

$$H_0 : \Sigma(\mathbf{s}) = \Sigma \quad \forall \mathbf{s} \quad against \quad H_1 : \Sigma(\mathbf{s}) \ not\ all\ the\ same\ \forall \mathbf{s} \tag{7}$$

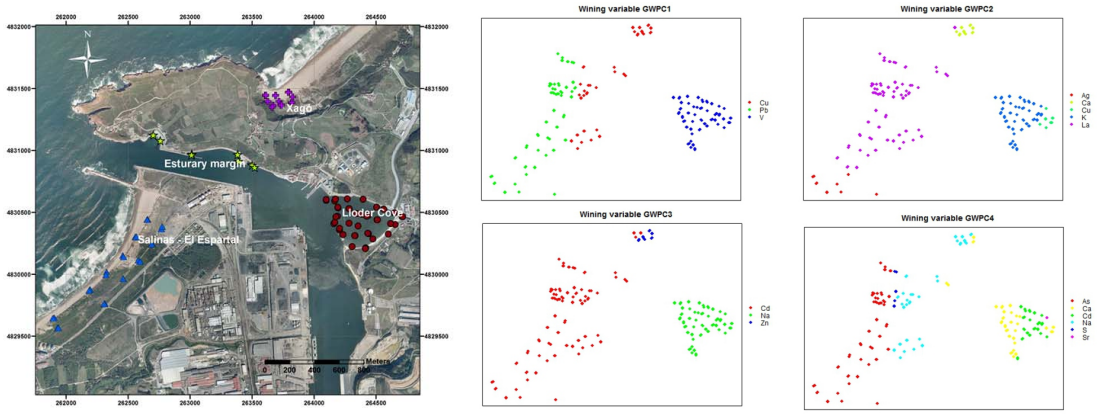


Figura 1: Left: Location of the samples overlaid on an orthophotography of the study area (Coordinate system: ETRS1989 UTM Zone 30N). Right: Winning variables for the first four geographically weighted components (GWPCs).

An Omnibus test to detect any departure from the null hypothesis would be desirable. A viable alternative, broad-spectrum test is given in Harris [3]. The statistic proposed is given by

$$T_1 = n^{-1} \sum_{i=1}^{n} \left(\hat{\lambda}_1(\mathbf{s}_i) - \bar{\lambda}_1 \right)^2 \tag{8}$$

where $\hat{\lambda}_1(\mathbf{s}_i)$ is the first eigenvalue of the matrix $\hat{\Lambda}(\mathbf{s}_i)$, and $\bar{\lambda} = n^{-1} \sum_{i=1}^{n} \hat{\lambda}_1(\mathbf{s}_i)$.
In this paper, we propose an alternative statistic based on eigenvectors instead of eigenvalues. We consider the transformed data $\mathbf{z}_1, \ldots, \mathbf{z}_n$ with $\mathbf{z}_i = \mathbf{P^t x}_i$ being $\mathbf{P}$ the projection matrix of $\mathbf{X}$ in a standard PCA. Denoting the covariance matrix of $\mathbf{z}$ as $\Sigma_{\mathbf{z}}(\mathbf{s})$ the following spectral decomposition is obtained

$$\Sigma_{\mathbf{z}}(\mathbf{s}) = \mathbf{P}_{\mathbf{z}}(\mathbf{s}) \Lambda_{\mathbf{z}}(\mathbf{s}) \mathbf{P}_{\mathbf{z}}^{t}(\mathbf{s}) \tag{9}$$

with $\mathbf{P}_{\mathbf{z}}(\mathbf{s})$ an orthogonal projection and $\Lambda_{\mathbf{z}}(\mathbf{s})$ diagonal matrix of corresponding eigenvalues. And, the proposed statistic is

$$T_2 = n^{-1} \sum_{i=1}^{n} \left(\hat{\mathbf{P}}_{\mathbf{z}}^{(1,1)}(\mathbf{s}_i) - 1 \right)^2 \tag{10}$$

where $\hat{\mathbf{P}}_{\mathbf{z}}^{(1,1)}(\mathbf{s}_i)$ represent the $(1,1)$th element of the matrix $\hat{\mathbf{P}}_{\mathbf{z}}(\mathbf{s}_i)$.
For a given significance α, the decision rule based on T consists of rejecting the null hypothesis if $T > T^{\alpha}$, where T^{α} represents the $(1-\alpha)$-percentile of T. This percentiles were obtained using Monte Carlo techniques [7, 8]: i.e. randomly re-assigning measurements to spatial locations. The power of this hypothesis test obtained with this procedure using statistic T_1, in our simulation study, is quite poor, at least in comparison with the power obtained with the proposed statistic T_2.

4. APPLICATION TO THE STUDY OF SOIL CONTAMINATION IN AVILÉS

GWPCA was applied to study the distribution of soil pollutants in a industrialized area located in Avilés (Northern Spain) represented in Figure 1 (left). Specifically, Cu, Pb, Zn, Ag, Ni, Co, Mn, Fe, As, Sr, Cd, V, Ca, La, Cr, Mg, Al, Na, K, S were the pollutants studied.
The statistical test proposed indicates the existence of spatial heterogeneity for a local vecinity below 0,30 of the whole study area.
Figure 1 (right) shows the spatial distribution of the winning variables (those with the highest absolute local loading) for the first four geographically weighted components analyzed. The winning variable in GWPC1 seems to be related with the presence of a plant of Zn. GWPC2 can be associated to lithogenic elements such as Al, K and La, reflecting the siliciclastic materials from the geology of the area. GWPC3 represents elements that are not present in the lithology of the zone, such as Na, an element of non-lithogenic origin. Finally, GWPC4 represents a group of mixed nature in which anthropogenic but mainly natural elements are present.

REFERENCIAS

Referencias

[1] Tipping,M.E., and Bishop, Ch.M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B* 61(3), 611–622.

[2] Charlton, M. Brunsdon, C., Demsar, U., Harris, P., Fotheringham, A.S. (2011). Principal Components Analysis: from Global to Local. *13th AGILE International Conference on Geographic Information Science*, Guimaraes, Portugal.

[3] Harris P., Brunsdon C., Charlton M. (2011). Geographically weighted principal component analysis. *International Journal of Geographical Information Science*, 25(10), 1717–1736.

[4] Gollini, I., Lu, B., Charlton, M., Brunsdon, Ch., Harris, P. (2015). GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models. *Journal of Statistical Software* 63(17), 1–50.

[5] Jolliffe, I.T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag New York.

[6] Demsar U., Harris P., Brunsdon Ch., Fotheringham A., McLoone A. (2013). Principal Component Analysis on Spatial Data: An Overview. *Ann. Assoc. Am. Geogr.*, 103 (1), 106—128.

[7] Brunsdon, C., Fotheringham, A.S., Charlton, M.E. (1998). Geographically weighted regression modelling spatial non-stationarity. *The Statistician,* 47 (3), 431–443.

[8] Hope, A. (2002). A Simplified Monte Carlo Significance Test Procedure. J. R. Stat. Soc. Series B (Methodological), 30(3), 582–598.

[9] Wand M.P., Jones M.C. (1995). *Kernel Smoothing*. Champan & Hall.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

STRUCTURAL TIME SERIES MODELING: AN APPLICATION TO ENVIRONMENTAL VARIABLES

Marco Costa[1], Olexandr Baturin[2] and A. Manuela Gonçalves[3]

[1]CIDMA-Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal (marco@ua.pt)
[2]DMA-Department of Mathematics and Applications, University of Minho, Portugal (olexandr.baturin@gmail.com)
[3]CMAT-Centre of Mathematics, DMA-Department of Mathematics and Applications, University of Minho, Portugal (mneves@math.uminho.pt)

ABSTRACT

A structural time series model is one which is set up in terms of components which have a direct interpretation. In this paper, the discussion focuses on the dynamic modeling procedure based on the state space approach (associated to the Kalman filter), in the context of surface water quality monitoring, in order to analyze and evaluate the temporal evolution of the environmental variables, and thus identify trends or possible changes in water quality (change point detection). The approach is applied to environmental time series: time series of surface water quality variables in a river basin. The statistical modeling procedure is applied to monthly values of physico-chemical variables measured in a network of 8 water monitoring sites over a 15-year period (1999-2014) in the River Ave hydrological basin located in the Northwest region of Portugal.

Keywords: River Ave, Water quality variables, Structure time series models, State space models, Kalman filter, Change point detection.

1. INTRODUCTION

Structural time series models are formulated directly in terms of unobserved components, such as trends, cycles and seasonals, that have a natural interpretation and represent the salient features of the series under investigation. The key to handling structural time series models is the state space form with the state of the system representing the various unobserved components. State space models provide a very flexible yet fairly simple tool for analyzing dynamic phenomena and evolving systems, and have significantly contributed to extend the classic domains of application of statistical time series analysis to non-stationary, irregular processes, etc. A structural model can therefore not only provide forecasts, but also, through estimates of the components, present a set of stylised facts and this formulation will allow making some useful interpretations [3] and [4]. In this study, it is proposed a dynamic modeling procedure based on the state space approach (associated to the Kalman filter) in time series of water quality variables. The data concerns the River Ave hydrological basin located in the Northwest of Portugal, where monitoring has become

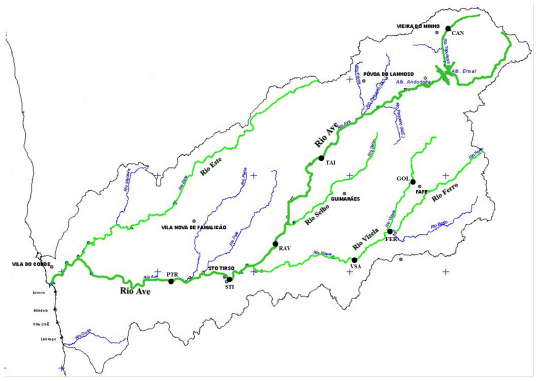


Figure 1: River Ave hydrological basin located in the Northwest of Portugal and its 8 water monitoring sites.

a priority in water quality planning and management because its water has been in a state of obvious environmental degradation for many years. As a result, the watershed is now monitored by eight sampling stations distributed along the River Ave and its main streams: PTR, STI, RAV, VSA, FER, GOL, TAI and CAN (Figure 1). In these water quality monitoring sites are held monthly measurements and analyses for a general assessment of basin's surface water quality (physico-chemical and microbiological variables).

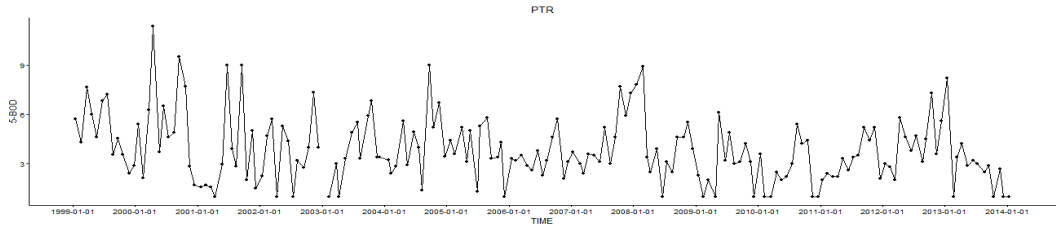


Figure 2: 5-BOD monthly concentration from January 1999 to January 2014 in Ponte Trofa (PTR) monitoring site.

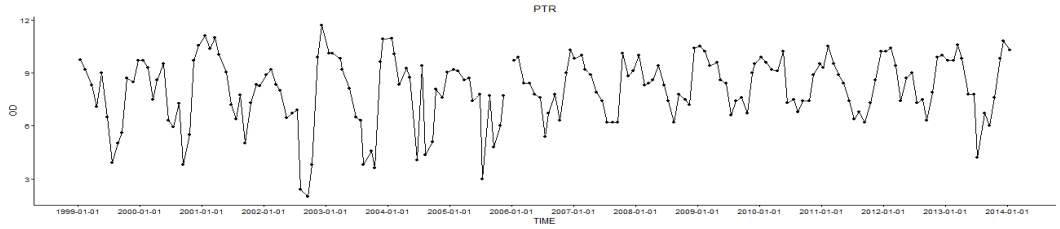


Figure 3: DO monthly concentration from January 1999 to January 2014 in Ponte Trofa (PTR) monitoring site.

For this study we selected the Dissolved Oxygen (OD) concentration and the 5-day Biological Oxygen Demand (5-BOD) due to their importance in the evaluation of this river's water quality, both measured in mg/l. Dissolved oxygen (OD) refers to the level of free, non-compound oxygen present in water or other liquids. It is an important variable in assessing water quality because of its influence on the organisms living within a body water. Dissolved oxygen (OD) can range from less than $1mg/l$ to more than $20mg/l$. Biochemical Oxygen Demand (5-BOD) refers to the amount of oxygen that would be consumed if all the organics in one liter of water were oxidized by bacteria and protozoa. Microorganisms such as bacterias are responsible for decomposing organic waste. When organic matter such as dead plants, sewage, etc. is present in a water supply, the bacteria will begin the process of breaking down this waste. Biochemical Oxygen Demand (BOD)

is a measure of the oxygen used by microorganisms to decompose this waste. The range of possible readings can vary considerably: water from an exceptionally clear water might show a 5-BOD of less than $2mg/l$ of water. In this paper we consider data series from one water monitoring site PTR (Ponte Trofa) on pretence of example of the modeling approach (Figures 2 and 3). For the modeling process we consider data between January 1999 to January 2014 ($n = 181$ observations).

2. MODELING APPROACH

An exploratory analysis was performed and showed that the data exhibit seasonal behavior (by analyzing the monthly averages) and a moderate temporal dependency [2] and [3]. The linearity of the possible trend that was suggested in [2] by a graphical analysis of the 12 times series corresponding to each month was found not suitable for the time series. However, we must explore a stochastic level as p-dimensional state space in the modeling. The parameters of the state space models must be estimated for each environmental series, and they are estimated by Gaussian maximum likelihood estimation in the state space framework. As the series of observations present intrinsic environmental data proprieties, the initial model is very versatile since it can accommodate several statistical properties often presented in environmental data, such as pth order polynomial component (a stochastic local level is considered), seasonality and temporal correlation. It is proposed an application of a structural time series model by taking into account this data structure. So, the monthly time series are modeled by equations

$$y_t = \mu_t + \gamma_t + \epsilon_t$$
$$\mu_{t+1} = \mu_t + \xi_t$$
$$\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} \ldots - \gamma_{11,t} + \omega_t$$
$$\gamma_{2,t+1} = \gamma_{1,t}, \gamma_{3,t+1} = \gamma_{2,t}, \ldots, \gamma_{11,t+1} = \gamma_{10,t}$$
$$\epsilon_t = \phi_1 \epsilon_{t-1} + \varepsilon_t$$

with

$$\xi_t \sim N(0, \sigma_\xi^2), \omega_t \sim N(0, \sigma_\omega^2) \quad \text{and} \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2).$$

The first equation represents the structure for the observed data, the observation equation, where y_t is the monthly environmental variable, in this case 5-BOD and OD variables with $t = 1, 2, ..., 181$. The observations are driven by two state variables, one being the level variable and the second being the seasonality variable. The second equation is a state space equation representing a stochastic level as an 1-dimensional state space. The seasonal behavior is represented by γ_t, an 11-dimensional state space vector of seasonal deviations. And to deal with the temporal correlation structure it is considered ϵ_t following a 1st order autoregressive process, AR(1). Since the model has a state space representation, it allows obtaining forecast or other predictions of interest (filtered or smoother predictions). Since the state process is unobserved, both forecast and filtered predictions are obtained through the Kalman filter algorithm.

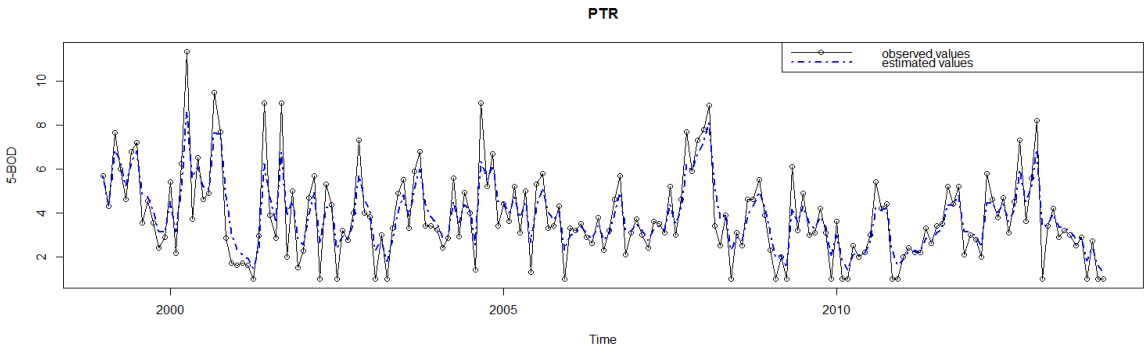


Figure 4: Observed values and filtered state estimates of 5-BOD in PTR monitoring site.

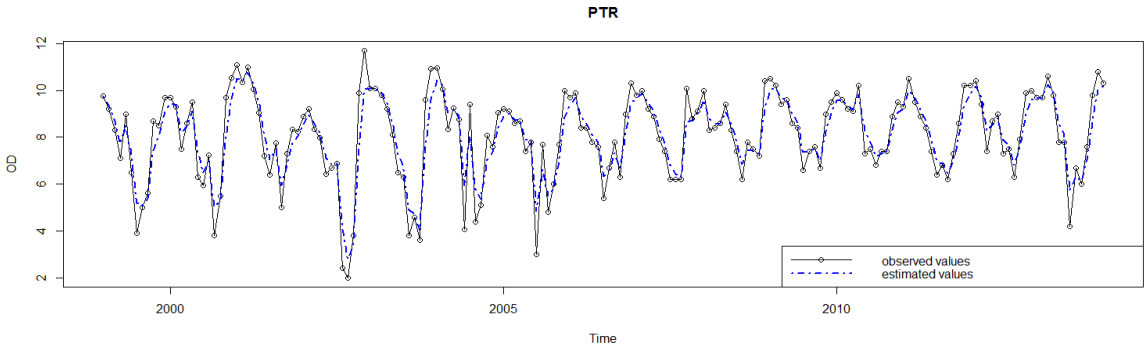


Figure 5: Observed values and filtered state estimates of OD in PTR monitoring site.

The referred methodology was applied in R environment with dlm package (http://www.r-project.org/) [5] and [6]. A first analysis of adjustment models is done by comparing data and predictions. Figures 4 and 5 show observed data and the filtered predictions in Ponte Trofa (PTR) for 5-BOD and OD.

3. CONCLUSIONS

The analysis performed in this study shows that the structural time series models (the state space models associated to the Kalman filter) are suitable to model 5-BOD and OD concentration series at the PTR water monitoring site and they allow to obtain pertinent findings concerning water surface quality interpretation and of change point of view [1] and [4], thus highlighting the potential value of this type of analysis. The next step is to analyze the filtered and smoothed predictions (forecasts) of states given by the Kalman filter, which allows an interesting analysis of the structural components, and further research is in progress to improve the modeling process and the results obtained.

ACKNOWLEDGEMENTS

Marco Costa was supported by Portuguese funds through the CIDMA-Centre for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology "FCT-Fundação para a Ciência e a Tecnologia", within project UID/MAT/04106/2013. A. Manuela Gonçalves was supported by the Research Centre of Mathematics of the University of Minho with the Portuguese Funds from the "FCT-Fundação para a Ciência e a Tecnologia", through the Project PEstOE/MAT/UI0013/2014.

References

[1] Costa M., Gonçalves A.M. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk assessment* 25(2):151–163.

[2] Gonçalves A.M., Alpuim T. (2011). Water Quality Monitoring using cluster analysis and linear models. *Environmetrics* 22(8), 933–945.

[3] Gonçalves A.M., Costa M. (2013). Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. *Stochastic Environmental Research and Risk assessment* 127(5):1021–1038.

[4] Gonçalves A.M., Costa M. (2011). Application of Change-Point Detection to a Structural Component of Water Quality Variables. In Simos, T.E., Psihoyios, G., Tsitouras, Ch. (eds.): *AIP Conference Proceedings* Vol 1389, 1565–1568, AIP American Institute of Physics Publishing.

[5] Petris, G., Petrone, S., Campagnoli, P. (2009). *Dynamic Linear Models with R.* Springer, New York.

[6] R Team, D. (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

UNA APROXIMACIÓN PARA DETERMINAR GRUPOS DE MÚLTIPLES CURVAS REGRESIÓN. APLICACIÓN AL ESTUDIO DE LA GEOMETRÍA DE UN TÚNEL

Nora M. Villanueva[1,3], Marta Sestelo[2,3], Celestino Ordóñez[4] y Javier Roca-Pardiñas[1,3]

[1] Departamento de Estadística e I. O., Universidade de Vigo, España.

[2] Centro de Matematica, Universidade do Minho, Portugal.

[3] Grupo de investigación SiDOR y CINBIO, Universidade de Vigo, España.

[4] Departamento de Prospección y Explotación Minera, Universidad de Oviedo, España.

RESUMEN

En muchas situaciones prácticas puede resultar interesante realizar grupos de curvas de regresión, sobretodo en situaciones donde existe un elevado número de ellas. La aproximación que se propone en este trabajo permite determinar grupos a través de una selección automática de su número. El procedimiento propuesto consiste en: (i) estimar las curvas de regresión mediante suavizadores tipo kernel, (ii) dado un número de grupos K, asignar de manera óptima las curvas a los K grupos correspondientes mediante el uso de un algoritmo heurístico, (iii) determinar el número óptimo de grupos utilizando un procedimiento automático basado en técnicas bootstrap. Con el fin de determinar la validez de la metología y compararla con otras aproximaciones existentes se han realizado distintos estudios de simulación.
La metodología propuesta se ha utilizado para estudiar la geometría de un túnel mediante el análisis de un conjunto de secciones del mismo. Estas secciones se obtuvieron ajustando una superficie a una nube de puntos medida con un escáner láser terrestre. La existencia de diferentes grupos de secciones indica que el túnel tiene una geometría heterogénea. Dicha heterogeneidad podría ser debida a una mala construcción o a deformaciones posteriores.

Palabras e frases clave: Múltiples curvas de regresión, número de grupos, suavizadores kernel, geometría de un túnel.

AGRADECIMIENTOS

Los autores agradecen el apoyo financiero del Ministerio de Ciencia e Innovación de España (MTM2011-23204, fondos FEDER incluidos) y de la Xunta de Galicia (CN2012/180). Además, este trabajo trabajo ha sido financiado por el proyecto BIA2011-26915 del Ministerio de Ciencia e Innovación de España, por la beca SFRH/BPD/68328/2010 de la Fundación de Ciencia y Tecnología Portuguesa (FCT), por fondos FEDER a través del "Programa Operacional Factores de Competitividade - COMPETE" y por fondos nacionales de Portugal a través de la FCT en el ámbito del proyecto PEst-OE/MAT/UI0013/2014.

Referencias

[1] Macqueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

[2] Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society Series B* 63(Part 2), 411–423.

[3] Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall.

[4] Wu, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics* 14(4), 1261–1295.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

USO DE MÉTODOS BIOMÉTRICOS EN TAXONOMÍA VEGETAL: EL CASO DE LA SUBTRIBU ANTHOXANTHINAE (POACEAE)

Lema-Suárez I.[1], Sahuquillo E.[1], Estévez G.[2] y Pimentel M.[1]

[1] Departamento de Bioloxía Animal, Bioloxía Vexetal e Ecoloxía. Universidade da Coruña
[2] Departamento de Matemáticas. Universidade da Coruña

RESUMEN

En este estudio se analizan 26 características macro y micromorfológicas en 31 especies de gramíneas pertenecientes a la Subtribu Anthoxanthinae. Se aplica métodos de análisis multivariantes para seleccionar las características fenotípicas que permitan clarificar las relaciones taxonómicas entre los representantes de este complejo grupo de plantas.

Palabras y frases clave: *Anthoxanthum*, *Hierochloe*, morfometría, taxonomía numérica

1. INTRODUCCIÓN

El estudio de los procesos evolutivos que determinan la aparición de nuevas especies implica la integración de diversas herramientas que permitan aproximaciones complementarias en el análisis de la biodiversidad [1]. La subtribu Anthoxanthinae está constituida por dos géneros (*Anthoxanthum* e *Hierochloe*) que incluyen complejos poliploides de origen reciente y difícil diferenciación morfológica [2, 3]. Tradicionalmente, la diferenciación entre ambos géneros se realizaba en base al número de flores fértiles presentes en las espiguillas, pero el descubrimiento de nuevas especies que presentan características morfológicas intermedias entre ambos ha llevado a algunos autores a cuestionarse su separación taxonómica [4]. Los estudios moleculares pusieron de manifiesto que esta subtribu forma un grupo monofilético, pero tanto el número de géneros como de especies que la constituyen sigue siendo fuente de controversias [5].

El objetivo de este estudio es aplicar métodos estadísticos en el análisis de características morfológicas y anatómicas de uso habitual en la taxonomía de la subtribu Anthoxanthinae. Los métodos utilizados han sido empleados previamente en sistemática de angiospermas con resultados satisfactorios [6, 7].

2. MATERIAL Y MÉTODOS

Se examinaron muestras de 31 especies de la subtribu Anthoxanthinae (10 del género *Anthoxanthum*, 17 del género *Hierochloe* y 4 de la sección *Ataxia*, cuya posición en la subtribu es incierta). En total, se analizaron 920 especímenes en los estudios morfológicos y 348 en los anatómicos.

Datos macro y micromorfológicos

Se midieron 15 caracteres macromorfológicos, 8 cuantitativos (transformados logarítmicamente) y 7 cualitativos (variables binarias). Para estandarizar la medición de los datos foliares, ésta se realizó en la segunda hoja basal. Las características referidas a las flores e inflorescencias se tomaron siempre en las espiguillas basales. Para la toma de los datos micromorfológicos se siguió el protocolo descrito por Devesa [7] y Pimentel & Sahuquillo [2]. Todos los caracteres anatómicos fueron tratados como variables binarias indicando la presencia/ausencia de dicho carácter para cada espécimen examinado.

Análisis estadísticos

El análisis de los caracteres morfológicos cuantitativos se planteó a dos niveles taxonómicos: géneros y especies (aplicado solo al género *Hierochloe*). Tras la realización de un análisis exploratorio de los distintos caracteres se aplicó el test ANOVA y la prueba de Tuckey (como método de comparaciones

múltiples) para seleccionar las características responsables de las diferencias entre las especies/géneros. Para el análisis conjunto de los 8 caracteres fue realizado un análisis de componentes principales (PCA) para resumir los patrones de variación y obtener los componentes principales no correlacionados que mejor explican la variabilidad observada. Con los resultados del PCA se realizó un análisis discriminante (CDA) para caracterizar las diferencias entre grupos (género/sección) y determinar aquel al que pertenecerían los especímenes con mayor probabilidad. Por último, también se realizó un análisis Cluster utilizando la distancia euclídea al cuadrado y el método de Ward.
Para el análisis de los datos cualitativos y anatómicos, se recurrió a técnicas de ordenación geométrica, el análisis de Coordenadas Principales (PCoA) y sobre todo, debido al elevado número de caracteres analizados, Escalado Multidimensional No Métrico (NMDS) para representar la similitud entre individuos y realizar un agrupamiento de los mismos**.**

3. RESULTADOS Y DISCUSIÓN

3.1. Diferenciación entre géneros de la subtribu *Anthoxanthinae*

Las características morfológicas analizadas presentan en su mayoría un gran solapamiento de valores entre los dos géneros y los representantes de la sección *Ataxia*. En general, las diferencias fueron significativas entre ambos géneros para cinco características: *longitud de la hoja, de la rama basal de la inflorescencia, de la espiguilla* y *del lema estéril* (Fig. 1). Solo una característica presentó diferencias significativas entre los tres grupos la *longitud del lema estéril*.

Como resultado del PCA se obtuvieron 3 componentes que explicaban el 85.25% de la varianza. En el primer componente las características que presentaban mayor peso fueron fundamentalmente las referidas a caracteres vegetativos, mientras que en el segundo y tercer componente se encontraban los caracteres florales. Al utilizar como variables para realizar el CDA los componentes obtenidos en el PCA y así evitar la correlación existente entre los caracteres analizados, se consiguió discriminar de forma clara los individuos pertenecientes a ambos géneros, *Anthoxanthum* con un 97% de los casos correctamente asignados e *Hierochloe* con un 83%, mientras que los individuos pertenecientes a la sección *Ataxia* se repartieron entre ambos géneros al 50%. El primer eje discriminante separa perfectamente el género *Anthoxanthum* de *Hierochloe* y los representantes de la sección *Ataxia* (Fig. 2).

El análisis cluster con los datos cuantitativos, que permitió formar 7 grupos, no consiguió definir grupos coherentes con la estructura de los géneros, apareciendo muy entremezclados sus representantes (Fig. 3 A). El 40% de los individuos pertenecientes al género *Hierochloe* se incluyen en el grupo 7; el 19,7%, en el 2, y el 10,9%, en el 3 ,y todos los de la sección *Ataxia*, en los grupos 2 y 3. Estos grupos se separan de la mayoría de los individuos del género *Anthoxanthum* por la combinación de tres caracteres: *longitud del lema, de la inflorescencia* y *de la hoja basal* (Fig. 3B).

Tanto los datos macromorfológicos cualitativos como anatómicos no presentaron valor discriminante entre géneros aunque sí se pudo observar que los representantes de la sección *Ataxia* se mezclan con los del género *Hierochloe*.

3.2. Diferenciación entre especies del género *Hierochloe*

Los caracteres analizados presentaron menor valor discriminante para las especies de *Hierochloe*. Solo dos características presentaron diferencias significativas entre las especies, una vegetativa, *talla,* y otra referida a la inflorescencia, *ángulo de la rama basal de la inflorescencia* (Fig. 4).

Los componentes extraídos mediante el PCA explican un 91,47% de la varianza. El primero incluye los siguientes caracteres vegetativos: *talla, longitud y anchura de la hoja y longitud de la rama basal de la inflorescencia* y *de la inflorescencia*. El segundo factor engloba solo caracteres florales: *longitud de la espiga, del lema estéril, de la gluma superior y gluma inferior y de la arista superior* e *inferior.* Al utilizar los factores obtenidos para realizar el CDA solo se obtiene una buena discriminación de 5 especies: *H. australis, H. glabra, H. gunkelii, H. pauciflora* y *H. pusilla* (Fig. 5). El resto de los individuos forman una masa de puntos continuos en la que es difícil establecer grupos discretos.

El dendrograma obtenido utilizando los 3 primeros componentes del ACP para realizar el análisis cluster pone de manifiesto una clara gradación entre las características morfológicas estudiadas en las 17 especies de *Hierochloe* (Fig. 6). Tres grupos de especies parecen bien definidos: el grupo 1, que incluye

las especies *H. pauciflora* (100%) e *H. pusilla* (100%) y presentan valores similares en *longitud de la hoja, de la espiguilla, del lema* y *de las glumas*; el grupo 2, con las especies *H. glabra* (84,6%) e *H. australis* (12,5%), relacionadas por *longitud de la inflorescencia, ángulo de la rama basal de la inflorescencia* y *longitud de las aristas*; el grupo 9, con *H. laxa* (14,3%) e *H. redolens* (21,7%), con valores similares en *anchura de la hoja* y *de la inflorescencia en 2/3, longitud de la rama basal de la inflorescencia* y de los caracteres relativos a la espiguilla; y el grupo 8, con *H. altissima* (81,2%), *H. redolens* (39,1%) e *H. utriculata* (40,0%), relacionados por *anchura de la hoja* y *longitud de la inflorescencia* y *del lema*.

3. CONCLUSIONES

- La caracterización morfológica entre *Anthoxanthum* e *Hierochloe* es compleja y se debe basar en una combinación de caracteres tanto vegetativos como florales.
- La sección *Ataxia* se encuentra morfológicamente más próxima al género *Hierochloe* que al género *Anthoxanthum*.
- Solo 5 especies de las 17 analizadas del género *Hierochloe* se pueden diferenciar mediante los caracteres estudiados. El resto de las especies forman una masa continua de puntos en las que es difícil establecer grupos discretos.

REFERENCIAS

[1] Duminil, J., Kenfack, D., Viscosi, V., Grumiau, L., Hardy, O.J. (2012). Testing species delimitation in sympatric species complexes: the case of an African tropical tree, *Carapa* spp. (Meliaceae). *Molecular Phylogenetics and Evolution* 62, 275-285.

[2] Pimentel, M., Estévez, G., Sahuquillo, E. (2007). European Sweet Vernal Grasses (*Anthoxanthum*: Poaceae, Pooideae, Aveneae): A Morphometric Taxonomical Approach. *Systematic Botany* 32(1), 43-59.

[3] Weimar, G. (1971). Variation and Taxonomy of *Hierochloe* (Gramineae) in the Northern hemisphere. *Botaniska Notiser* 124, 129-175

[4] Schouten, Y., Veldkamp, J.F. (1985). A revision of *Anthoxanthum* including *Hierochloe* (Gramineae) in Malesia and Thailand. *Blumea* 30, 319-351.

[5]Pimentel, M., Sahuquillo, E., Torrecilla, Z., Popp, M., Catalán, P., Brochmann, C. (2013). Hybridizations and long-distance colonization at different time scales: towards resolution of longterm controversies in the sweet vernal grasses (Anthoxanthum). *Annals of Botany* 112, 1015-1030.

[6] Baum, B.R. (1986). Numerical taxonomic analyses of the Poaceae. In Soderstrom, T.R., Hilu, K.W., Campbell, C.S., Barkworth, M.E. (eds.): *Grass systematics and evolution* 334-342, Smithsonian Institution Press, Washington, D.C.

[7] Devesa, J.A. (1992). *Anatomía foliar y palinología de las gramíneas extremeñas*. Universidad de Extremadura, Badajoz.

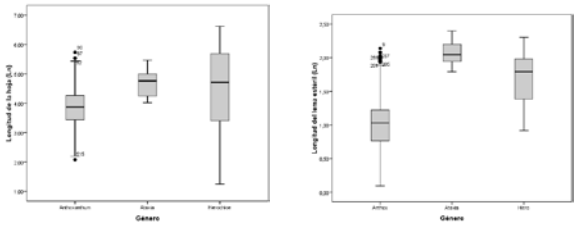


Figura 1: Box-plots de *longitud de la hoja* y *longitud del lema estéril*.

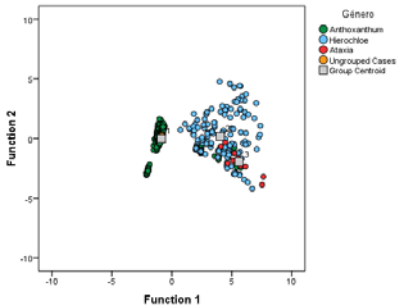
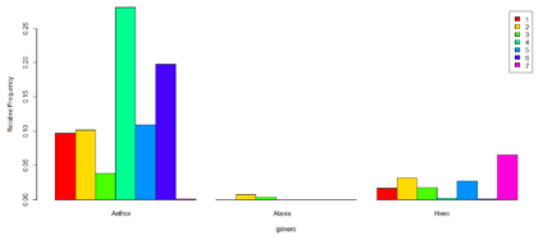


Figura 2: Gráfico CDA realizado con los tres primeros componentes obtenidos en el PCA.

A

Column percentages:

género			
hclus.Ward.7	Anthox	Ataxia	Hiero
1	11.8	0	10.2
2	12.3	70	19.7
3	4.7	30	10.9
4	34.0	0	1.4
5	13.2	0	17.0
6	23.9	0	0.7
7	0.1	0	40.1
Total	100.0	100	100.0
Count	748.0	10	147.0



B

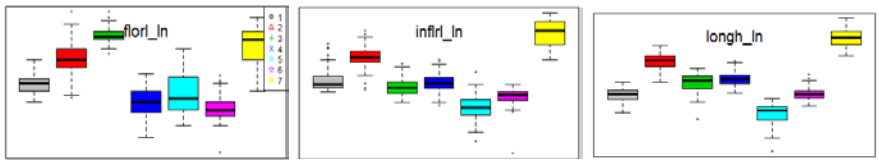


Figura 3: Gráficos del cluster entre géneros.

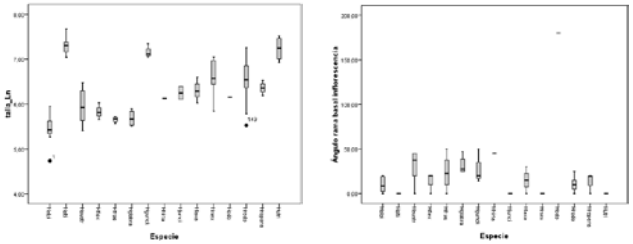


Figura 4: Box-plot de *talla* (izda.) y *ángulo de la rama basal de la inflorescencia* (dcha.).

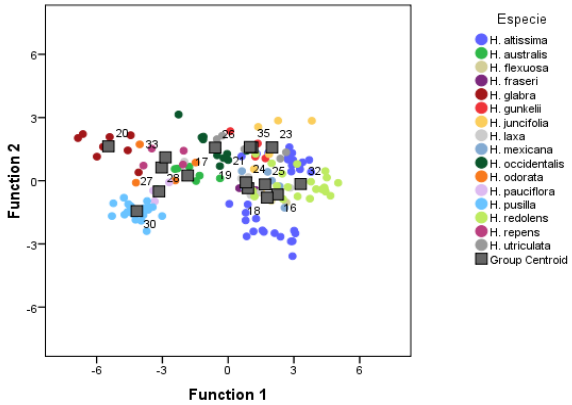


Figura 5: Gráfico CDA realizado con los tres primeros componentes obtenidos en el PCA.

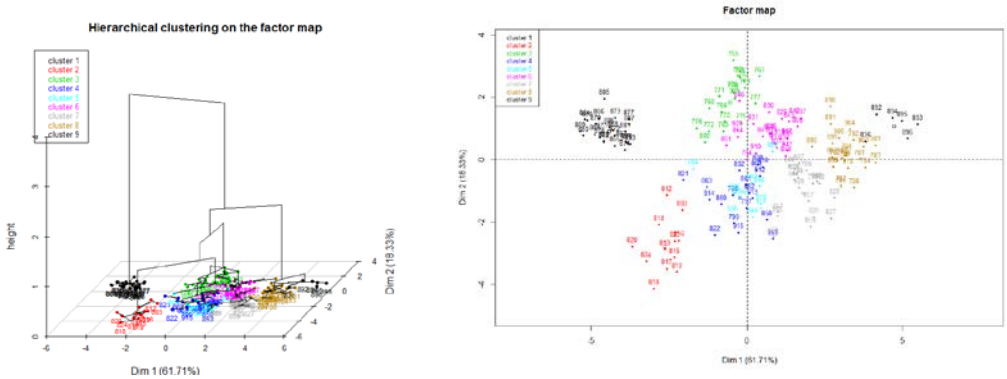


Figura 6: Gráficos del clúster entre especies.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

Técnicas no paramétricas en el riesgo sísmico: estimación de probabilidades de magnitudes de terremotos en el sur de la Península Ibérica

Quintela-del-Río, Alejandro 1[1] y Francisco-Fernández, Mario 2[2]

[1] Departamento de Matemáticas
Universidade da Coruña
15071 A Coruña, España
E-mail: aquintela@udc.es

[2] Departamento de Matemáticas
Universidade da Coruña
15071 A Coruña, España
E-mail: mariofr@udc.es

RESUMEN

En Francisco-Fernández y Quintela-del-Río (2011) se utilizaron técnicas de estimación no paramétrica para el estudio de parámetros de riego sísmico característicos de un área geográfica. En este trabajo, usaremos la técnica desarrollada por Choi y Hall (1999) para la construcción de los clusters de aquellos datos que están relacionados entre sí de manera espacial. Posteriormente, construiremos un mapa de riesgo, o de probabilidad de excedencia de la magnitud máxima anual, en la zona del sur de España (Andalucía y norte de Marruecos).

Palabras y frases clave: estimación no paramétrica, riesgo sísmico, data-sharpening, clustering.

1. INTRODUCCIÓN

La utilización de técnicas de estimación no paramétrica en el análisis de datos sísmicos ha adquirido una popularidad notable en la última decada. En el presente trabajo, utilizamos técnicas desarrolladas en Francisco-Fernández y Quintela-del-Río (2011) para la construcción de mapas de riesgo sísmico en una zona geográfica. En primer lugar, haremos uso de técnicas de data-sharpening (Choi y Hall, 1999) para la selección de aquellos datos que puedan ser utilizados como centros de los correspondientes clusters (datos dependientes desde el punto de vista espacial). A continuación, estos centros se usaran para la construcción de una grid que nos permita caracterizar su entorno en función de la mayor o menor probabilidad de ocurrencia de eventos de una magnitud elevada, creando así un mapa de riesgo sísmico. Para ello se utilizarán métodos de estimación no paramétricos de la fución de distribución. En la sección 2 se hará un breve resumen de las técnicas de data sharpening y como serán aplicadas en este contexto. En la sección 3, se describirán brevemente los estimadores no paramétricos utilizados para el análisis de los datos transformados mediante data sharpening.

2. DATA SHARPENING

La técnica de *data sharpening*, introducida por Choi y Hall (1999), consiste en transformar los datos originales de una muestra para posteriormente construir un estimador de la función de densidad a partir de los datos transformados. Se comprueba que, de esta forma, el estimador de la densidad obtenido tiene un sesgo menor que el que se obtiene con los datos originales. Esta técnica puede utilizarse con distintos tipos de estimadores no paramétricos. El interés de la misma para el estudio de datos de terremotos consiste en la posibilidad de transformar los datos originales en otros más cercanos a los centros espaciales donde se originan los terremotos de mayor magnitud (centros de los cluster de terremotos). A partir de estos datos transformados, se pueden obtener estimadores más precisos de características como la probabilidad de excedencia o el intervalo de recurrencia de terremotos de gran magnitud, entre otros. El método desarrollado es como sigue:

1. Cada dato se mueve según un promedio ponderado de las observaciones en un entorno del mismo.
2. La distancia vendrá determinada por el producto de densidades Gausianas, con diferente ponderación en cada dimensión (latitud, longitud y/o profundidad).
3. Cada dato x_i en la muestra de tamaño n se traslada hacia $x_i^\star$ utilizando la siguiente fórmula:

$$x_i^\star = \frac{\sum_{j=1}^n K_h(x_i - x_j)x_j}{\sum_{j=1}^n K_h(x_i - x_j)}, \text{ para } i = 1, 2, \ldots, n$$

3. MÉTODOS NO PARAMÉTRICOS

Sea Y una variable aleatoria continua, con función de densidad f y función de distribución F. Dada una muestra de datos $Y_1, \ldots, Y_n$ de Y, es posible construir un estimador no paramétrico de tipo núcleo (kernel) para estimar la función de distribución mediante la fórmula:

$$F_h(y) = P(Y \leq y) = \int_{-\infty}^{y} f_h(t)dt = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{y - Y_i}{h}\right),$$

donde $H(u) = \int_{-\infty}^{u} K(t)dt$, siendo K la función núcleo y h el parámetro de suavización o ventana. Ahora, supongamos que disponemos de una base de datos de terremotos, y sean $m_1, \ldots, m_n$ sus magnitudes. Denotemos por X la variable que mide el número de terremotos por año. A partir de un estimador no paramétrico de la función de distribución, puede construirse un estimador del riesgo sísmico mediante (Quintela-del-Río, 2010):

$$\hat{G}(y) = \sum_{k=o}^{\infty} \left(\hat{F}_M(y)\right)^k f_k, \tag{1}$$

donde denotamos por f_k la frecuencia observada de la variable X (número de años con k terremotos, dividido por el total de eventos n). En la práctica, la suma (1) será finita, hasta un entero n_0, de modo que $f_k = 0, \forall k \geq n_0$. Por consiguiente, el estimador adopta la forma

$$\hat{R}_h(\cdot) = 1 - \hat{G}(\cdot).$$

4. ANÁLISIS DE DATOS

La zona seleccionada para su estudio (Sur de España) es una zona de interacción de fallas entre Europa y África. Está limitada por las coordenadas 35.5°N - 38°N y 0°W - 10°W. Corresponde a la zona más importante de España desde el punto de vista de número de terremotos y de su magnitud. Para nuestro análisis, hemos seleccionado los datos disponibles a través de la página web del Instituto Geográfico Nacional (IGN), cuya dirección es *http://www.ign.es/ign/es/IGN/SisCatalogo.jsp* (Figure 1).

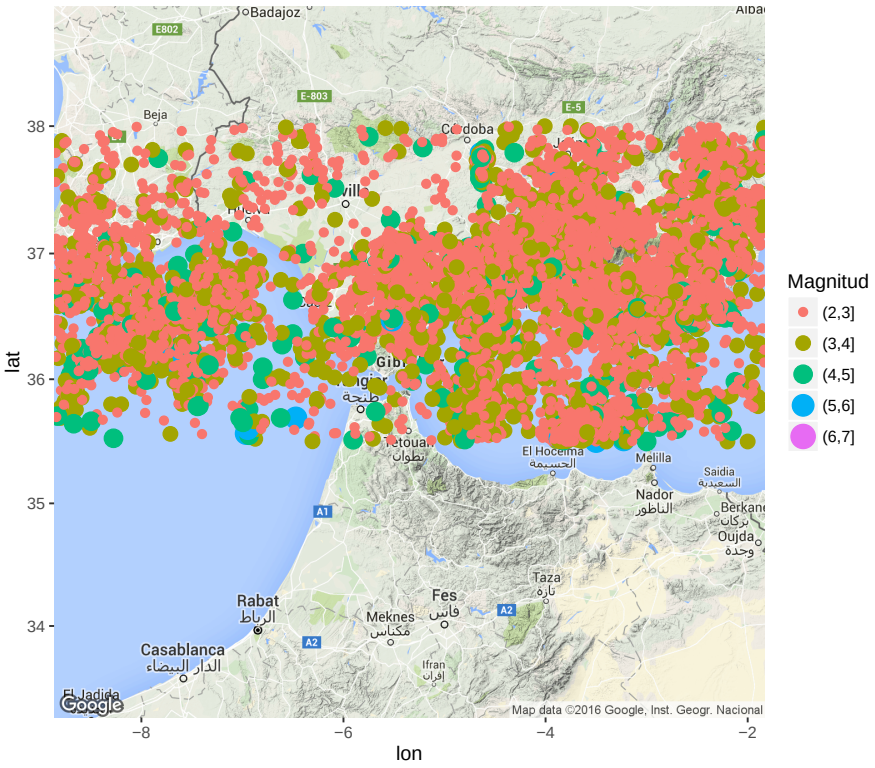


Figura 1: Mapa de la zona geográfica donde se analizará el riesgo sísmico, junto con los datos utilizados.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto del Ministerio de Economía y Competitividad (MINECO) MTM2014-52876-R.

Referencias

[1] Choi, E., Hall, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika* 86, 941–947.

[2] Francisco-Fernández, M., Quintela-del-Río, A. (2011). Nonparametric seismic hazard estimation: A spatio-temporal application to the northwest of the Iberian Peninsula. *Tectonophysics* 505, 35–43.

[3] Quintela-del-Río, A. (2010). On non-parametric techniques for area-characteristic seismic hazard parameters. *Geophysical Journal International* 180, 339—346.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

EVOLUCIÓN ESPACIAL DOS SUICIDIOS EN GALICIA

G Naveira Barbeito[1], MI Santiago Pérez[1] e A Malvar Pintos[1]

[1] Servizo de epidemioloxía, Dirección Xeral de Saúde Pública, Xunta de Galicia

RESUMO

Durante os últimos anos a mortalidade por suicidio en Galicia sitúase como primeira causa de morte violenta. O obxectivo proposto neste traballo foi estudar o patrón xeográfico da mortalidade por esta causa en Galicia durante os quinquenios 2000-04, 2005-09 e 2010-14 e analizar a súa tendencia temporal. Para isto aplicáronse modelos xerárquicos bayesianos baseados na aproximación por integradores de Laplace aniñados (INLA). Como resultado do axuste dos modelos, apréciase un claro patrón norte-sur no risco de morte por suicidio, sendo os concellos da metade norte os que presentan un maior risco mentres que os da metade sur presentan un risco menor. A tendencia temporal, de forma xeral, foi lixeiramente crecente ao longo dos quinquenios.

Palabras e frases chave: Suicidios, mortalidade, modelo Besag-York-Mollié, INLA.

1. INTRODUCIÓN

O suicidio constitúe un importante problema de saúde pública, sendo a causa pola que máis persoas morren en Galicia dende o ano 2009, dentro das causas externas de defunción (como poden ser os accidentes de tráfico, as caídas, os homicidios,...).
O obxectivo proposto neste traballo foi estudar o patrón xeográfico (a nivel municipal) da mortalidade por suicidio en Galicia durante os quinquenios 2000-2004, 2005-2009 e 2010-2014, e analizar a súa tendencia temporal.
De cara a axustar as taxas de mortalidade por suicidio a nivel municipal é importante ter en conta o escaso número de defuncións que se producen en parte dos concellos de Galicia. Neste tipo de situacións o método máis adecuado no axuste é o método indirecto, xa que elude o problema da inestabilidade asociada ás taxas específicas por idade para cada concello empregando unha taxa estándar de referencia dunha poboación cun tamaño suficiente. Este axuste proporciona, como indicador, a razón de mortalidade estandarizada (RME) que se calcula como o cociente entre o número observado de defuncións ocorridas nun determinado concello e o número que cabería esperar nese concello, no suposto de que tivera as mesmas taxas específicas que a poboación de referencia considerada (neste estudo, a poboación de Galicia no período 2000-2014 en función do sexo). Por definición, este estimador non debe ser empregado para comparar distintos concellos entre si, senón para comparar cada un deles coa poboación de referencia [1], identificando os concellos que presentan un exceso ou defecto de risco en comparación con dita poboación de referencia.
Este indicador aínda presenta limitacións, xa que a representación directa nun mapa das RMEs pode dar lugar a unha mala interpretación do patrón xeográfico de mortalidade, debido a que aqueles concellos cun moderado tamaño de poboación dan lugar as RMEs menos precisas e máis extremas que predominarían no mapa [2]. Co fin de solucionar esta limitación, neste traballo suavizáronse as RMEs a través dun modelo xerárquico bayesiano, baseado na aproximación por integradores de Laplace aniñados (INLA) [3], que permite empregar a información de áreas veciñas para a estimación de cada concello. Este modelo foi proposto por Clayton e Kaldor [2] e, posteriormente, desenvolto por Besag, York e Mollie [4].

2. MATERIAL E MÉTODOS

O número de defuncións por suicidio e lesións autoinflixidas[1], ocorridas na poboación residente en Galicia durante os quinquenios 2000-2004, 2005-2009 e 2010-2014, obtívose do Rexistro Galego de Mortalidade. A poboación de Galicia para eses quinquenios obtívose do Instituto Galego de Estatística, a partir da información do Padrón Municipal de Habitantes. Esta información dispúxose por quinquenio, concello, sexo e por grupos de idade quinquenais.

Modelo espacial:
Para cada concello i (con i=1, 2,..., 314), asúmese que o número de defuncións observadas, O_i, no concello segue unha distribución de Poisson de parámetro $\lambda_i = E_i \cdot \rho_i$, onde E_i é o número de defuncións esperadas no concello e ρ_i é a RME do concello.
As RME foron modelizadas do seguinte xeito:

$$\log(\rho_i) = \alpha + u_i + v_i$$

onde α é a media do logaritmo para todos os concellos, u_i é o efecto aleatorio non estruturado que recolle a heteroxeneidade espacial e, por último, v_i representa o termo aleatorio estruturado que explica a agrupación espacial de casos ao redor dun concello. Para o efecto aleatorio u_i considerouse unha distribución normal de media cero e desvío típico τ_u. Para o efecto aleatorio v_i considerouse unha distribución autoregresiva condicional intrínseca(iCAR) normal de media, a media do efecto dos concellos veciños (é dicir, adxacentes), e desvío típico τ_v. Tanto para α como para τ_u e τ_v, consideráronse distribucións a priori pouco informativas, en concreto para os desvíos típicos considerouse a distribución gamma.
A partir deste modelo obtivéronse as RME suavizadas para cada concello como $\exp(u_i + v_i)$. Estes valores foron representados nun mapa.

Modelo espazo-temporal:
Partindo do modelo espacial engadiuse a variable tempo. O modelo axustado foi o modelo clásico proposto por Bernardinelli [5]:

$$\log(\rho_{ik}) = \alpha + u_i + v_i + (\beta + \delta_i)t_k$$

Onde β é a media da tendencia lineal para o conxunto de Galicia, δ_i representa a desviación da tendencia de cada concello con respecto á tendencia global β, e t_k é a variable tempo, onde k é o quinquenio (k=1, 2, 3). Asúmese que δ_i segue unha distribución normal de media cero e desvío τ_δ. Por último, para solventar o problema de identificabilidade do modelo impúxose a restrición de que os termos δ e u sumen cero.
A partir deste modelo obtivéronse as tendencias temporais (TT) para cada concello como $\exp(\beta + \delta_i)$. Estes valores foron representados nun mapa.

O modelo espacial foi aplicado a cada quinquenio e sexo e o modelo espacio-temporal a cada sexo. Todos eles foron implementados por medio do paquete R-INLA de R [6].

Mapas:
Nun mapa de concellos de Galicia foron representadas as RME suavizadas e as TT. Os rangos de valores foron divididos en 7 intervalos: [0;0,85]; (0,85;0,95]; (0,95;0,99]; (0,99;1,01]; (1,01;1,05]; (1,05;1,15] e (1,15;∞). Para as RME suavizadas, a cor vermella identifica aos concellos cun exceso de risco, a cor azul indica un menor risco e a cor branca indica un risco similar ao de Galicia no período 2000-2014. Para as TT, a cor laranxa representa unha tendencia crecente e a cor verde unha tendencia decrecente. En ambos casos, a intensidade da cor marca o grao de importancia.

3. RESULTADOS

No período 2000-2014 producíronse en Galicia 4.707 defuncións por suicidio, un 23,4% do total de falecidos por causas externas. O número de mortes foi en lixeiro aumento ao longo dos tres quinquenios estudados debido, fundamentalmente, ao comportamento dos homes xa que acumulan máis do 70% das defuncións en todos os quinquenios. No caso dos homes, as taxas brutas aumentaron un punto por quinquenio pasando de 16,4 casos por 100.000 habitantes no quinquenio 2000-04 a 18,5 casos por

[1] Rúbricas E950-E959 da Clasificación Internacional de Enfermidades, 9ª revisión (CIE-9) ata o ano 1998, e rúbricas X60-X84 da CIE-10 dende o ano 1999.

100.000 habitantes no quinquenio 2010-14 (Táboa 1). As mulleres mantivéronse en torno aos 5,7 casos por 100.000 habitantes.

	2000-04	2005-09	2010-14	GLOBAL
HOMES	1.083 (16,4)	1.148 (17,1)	1.243 (18,5)	3.474 (17,4)
MULLERES	389 (5,5)	428 (6,0)	416 (5,8)	1.233 (5,7)
TOTAL	1.472 (10,7)	1.576 (11,4)	1.659 (11,9)	4.707 (11,3)

Táboa 1: Número de mortes e taxas brutas por 100.000 habitantes (entre paréntese) por suicidio en Galicia segundo quinquenio e sexo.

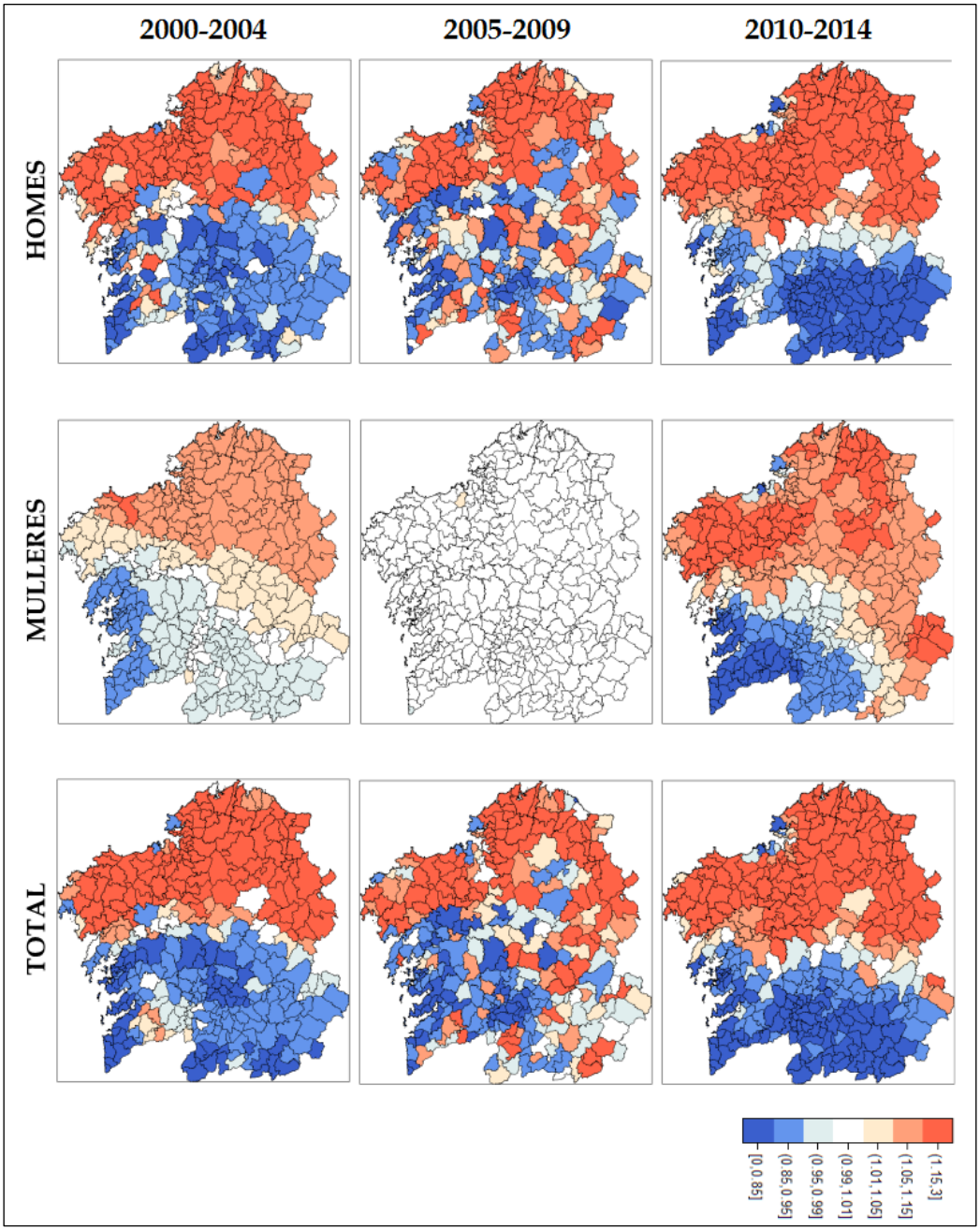


Figura 1: Distribución das razóns de mortalidade estandarizadas municipais suavizadas por medio do modelo espacial, para cada quinquenio e sexo. Taxas de referencia: taxas específicas de mortalidade por suicidio en Galicia no período 2000-2014.

Os mapas das RME suavizadas (Figura 1) amosan un comportamento norte-sur, no caso dos homes, onde os concellos situados na metade norte presentan un risco máis elevado en comparación ao conxunto de Galicia no período 2000-14, mentres que os situados na metade sur presentan un menor risco. É no quinquenio 2005-2009 cando ese patrón se desvirtúa lixeiramente, sobre todo na zona sur de Galicia.
No caso das mulleres obsérvase un patrón nordeste-suroeste (Figura 1). Compre destacar o comportamento no quinquenio 2005-2009, xa que os casos que cabería esperar neste quinquenio son similares aos observados polo que no mapa predomina totalmente a cor branca.

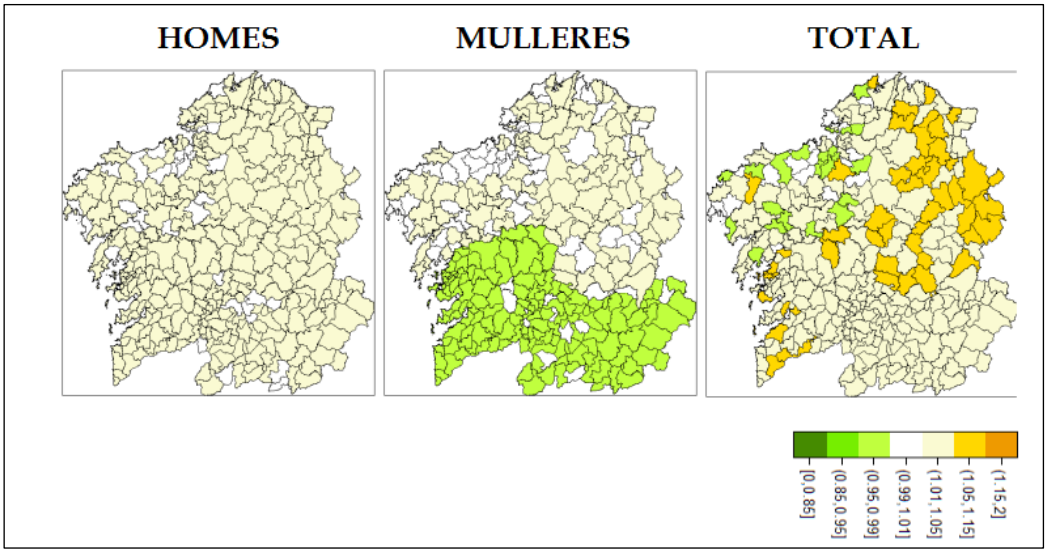


Figura 2: Tendencia temporal das razóns de mortalidade estandarizadas municipais por medio do modelo espazo-temporal, para cada sexo.

A tendencia temporal estimada en homes foi lixeiramente crecente en practicamente todos os concellos de Galicia. Mentres que no caso das mulleres, os concellos da rexión norte foron en lixeiro aumento, en contraposición do que se observa nos concellos da rexión sur (Figura 2).

4. CONCLUSIÓNS

A mortalidade por suicidios durante os quinquenios 2000-04, 2005-09 e 2010-14 en Galicia presenta un claro patrón norte-sur, cun exceso de risco nos concellos situados na metade norte e un menor risco nos concellos situados na metade sur. De forma xeral, a tendencia ao longo dos quinquenios a estudo foi lixeiramente crecente.

REFERENCIAS

[1] Rué M, Borrell C. Los métodos de estandarización de tasas. Revisiones en Salud Pública. 1993;3:263-95.

[2] Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics. 1987;43(3):671-81.

[3] Rué H, Martimo S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B. 2009;71:319-92.

[4] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Statist Math. 1991;43(1):1-59.

[5] Bernardinelli L, Clayton D, Montomoli C. Bayesian estimates of disease maps: how important are priors? Stat Med. 1995;14:2411-31.

[6] Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA. Spat Spatio-Temporal Epidemiol. 2013;4:33–49.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de julio, 1 y 2 de julio de 2016

INCIDENCIA DE CÁNCER EN PACIENTES VIH ¿ES MAYOR EN PACIENTES COINFECTADOS POR EL VIRUS DE LA HEPATITIS C?

S Pértega Díaz[1], MT Seoane Pillado[1], I Rodriguez Osorio[2], A Castro, H Meijide, E Poveda, A Mena de Cea

[1] Unidad de Epidemiología Clínica y Bioestadística, Complexo Hospitalario Universitario A Coruña, Instituto de Investigación Biomédica A Coruña
[2] Servicio de Medicina Interna, Complexo Hospitalario Universitario A Coruña
[3] Grupo de Virología Clínica, Instituto de Investigación Biomédica A Coruña

RESUMEN

Se llevó a cabo un estudio de seguimiento ambispectivo para determinar la incidencia de cáncer en pacientes VIH según la presencia o no de coinfección por el virus de la hepatitis C. La tasa de incidencia de cáncer fue de 696 tumores por 100.000 personas/año, significativamente más elevada que en población general (Razón de incidencias estandarizada=3,8; 95% IC 3,3-4,4). La coinfeccion por VHC se asoció a una mayor tasa de incidencia de tumores no definitorios de SIDA.

Palabras y frases clave: Neoplasias; VIH; Virus de la hepatitis C.

1. OBJETIVOS

El objetivo de este estudio fue determinar la incidencia de cáncer en pacientes VIH según la presencia o no de coinfección por el virus de la hepatitis C (VHC), y compararla con la incidencia en población general.

2. MÉTODOS

Estudio de seguimiento ambispectivo de pacientes VIH diagnosticados en el periodo 1985-2013 y seguidos en el Complexo Hospitalario Universitario A Coruña. La variable principal fue la confirmación anatomopatológica de diagnóstico de cáncer (tumores definitorios de SIDA y tumores no definitorios) en el seguimiento, que se extendió hasta Julio de 2015.
Se calculó la tasa de incidencia de cáncer cruda y ajustada por el método indirecto, utilizando los datos de incidencia de cáncer en España publicados en el informe GLOBOCAN 2012. Para comparar la incidencia observada con la publicada en población general, se calculó la razón de incidencias estandarizada (SRI) y su intervalo de confianza al 95%.
Para analizar la incidencia acumulada de cáncer en el seguimiento tras el diagnóstico de VIH se aplicó una metodología de riesgos competitivos, considerando la muerte por cualquier causa como evento que "compite" con la incidencia de cáncer.

3. RESULTADOS

Se incluyeron n=2318 pacientes (75.8% hombres, edad media al diagnóstico del VIH 32,3±10,1 años), que dieron lugar a 26.580 personas-año de seguimiento. De ellos, 857 (37,0%) eran pacientes coinfectados por VHC y 1461 (63,0%) pacientes monoinfectados por VIH.

Durante el seguimiento, se diagnosticaron 185 tumores, con una tasa de incidencia (TI) de 696 tumores por 100.000 personas/año, significativamente más elevada que en población general (SIR=3,8; 95% IC 3,3-4,4).
La incidencia de tumores no definitorios de SIDA fue de 349,4 tumores/100.000 personas/año, más elevada en pacientes coinfectados (TI=415,0; SIR=3,4) que en monoinfectados (TI=377,3; SIR=1,8). La incidencia de tumores definitorios, por el contrario, fue más elevada en pacientes monoinfectados (TI=442,2; SIR=37,4) que en pacientes coinfectados (TI=127,2; SIR=27,2).
La incidencia acumulada de cáncer es del 3,6% a los 5 años del diagnóstico de VIH, del 5,5% a los 10 años y del 10,4% a los 20 años. Un 50,2% de los pacientes siguen vivos y libres de tumores a los 20 años tras el diagnóstico. Tras ajustar por edad al diagnóstico, sexo y vía de transmisión, los pacientes coinfectados por VHC presentan significativamente una mayor incidencia de tumores no definitorios de SIDA (HR ajustado=1,7; 95% IC=1,1-2,8) y menor incidencia de tumores definitorios (HR ajustado=0,44; 95% IC=0,25-0,74).

4. CONCLUSIONES

Los pacientes VIH presentan una incidencia de cáncer superior a la de la población general. La coinfeccion por VHC se asocia a una mayor tasa de incidencia de tumores no definitorios de SIDA.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

INCIDENCIA DE LESIONES MEDULARES EN LA COMUNIDAD AUTÓNOMA GALLEGA EN 1995-2014

V Balboa Barreiro[1], A Montoto Marqués[2], ME Ferreiro Velasco[2], S Salvador de la Barrera[2], A Rodríguez Sotillo[2], S Pértega Díaz[1] y MT Seoane Pillado[1]

[1] Unidad de Epidemiología Clínica y Bioestadística, Complexo Hospitalario Universitario A Coruña, Instituto de Investigación Biomédica A Coruña

[2] Unidad de Lesiones Medulares, Complexo Hospitalario Universitario A Coruña

RESUMEN

Estudio de la incidencia de lesión medular según la etiología de la lesión y las causas más frecuentes durante el período 1995 – 2014 en la Comunidad Autónoma de Galicia. Se calcularon las tasas crudas, las tasas estandarizadas por edad (según población mundial estándar) y se estimó el cambio porcentual anual mediante regresion “joinpoint”.

Palabras y frases clave: Lesión medular; Incidencia; Regresión joinpoint.

1. OBJETIVOS

El objetivo de este estudio fue determinar la incidencia de lesiones medulares en la población gallega en el período comprendido entre los años 1995 y 2014.

2. MÉTODOS

En este estudio se incluyeron los pacientes afectados de lesiones medulares ingresados en el Complexo Hospitalario Universitario A Coruña (CHUAC) en el período 1995 - 2014. La variable principal del estudio fue la etiología de la lesión medular, siendo esta de tipo traumático o médico. Entre las lesiones traumáticas se analizaron las causadas por accidentes de tráfico y por caídas.
Se calculó la tasa de incidencia cruda para cada tipo de lesión, como denominadores se han utilizado las estimaciones de población en Galicia a 1 de julio de cada año, por grupos de edad y sexo, facilitadas por el INE.
Se calcularon las tasas estandarizadas ajustadas por edad por el método directo, usando como referencia la población mundial. Todas las tasas fueron calculadas para la población total y para cada sexo, siendo expresadas como tasas por 100.000 personas-año.
Para el análisis de tendencias de las tasas crudas se utilizaron modelos de regresión 'joinpoint', útiles para identificar cambios significativos en las tendencias y estimar la magnitud del aumento o descenso en cada intervalo. Dicho aumento o descenso fue expresado como porcentaje de cambio anual (PCA) con su correspondiente intervalo de confianza al 95 %.

3. RESULTADOS

En el estudio se incluyeron n=1567 pacientes con lesiones medulares (1132 (72,20 %) hombres, edad media en el momento de la lesión 51,87 ± 20,56 años). El 76.26 % de los pacientes presentaban lesiones de tipo

traumático (913 (76,40 %) hombres) y el 23.74 % de las lesiones eran de tipo médico (219 (58,87 %) hombres).
Entre las lesiones traumáticas, las causas más comunes fueron las caídas (648 (54,22 %)) y los accidentes de tráfico (442 (37 %)).

Las tasas crudas de lesiones presentaron un descenso anual significativo de un -0,9 % (IC 95 % =(-1,7 -0,1)), pasando de 3,05 lesiones / 100.000 personas-año en el 1995, a un 2,66 lesiones / 100.000 personas-año en el 2014, siendo más elevada entre el sexo masculino -1,4 (IC 95 % = (-2,2 -0,6)). Las lesiones traumáticas presentaron un mayor descenso, siendo el procentaje de cambio de un -1,4 % anual (IC 95 % = (-2,4 -0,5)), pasando de una tasa de 2,5 lesiones traumáticas / 100.000 personas-año en 1995 a 1,89 lesiones traumáticas / 100.000 personas-año en 2014. Las lesiones médicas mostraron un ligero aumento no significativo de 1 % anual (IC 95 % = (-0,3 2,2)), pasando de 0,55 a 0,79 lesiones / 100.000 personas-año desde 1995 a 2014. Las lesiones traumáticas causadas por accidentes de tráfico presentaron un descenso significativo de un -4,9 % (IC 95 % = (-6,9 -2,7)), siendo de 1,25 la tasa cruda por 100.000 personas-año en 1995 y 0,25 en 2014. El sexo femenino mostró un descenso más pronunciado en este período -5,9 % (IC 95 % = (-9,2 -2,5)) frente a un descenso anual de -4,3 % (IC 95 % = (-7 -1,4)) entre los hombres.
En ninguno de los casos se identificaron cambios de tendencia de forma significativa en las tasas.
Las tasas de incidencia ajustadas por edad, tomando como denominador la población muncial, mostraron una incidencia anual de 2,42 lesiones totales / 100.000 personas-año (3,71 en hombres y 1,14 en mujeres), 1,88 lesiones traumáticas (3,09 en hombres y 0,73 en mujeres), 0,54 lesiones médicas (0,66 en hombres y 0,42 en mujeres), 0,79 lesiones traumáticas a causa de accidentes de tráfico (1,30 en hombres y 0,30 en mujeres) y 0,90 lesiones traumáticas por caídas (1,44 en hombres y 0,39 en mujeres).

4. CONCLUSIONES

El estudio ha revelado que a lo largo del período 1995 – 2014 tuvo lugar un descenso significativo de las lesiones medulares en general y, particularmente en las lesiones de tipo traumático, no mostrando cambios significativos la incidencia de lesiones médicas. En este período se puede apreciar un considerable descenso de los accidentes de tráfico, manteniéndose esta tendencia en ambos sexos.

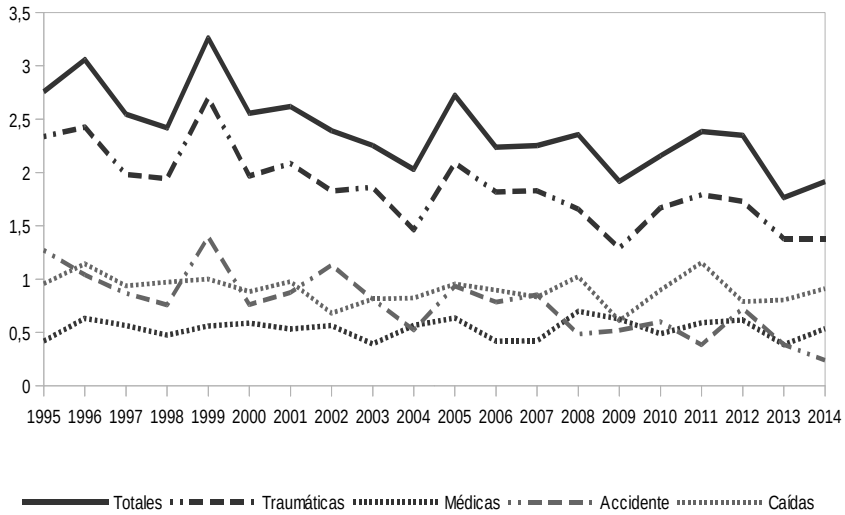


Figura 1: Tasas estandarizadas por edad, tomando como referencia la población mundial. Expresadas en tasas por 100.000 personas-año.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

INFLUÊNCIAS DAS VARIAÇÕES CLIMÁTICAS NA INCIDÊNCIA DE CASOS DE DENGUE NA CIDADE DE GOIÂNIA, BRASIL

Antônio Neco[1], Raquel Menezes[2] e Susana Faria[2]

[1]Instituto Federal Goiano / Escola de Ciências - Universidade do Minho
[2]Departamento de Matemática e Aplicações, Escola de Ciências, Universidade do Minho

RESUMO

A dengue é uma doença transmitida pelo mosquito *Aedes aegypti*, típica das regiões tropicais e subtropicais, onde as variações climáticas propiciam a proliferação do vetor transmissor. A cidade de Goiânia, com uma população estimada em 1,5 milhões de habitantes, apresenta temperatura média anual entre 18°C e 26°C e duas estações bem definidas: verão húmido (dezembro a março) e inverno seco (junho a agosto), propícias para o desenvolvimento do mosquito transmissor da dengue. Neste trabalho, utiliza-se a metodologia das equações de estimação generalizadas (EEG) para estudar a relação entre o número de casos de dengue e as variações climáticas. Os resultados indicam que as variações na precipitação acumulada, na temperatura mínima e na humidade relativa do ar, contribuem positivamente para o aumento do número de casos de dengue, exigindo ações do poder público e da população em geral para eliminar os criadouros do mosquito e evitar períodos epidêmicos.

Palavras e frases chave: Variações climáticas, dengue, distribuição binomial negativa, equações de estimação generalizadas.

1. INTRODUÇÃO

A febre dengue é uma doença viral transmitida pelo mosquito *Aedes aegypti*, com rápida disseminação em centros urbanos. O Brasil concentra mais de 80% da população em zonas urbanas com lacunas de infraestrutura de saneamento básico, além dos fatores climáticos favoráveis que dificultam as ações para a erradicação do vetor transmissor da dengue e de seus criadouros [1].
No Brasil, as primeiras epidemias de dengue ocorreram em 1981-1982 em Boa Vista (estado de Roraima) e em 1986 no Rio de Janeiro e em algumas capitais do Nordeste. Desde então, ocorrem epidemias associadas à circulação dos quatro sorotipos denominados DEN-1, DEN-2, DEN-3 e DEN-4, sendo todos causadores de doenças graves e fatais. Para acompanhar o padrão de transmissão da doença e possibilitar ações preventivas, o agravo de notificação compulsória (Portaria GM/MS nº 5 de 21/02/2006) e todos os casos suspeitos devem ser notificados à vigilância epidemiológica do município pelas unidades de saúdes.
A cidade de Goiânia localiza-se na região central do Brasil (Latitude: -15,91°, Longitude: -50,13°, Altitude: $512,22m$), tem uma área de $729,02km^2$ e população estimada em 1,5 milhões de habitantes. A temperatura média varia entre 18°C e 26°C, tendo um verão húmido entre os meses de dezembro a março, o que favorece a proliferação do mosquito *Aedes aegypti*, sendo que a primeira epidemia de dengue foi registrada em 1994, com repetições nos anos de 2008, 2010, 2013 e 2015.

Neste trabalho utiliza-se a metodologia das equações de estimação generalizadas para estudar como as variações climáticas influenciam positivamente a incidência dos casos de dengue. Ajusta-se um modelo de regressão binomial negativa para predizer o número de casos de dengue em função das variações climáticas, o que possibilitará ações públicas na definição de estratégias para eliminar os criadouros do vetor transmissor da dengue, com medidas preventivas que possam ser adotadas pela população para evitar períodos epidêmicos e mortes por febre dengue.

2. METODOLOGIA E DISCUSSÃO DOS RESULTADOS

A base de dados foi construída a partir das informações de notificações de casos de dengue[1], das informações meteorológicas[2] e dos números da população[3] goianiense para o período de 2008 a 2015. O número total de observações é 377 representando as informações semanais para o número de casos de dengue, precipitação (mm), temperatura mínima (°C), temperatura máxima (°C), humidade relativa do ar (%) e população. A partir das variáveis preditoras precipitação (*prect*), temperatura mínima (*tmin*) e humidade relativa do ar (*hram*), foram criadas variáveis desfasadas em 0, 1, 2, ..., 10 semanas. Inicialmente foram ajustados Modelos Lineares Generalizados (MLG) com distribuição de Poisson por tratar-se de dados de contagem. Ao identificar sobredispersão, optou-se pela distribuição binomial negativa [2] para ajustar um modelo que melhor explique a incidência dos casos de dengue semanal a partir das variações climáticas desfasadas no tempo [3]. Devido à existência de correlação temporal, utiliza-se o método de EEGs que ampliam os MLGs para permitir a análise de dados longitudinais e correlacionados [4]. A análise estatística foi realizada com os softwares estatísticos R e IBM-SPSS.

2.1. Análise das variações meteorológicas para o período de 2008 a 2015

Para o período em análise, a cidade de Goiânia apresenta temperatura mínima a variar entre 9,8°C e 21,5°C e a temperatura máxima entre 29,4°C e 39,6°C, sendo a temperatura entre 21,5°C e 29,4°C correspondente ao período que registra o maior número de casos dengue. A humidade relativa do ar média varia entre 26,04% e 84,68%, sendo que ocorre maior incidência de dengue para o período em que se mantém acima 60,31% (média para o período observado). E o período de chuva com maior incidência de casos de dengue corresponde aos meses de outubro a abril, com a precipitação acumulada semanal acima da média para o período (34,04 mm). A Figura 1 apresenta as variações climáticas e a incidência dos casos de dengue para o período de 2008 a 2015.

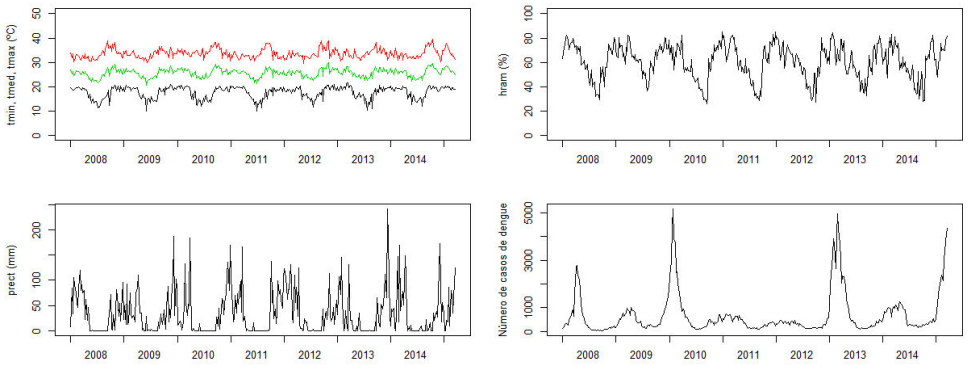


Figura 1: Variações meteorológicas semanais e número de casos de dengue.

O mosquito transmissor da dengue desenvolve-se em águas limpas e paradas, em criadouros formados após o início do período de chuva. O ciclo de desenvolvimento do *Aedes aegypti*, incluindo a transmissão da dengue e a manifestação da doença na pessoa infectada, dura aproximadamente 40 dias dependendo das variações climáticas. A partir de então, caso não haja ações conjuntas

[1]Disponibilizadas pela Superintendência de Vigilância em Saúde do Estado de Goiás (SUVISA-GO).
[2]Disponíveis em http://www.inmet.gov.br/projetos/rede/pesquisa/.
[3]Disponibilizadas pelo Instituto Brasileiro de Geografia e Estatística (IBGE), www.ibge.gov.br

entre o poder público e a população, os casos de dengue poderão evoluir para uma epidemia, com o período mais crítico correspondendo aos meses de janeiro a abril de cada ano.

Para examinar o grau de associação existente entre a incidência dos casos de dengue e as variações climáticas, desfasadas de 0 a 10 semanas, utilizou-se o coeficiente de correlação de *Spearman* (ρ). Na Tabela 1 observam-se as associações positivas com precipitação, temperatura mínima e humidade relativa do ar, com as maiores correlações para os desfasamentos de seis, sete e cinco semanas, respetivamente. Temperatura máxima apresenta baixa associação negativa, invertendo-se a partir da sexta semana de desfasamento, não sendo incluída no modelo ajustado.

Lag	prect (mm)	tmin (°C)	tmax (°C)	hram (%)
0	0.4040**	0.4162***	**-0.2895***	0.5716***
1	0.4651**	0.4820***	-0.2206	0.6026***
2	0.5064**	0.5392***	-0.1609	0.6156***
3	0.5459***	0.5915***	-0.1233	0.6401***
4	0.5861***	0.6403***	-0.0725	0.6476***
5	0.6044***	0.6652***	-0.0305	**0.6569*** **
6	**0.6333*** **	0.6821***	0.0152	0.6513***
7	0.6308***	**0.6988*** **	0.0726	0.6238***
8	0.6254***	0.6860***	0.1138	0.5960***
9	0.6152***	0.6749***	0.1519	0.5610***
10	0.5849***	0.6556***	0.1978	0.5179***

Nível de significância: 0.05; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.0001$.

Tabela 1: Correlações das variáveis meteorológicas com o número de casos de dengue.

2.2. Abordagem EEG com a distribuição binomial negativa

A incidência de dengue semanal será modelada utilizando a metodologia das equações de estimação generalizadas (EEG) [4] com a distribuição binomial negativa para acomodar a sobredispersão e a correlação temporal dos dados. Considere-se então y_{ij} como sendo o número de casos de dengue na semana $i = 1, ..., 52$ e ano $j = 1, ..., 8$. Numa primeira fase deste trabalho, cujos resultados são aqui apresentados, assume-se a matriz identidade como matriz de correlação entre as observações, ou seja, assume-se que as observações repetidas para uma dada semana são independentes entre si. O logaritmo da população (*popul*) foi adicionado ao modelo como variável *offset*. A seleção do modelo que melhor se ajusta aos dados baseia-se no critério *QIC (Quasi Likelihood Under Independence Model Criterion)*. Considerando a distribuição binomial negativa com função de ligação logarítmica, parâmetro escalar e o preditor linear dado por $\eta_{ij} = x_{ij}^T\beta$, tem-se:

$$y_{ij}|x = x_{ij} \sim BinNeg(\mu_{ij}, k)$$

$$ln(\mu_{ij}) = \beta_0 + \beta_1 * prect_{(i-l)j} + \beta_2 * tmin_{(i-l)j} + \beta_3 * hram_{(i-l)j} + ln(popul_j)$$

onde μ_{ij} é a respetiva média de casos de dengue. O modelo ajustado com as variáveis que apresentaram maior correlação com a variável resposta (Tabela 1) não atendeu aos critérios de seleção. Procurou-se ajustar um modelo com uma estrutura de variáveis mais simples, obtendo-se um modelo com precipitação, temperatura mínima e humidade relativa do ar (*prect*, *tmin* e *hram*) desfasadas em seis, quatro e dez semanas, respetivamente. As estimativas dos parâmetros são apresentadas na Tabela 2.

Variáveis	Estimate	Std.err	P
$(Constante)$	-13.714	0.5927	< 0.0001
$prect(lag = 6)$	0.004	0.0009	< 0.0001
$tmin(lag = 4)$	0.233	0.0279	< 0.0001
$hram(lag = 10)$	0.027	0.0038	< 0.0001
$QIC = 433.367$			

Tabela 2: Estimativas dos parâmetros utilizando EEG-BN e estrutura de correlação independente.

A partir dos valores estimados para os parâmetros preditores (Tabela 2), a alteração no valor de uma variável preditora, mantendo as demais constantes, temos: O aumento de 1°C na temperatura mínima provoca um aumento de 26% no número de casos de dengue passadas quatro semanas; A cada incremento de 1% na humidade relativa do ar, temos um aumento aproximado de $2,7\%$ no número de casos de dengue passadas dez semanas; Para a precipitação acumulada de 100 mm, temos um aumento de 40% nos casos de dengue, após seis semanas.
A Figura 2 apresenta o número de casos de dengue observados, juntamente com os valores ajustados pelo modelo EEG-BN, ao longo do período em estudo.

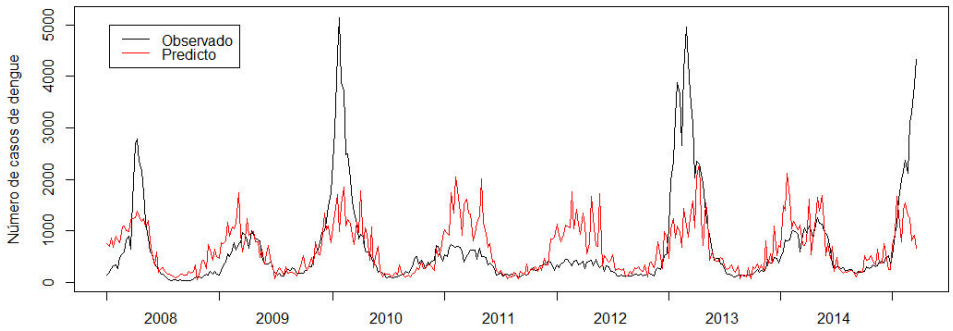


Figura 2: Número de casos de dengue observados e ajustados pelo modelo.

3. CONCLUSÕES

O mosquito *Aedes aegypti*, desenvolve-se em ambientes com água parada e temperatura média propícia entre 21°C e 29°C. Entre os meses de outubro a abril, a cidade de Goiânia apresenta um verão quente e húmido com precipitações acumuladas mensais acima de 100 mm, contribuindo para o acúmulo de água parada e a formação de criadouros para o vetor transmissor da dengue.
Neste trabalho, a incidência de dengue semanal foi modelada utilizando o método das equações de estimação generalizadas (EEGs) e a distribuição binomial negativa, com a seleção do modelo definida pelo menor *QIC*. De acordo com os preditores estimados, temos a indicação que as variações climáticas contribuem positivamente para o aumento dos casos de dengue nos centros urbanos.
Sob as condições climáticas favoráveis que foram indicadas, tornam-se necessárias ações públicas para a eliminação dos criadouros e dos mosquitos *Aedes aegypti* para evitar a ocorrência de epidemias e mortes por dengue.

AGRADECIMENTOS

Departamento de Informática do Instituto Federal Goiano, Campus Morrinhos (GO);
Departamento de Matemática e Aplicações da Escola de Ciências da Universidade do Minho.

Referências

[1] Brasil. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. (2009). Diretrizes nacionais para prevenção e controle de epidemias de dengue. Brasília: Ministério da Saúde.

[2] Paula, G. A. (2004). Modelos de regressão: com apoio computacional. São Paulo: IME-USP.

[3] Gomes, A. F., Nobre, A. A., & Cruz, O. G. (2012). Temporal analysis of the relationship between dengue and meteorological variables in the city of Rio de Janeiro, Brazil, 2001-2009. Cadernos de Saúde Pública, 28(11), 2189-2197.

[4] Diggle, P., Heagerty, P., Liang, K. Y., & Zeger, S. (2013). Analysis of longitudinal data. OUP Oxford.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Sample size calculators for estimating antimalarial seroconversion rate and its putative changes over time

Nuno Sepúlveda[1,2], Carlos Daniel Paulino[3] and Chris Drakeley[1]

[1]London School of Hygiene and Tropical Medicine, United Kingdom
[2]Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[3]Departamento de Matemática, Instituto Superior Técnico, Portugal

ABSTRACT

According to the 2015 World Malaria Report, global disease incidence and mortality rates have been going down over the last 10 years. The same report compiles a list of more than 30 countries where malaria elimination and eradication might be achieved in a near future. With the decrease of disease burden in these countries, there is a high chance of sampling a few infected individuals in surveillance studies, thus, limiting the use of infection-based epidemiological measures for estimating malaria transmission intensity and its putative change over time. To overcome this problem, an alternative epidemiological approach based anti-malarial antibody data has been proposed and tested in different epidemiological settings. In this approach, it is of interest to estimate the seroprevalence - the prevalence of seropositive individuals in a sample - and the seroconversion rate - the rate by which seronegative individuals become seropositive that can be seen as a proxy of a malaria transmission intensity. However, the use of these serological measures brings important statistical challenges in terms of sample design. With this in mind, we propose sample size calculators (i) to estimate these measures with a given precision under the assumption of a stable disease transmission [1] and (ii) to detect changes in disease transmission intensity with a given probability [2]. The use of these calculators is illustrated with hypothetical studies to be conducted in Africa and elsewhere.

Keywords and key sentences: study design, age structure, seroprevalence, precision, sample size, reversible catalytic model.

References

[1] Sepúlveda N., Drakeley C. (2015). Sample size determination for estimating antibody seroconversion rate under stable malaria transmission intensity. *Malaria Journal* 14, 141.

[2] Sepúlveda N., Paulino C.D., Drakeley C. (2015). Sample size and power calculations for detecting changes in malaria transmission using antibody seroconversion rate. *Malaria Journal* 14, 529.

II Encontro Galaico-Português de Biometría
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Análise de tempo até recidiva em cancro da mama

Liliana Coelho[1], Ana Borges[2], Luís Castro[3] e Inês Sousa[4]

[1] DMA-ECUM, Universidade do Minho
[2] CIICESI, ESTGF - Porto Politécnico
[3] Hospital de Braga
[4] CMAT e DMA-ECUM, Universidade do Minho

RESUMO

O presente trabalho objetiva contribuir para a compreensão do comportamento do cancro da mama, uma doença com contornos complexos, na população portuguesa, especificamente na população da Unidade de Senologia do Hospital de Braga, localizado no norte de Portugal.

Com o particular intuito de compreender quais os fatores de risco relacionados com a recidiva do cancro da mama, na referida população, foi realizada uma análise de sobrevivência de dados de pacientes do Hospital de Braga. Para tal foi recolhida informação dos registos médicos do Hospital, relativa a 540 pacientes diagnosticadas com tumor maligno da mama no período de 2008 a 2012, ou em acompanhamento à data de 2008, com idade ao diagnóstico a variar entre os 20 e os 89 anos.

Os dados recolhidos englobam um conjunto de 50 variáveis explicativas tanto ao nível de características individuais como ao nível do tumor. Para inferir relativamente a possíveis fatores de risco de recidiva de entre as variáveis recolhidas, recorreu-se ao modelo de riscos proporcionais de Cox, considerando como tempo de referência o tempo entre o diagnóstico do tumor e a data da primeira recidiva, ou até ao final do estudo onde considerada a data de 30/11/2014.

Para estimar o risco relativo de recidiva de cancro da mama, procedeu-se inicialmente com o cálculo das estimativas não paramétricas de Kaplan-Meier, para cada variável, estratificadas por categoria, comparando-se estas através da sua representação gráfica. Para avaliar a significância estatística das diferenças entre as curvas de sobrevivência de cada categoria de variável, fez-se uso do teste de log-rank ou teste de Mantel-Haenszel, onde a hipótese nula a ser testada é: não há diferença entre as (verdadeiras) curvas de sobrevivência.

Considerando as variáveis estatisticamente significativas, obtidas nos testes de log-rank, ajustaram-se vários modelos de sobrevivência com múltiplas covariáveis para estimar o efeito conjunto de diversas variáveis independentes na probabilidade de recidiva dos pacientes. O modelo final escolhido foi estimado por "step-wise backwards", ou seja, começando com o modelo saturado considerando todas as variáveis significativas, e, em seguida, eliminando uma-por-uma as variáveis menos significativas.

Para diagnóstico do modelo ajustado foi representada graficamente uma sobreposição da curva de sobrevivência não paramétrica (Kaplan-Meier) versus a curva do modelo teórico, para uma determinada combinação de valores de covariáveis. Representou-se, também, graficamente os resíduos de Cox-Snell para avaliar a qualidade global do ajuste.

Palabras e frases chave: Cancro da mama, sobrevivência, recidiva.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de julio, 1 y 2 de julio de 2016

ASOCIACIÓN DE LOS HAPLOGRUPOS MITOCONDRIALES CON LA PROGRESIÓN RADIOGRÁFICA DE LA ARTROSIS DE RODILLA

I Rego Pérez[1], S Pértega Díaz[2], M Fernández Moreno[1], A Soto Hermida[1], ME Vázquez Mosquera[1], E Cortés Pereira[1], S Relaño[1], N Oreiro Villar[1], C Fernández López[1], FJ Blanco[1]

[1] Servicio de Reumatología. Instituto de Investigación Biomédica de A Coruña (INIBIC). Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC).
[2] Unidad de Epidemiología Clínica y Bioestadística. Instituto de Investigación Biomédica de A Coruña (INIBIC). Complexo Hospitalario Universitario de A Coruña (CHUAC), Sergas. Universidade da Coruña (UDC).

RESUMEN

Estudio de asociación y metaanálisis de la relación entre los diferentes haplogrupos mitocondriales y la probabilidad de progresión radiológica en pacientes con artrosis de rodilla, utilizando datos de diferentes cohortes prospectivas estudiadas a nivel internacional. Se objetiva como las variantes del superhaplogrupo mitocondrial JT se asocian con una menor tasa de progresión en estos pacientes.

Palabras y frases clave: Osteoartritis; Haplogrupos, Mitocondria, Progresión radiológica.

1. INTRODUCCIÓN

El objetivo de este estudio fue estudiar la asociación entre los diferentes haplogrupos mitocondriales y la probabilidad de progresión radiológica en pacientes con artrosis de rodilla, a partir de los resultados de diferentes cohortes prospectivas estudiadas a nivel internacional.

2. MÉTODOS

Se determinó la asociación de los diferentes haplogrupos mitocondriales con la tasa de progresión radiológica a los 96 meses en n=431 pacientes de la cohorte CHECK (Cohort Hip and Cohort Knee). Se definió progresión como un aumento en la escala de Kellgren y Lawrence (KL) desde un grado ≥1 en la visita basal a cualquier grado mayor durante el seguimiento. Para el análisis, se consideró la extensión de Turnbull del método de Kaplan-Meier para datos censurados por intervalos. En el análisis multivariante, se realizó un modelo de regresión de Cox extendido.
Posteriormente, se llevó a cabo un meta-análisis de los resultados incluyendo además los hallazgos obtenidos al respecto en la cohorte de la Osteoarthritis Initiative (OAI) y en una cohorte de pacientes con artrosis de rodilla publicados con anterioridad por nuestro grupo en España. Se consideró un modelo de efectos aleatorios, tomando los valores de hazard ratio (HR) ajustados como medida del efecto.

3. RESULTADOS

Para la cohorte CHECK, los pacientes con un haplogrupo T mostraron la tasa de progresión radiológica más baja (HR=0.645; 95% IC=0.419-0.978, $p<0.05$). Considerando los pacientes en el superhaplogrupo JT se observó la misma tendencia (HR=0.707;95%CI:0.501–0.965;$p<0.05$). Un mayor índice de masa corporal (HR=1.046;95%CI:1.018–1.073;$p<0.05$) y el presentar artrosis bilateral (HR=2.266; 95%CI:1.733–2.954;$p<0.05$) en el momento basal se asociaron asimismo a una mayor probabilidad de progresión de la artrosis

Tras llevar a cabo el meta-análisis, se objetivó una menor tasa de progresión de la artrosis en pacientes con el haplogrupo T (HR=0.612;95%CI=0.454-0.824;p=0.001) o en el superhaplogrupo JT (HR=0.765;95%CI=0.624-0.938;p=0.009).

4. CONCLUSIONES

Las variantes del superhaplogrupo mitocondrial JT se asocian con una menor tasa de progresión en pacientes con artrosis de rodilla. Esta información es un potencial biomarcador genético a tener en cuenta a la hora de valorar la probabilidad de progresión de la enfermedad en estos pacientes.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

CRIB vs SNAPPE II - avaliação espacial do seu desempenho através do Comp2ROC

Maria Filipa Mourão[1], Ana Cristina Braga[2]

[1]Instituto Politécnico de Viana do Castelo, Escola Superior de Tecnologia e Gestão
[2] Centro ALGORITMI, Universidade do Minho

RESUMO

Neste trabalho apresenta-se a comparação do desempenho das escalas de gravidade clínica, Clinical Risk Index for Babies (CRIB) e Score for Neonatal Acute Physiology II (SNAPPE II), aplicadas nas Unidades de Cuidados Intensivos de Neonatologia de Portugal. Este estudo é efetuado a nível diferenciado pela Nomenclatura das Unidades Territoriais para fins Estatísticos II (NUTS II) que considera as seguintes regiões: Norte (Viana do Castelo, Braga, Porto, Vila Real e Bragança), Centro (Coimbra, Castelo Branco, Leiria, Viseu, Aveiro e Guarda), Lisboa/Vale do Tejo (Lisboa e Setúbal), Alentejo (Évora, Santarém, Beja e Portalegre), Algarve (Faro) e Ilhas (Açores e Madeira). As escalas CRIB e SNAPPE II aplicam-se a recém-nascidos de muito baixo peso (<1500 gramas) e/ou a recém-nascidos com menos de 32 semanas de gestação e discriminam entre recém-nascidos "falecidos" e "sobrevivos". Esta capacidade discriminante e o consequente desempenho das escalas na avaliação do risco de morte dos recém-nascidos é avaliada através da análise ROC (Receiver Operating Characteristic) recorrendo à representação gráfica dos pares de valores (sensibilidade, 1-especificidade), obtidos a partir dos resultados dessas escalas. Esta representação gráfica denomina-se curva ROC e permite, a partir da área abaixo dessa curva (AUC - Area Under the Curve), a comparação do desempenho entre diferentes testes de diagnóstico. A partir de dados demográficos e clínicos, recolhidos entre 2010 e 2012 pelo Registo Nacional de Muito Peso (RNMBP), associados à população servida pelas unidades de neonatologia, recorreu-se ao *software* IBM® SPSS® Statistics e ao *package* Comp2ROC do R, desenvolvido por Braga et al. [3], para estabelecer essa comparação com base na metodologia não paramétrica de estimação da curva ROC. O Comp2ROC tem a vantagem de permitir comparar curvas ROC, associadas a dois sistemas de diagnóstico, que se cruzam entre si com base no índice AUC e usa a metodologia proposta por Braga et al. [1]. É constituído por várias funções que permitem obter a curva ROC e medidas estatísticas de dispersão e localização.

Palavras e frases chave: curva ROC, Comp2ROC, CRIB (Clinical Risk Index for Babies), Score for Neonatal Acute Physiology II (SNAPPE II)

1. INTRODUÇÃO

Os cuidados intensivos pediátricos em Portugal são prestados por uma diversidade de Unidades de Cuidados Intensivos interessando neste estudo, em particular, as unidades que prestam cuidados

a recém-nascidos de muito baixo peso (<1500 g) e/ou com menos de 32 semanas de gestação. Atualmente, são usadas como escalas de avaliação da gravidade clínica de recém-nascidos, a Clinical Risk Index for Babies (CRIB) e a Score for Neonatal Acute Physiology II (SNAPPE II). Cada unidade procede a uma recolha individualizada dos dados referentes à sua atividade e que são enviados para o RNMBP, que é a entidade responsável pelo armazenamento desta informação. Os dados que integram o presente estudo foram recolhidos, entre 2010 e 2012, pelas unidades de cuidados intensivos neonatais do território português (Continente e Ilhas) e o objetivo passa pela avaliação e comparação do desempenho desses dois indicadores, no que respeita à sua capacidade discriminante na mortalidade neonatal, com recurso à metodologia ROC. A análise ROC foi desenvolvida nos anos 50, na Teoria da Decisão e a sua primeira aplicação foi motivada por problemas práticos na teoria de deteção de sinal [7] e [8]. O livro de Swets e Pickett [13] marcou o início da difusão desta técnica na área da biomedicina, inicialmente em radiologia. O desempenho de um teste, utilizado no auxílio de diagnósticos clínicos, é usualmente avaliado pela curva ROC, uma representação visual direta no plano unitário do desempenho de um teste de diagnóstico, de acordo com o conjunto das suas possíveis respostas, através da curva ROC. A AUC é uma medida resumo dessa curva, amplamente utilizada ([12], [9], [13]), para avaliar o desempenho de um teste diagnóstico, porque a sua estimativa leva em consideração todas as sensibilidades e especificidades associadas a cada valor de corte. Considerando-se a seleção aleatória de um indivíduo doente e um não-doente, a AUC é interpretada como a probabilidade de o indivíduo doente ter um resultado de teste de diagnóstico com uma magnitude maior do que a de um não doente. Um teste incapaz de discriminar entre o estado de saúde dos dois - doente e não doente - teria uma AUC igual a $0,5$. Quanto maior a capacidade do teste para distinguir os indivíduos desses dois grupos, mais perto a curva ROC se situará do canto superior esquerdo do plano. Uma das primeiras propostas para comparar duas curvas ROC, utilizando a AUC, foi apresentada por Hanley e McNeil [9] e baseia-se na relação crítica Z.

2. RESULTADOS POR NUTS II E EXEMPLO DE OUTPUT

Apresentam-se as taxas de mortalidade obtidas com base na classificação atribuída pelas duas escalas aos recém-nascidos (Tabela 1), nas UCIN incluídas em cada NUTS II e, a título meramente ilustrativo, nas figuras 1 e 2 um exemplo de output de cada software utilizado no estudo, IBM® SPSS® Statistics e Comp2ROC, respetivamente.

	Taxas de Mortalidade					
	Norte	Centro	Lisboa/Vale do Tejo	Alentejo	Algarve	Ilhas
CRIB	9.96%	9.34%	12.34%	6.05%	7.47%	16.90%
SNAPPE II	8.76%	6.93%	12.36%	6.28%	7.47%	14.29%

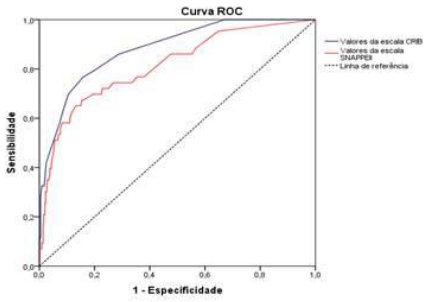
Tabela 1: Taxas de mortalidade por NUTS II.

3. CONCLUSÕES

Para o exemplo apresentado, e atendendo às Figuras 1 e 2, verifica-se existir diferença estatisticamente significativa no desempenho das duas escalas de avaliação da gravidade clínica dos recém-nascidos.

AGRADECIMENTOS

As autoras gostariam de agradecer ao Registo Nacional de Muito Baixo Peso pela disponibilização dos dados.
Este trabalho foi financiado pela FCT - Fundação para a Ciência e Tecnologia dentro do projeto Scope:UID/CEC/00319/2013.



	AUC	DP	Z	Valor p	Decisão
CRIB	0,881	0,026	2,46	0,0139	Diferença estatisticamente significativa
SNAPPE II	0,814	0,037			

Figura 1: Curva ROC comparativa nas UCIN da região Norte (output SPSS).

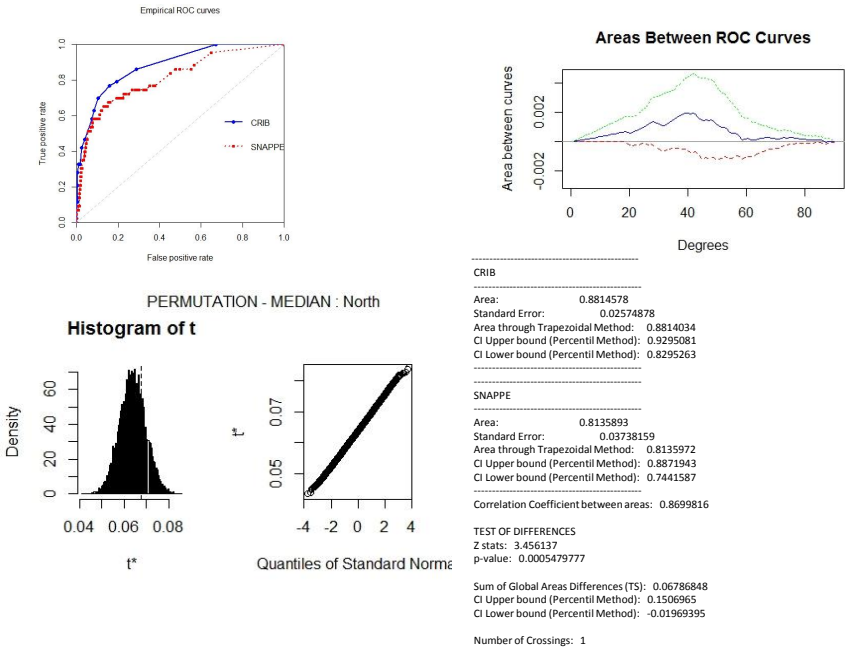


Figura 2: Comparação entre CRIB e SNAPPE II nas UCIN da região Norte (output Comp2ROC).

Referências

[1] Braga A. C., Costa L., Oliveira P.(2013). An alternative method for global and partial comparison of two diagnostic system based on ROC curves. *Journal of Statistical Computation and Simulation* 83, 307–325.

[2] Frade H., Braga A. C.(2013). Comp2ROC: R package to compare two ROC curves. *7th International Conference on Practical Applications of Computational Biology & Bioinformatics* 222, 127–135.

[3] Braga A. C.(2015). Compare two ROC curves that intersect. Repository CRAN.

[4] DeLong E. R., DeLong D. M., Clarke-Pearson D. L.(1988). Comparing the areas under two or more correlated receiver operating characteristic curves:a nonparametric approach. *Biometrics* 44, 837–845.

[5] Dorfman D. D., Jr. E. A.(1969). Maximum-likelihood estimation of parameters of signal-detection of confidence intervals-Rating-method data. *Journal of Mathematical Psychology* 6, 487–496.

[6] Dorling J. S., Field D. J., Manktelow B.(2005). Neonatal disease severity scoring system. Archive of Disease Childood. *Fetal and Neonatal Ed.* 90, F11–F16.

[7] Egan J. P.(1975). *Signal Detection Theory and ROC Analysis.* New York:Academic Press.

[8] Green D. M., Swets J. A.(1974). *Signal Detection Theory and Psychophysics(rev.ed).* Robert E. Krieger Publishing Company.

[9] Hanley J. A., McNeill B. J.(1982). The meaning and use of the area under receiver operating characteristic (ROC) curve. *Radiology* 143, 29–36.

[10] Hanley J. A., McNeill B. J.(1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 3, 839–843.

[11] Metz, C. E.(1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 8,283–298.

[12] Metz C. E.(1984). Statistical Analysis of ROC Data in Evaluating Diagnostic Performance. Multiple Regression Analysis: Applications in the Health Sciences, D. E. Herbert & R. H. myers (Eds), *American Institute of Physics*, 365–384.

[13] Swets J. A., Pickett R. M.(1996). *Signal detection theory and ROC analysis in Psychology and diagnostic: collected papers.* Lawrence Erlbaum Associates,Inc.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Disfunção cardíaca em ratos: Uma análise de bayesina

Magali T. R. Amaral[1], Maria C, Evangelista [2] e Antônio C, Cicogna [3], Carlos R. Padovani [4]

[1] Departamento de Bioestatística, Programa de pós-graduação em Biometria-UNESP, Universidade Estadual de Feira de Santana - UEFS.
[2,4] Departamento de Bioestatística, Programa de pós-graduação em Biometria-UNESP.
[3] Departamento de Clínica Médica, Faculdade de Medicina de Botucatu-UNESP.

RESUMO

O presente estudo tem como principal objetivo verificar por meio do método bayesiano a existência de diferença significativa na tensão máxima desenvolvida (TD) do músculo papilar isolado entre os diferentes níveis de doses de cálcio administrados, bem como quantificar a eficiência média cardíaca dos animais submetidos à cirurgia para indução de Estenose Aórtica supravalvar (EAo) com dieta rica em gorduras saturadas (hipercalórica). Para esse experimento, foram considerados o conjunto de dados que consiste nas medidas de 47 ratos Wistar machos, com 21 dias de idade, provenientes do Biotério do Laboratório Experimental do Departamento da Clínica Médica, Faculdade de Medicina de Botucatu. Conclui-se que o perfil médio de resposta da TD apresenta relação com os níveis de cálcio, os quais mostram que os níveis mais baixos são mais expressivos na diferenciação.

Palavras-chave: Método bayesiano, Estenose aórtica supravalvar.

1. INTRODUÇÃO

A cirurgia para indução de Estenose Aórtica supravalvar (EAo) é um procedimento que consiste em submeter os animais a toracotomia mediana, sendo a aorta ascendente dissecada e um clipe de prata, com 0.6 mm de diâmetro interno, colocado aproximadamente a 3 mm da sua raiz. No presente estudo, após a cirurgia os animais receberam, uma dieta rica em gorduras saturadas (hipercalórica).
Considerou-se a dieta rica em gorduras saturadas pois em trabalhos anteriores verificou-se que a administração de ácidos graxos no modelo de estenose aórtica juntamente com dieta hiperlípidica atenua a queda do desempenho do ventrículo esquerdo. Essa dieta foi elaborada pelo grupo de Pesquisa Experimental de Músculo Papilar Isolado do Departamento de Clínica Médica, Faculdade de Medicina de Botucatu, UNESP, sendo a manipulação realizada juntamente com a empresa Biotron Zootecnia, Rio Claro -SP [3].
O experimento consiste na manobra de cálcio com a elevação em sequência de 0.5 mM, a partir do nível basal de 0.5mM até 3.5 mM, totalizando sete doses $(0{,}5, 1, 1{,}5, 2{,}0, 2{,}5, 3{,}0)$, logo após a cirurgia. Porém, antes de submeter os animais a dieta, realiza-se um ecocardiograma resultando em 33 variáveis que oferecem importantes informações sobre o desempenho cardíaco do animal. Estas variáveis permitem avaliar de acordo com [7] como é a morfologia e a função do coração; acompanhar a evolução da disfunção cardíaca causada por diferentes graus de agressão e avaliar os efeitos de diferentes intervenções sobre o coração.

2.MODELO PROPOSTO

O caso mais simples de análise de variância (ANOVA), consiste em mensurações que são classificadas de acordo com um critério unidimensional no qual um único fator é subdividido em tratamentos (níveis do fator). Esse fator pode ser classificado como fixo ou aleatório. Para um fator fixo o modelo tem a seguinte estrutura:

$$y_{g,i} = \mu_g + \omega_{g,i} + \varepsilon \tag{1}$$

em que, $y_{g,i}$ é a i-ésima observação pertencente ao g-ésimo nível de fator, μ_g é a média de cada g-ésimo nível como constantes fixas e os valores $\omega_{g,i}$ como permutáveis com distribuição normal $N(0, \sigma)$[5]. Assim a análise consiste em quantificar a resposta média de cada combinação de níveis categóricos.
Para a aplicação da técnica no modelo proposto são requeridas algumas estatísticas:

$$SQ_g = \sum_{i=1}^{n_g} (y_{g,i} - \bar{y}_g)^2 \tag{2}$$

$$S_e^2 = \frac{\sum_{g=1}^{G} SQ_{SQ_g}}{\sum_{g=1}^{G} n_g - G} \tag{3}$$

em que SQ_g é a soma dos quadrados dos resíduos, S_e^2 é a variância das estimativas e $\bar{y}_g$ é a médias amostral do g-ésimo fator.

0.1. Método de Jeffreys

A densidade a priori de Jeffreys é definida por uma densidade proporcional a informação de Fisher [8], ou seja:

$$\pi(\theta) \propto |I(\theta)|^{\frac{1}{2}}; \tag{4}$$

em que θ é o parâmetro de interesse.
Considere o modelo 3.1 apresentado acima, o conjunto de medidas encontrado na TD e os níveis de cálcio (doses), $D = \{(y_{ig}; i = 1 \ldots, n_g, g = 1, \ldots, G)$, pode-se reescrever a equação como:

$$\ln y_{g,i} = \ln(\mu_g + \omega_{g,i} + e_g) \tag{5}$$

Considerando os erros independentes e identicamente distribuídos $e_g \sim N(0, \sigma^2)$, tem-se que $p(\mu_1, \mu_2, \ldots, \mu_G, \sigma) = \frac{1}{\sigma^2}$ é a priori não informativa de Jeffreys e $p(y_{g,i}|\mu_1, \mu_2, \ldots, \mu_G, \sigma)$ tem distribuição normal $N(\mu_g, \sigma), g = 1, \ldots, G$ e $i = 1, \ldots, n_g$, então:

$$p(\mu_g|D) \sim St(n - G, \overline{y}, \frac{S_e}{\sqrt{n_g}});$$

$$p(\sigma^2|D) \sim GInv(\frac{n - G}{2}, \frac{n - G}{2} S_e^2);$$

em que, S_e^2 expressa a variância dos resíduos.
Assim, por meio das amostras simuladas pode-se obter as distribuições posteriores marginais como segue:

1. Simular valores de $\tau_{(l)}$ para $l = 1, 2, \ldots, m$ da distribuição $Gamma(\frac{n-G}{2}, \frac{n-G}{2} S_e^2)$.

2. Calcular as m variâncias $\sigma_{(l)}^2$ que tem distribuição gama-inversa e correspondem ao inverso da precisão, ou seja,$\sigma_{(l)}^2 = \frac{1}{\tau_{(l)}}$.

3. Simular os vetores $(\mu_1^{(l)}, \ldots, \beta_G^{(l)})$ de distribuições normais G-dimensionais denotadas por $N_G((\bar{y}_1, \ldots, \bar{y}_G), \sigma_{(l)}^2 V_\mu)$, em que, $V_\mu = diag(\frac{1}{n_1}, \ldots, \frac{1}{n_G})$ é uma matriz diagonal.

3. RESULTADOS

A estatística a posteriori da dose de cálcio, obtida após a simulação é apresentada na tabela 1 e por meio dos resultados apresentados pode-se notar que os intervalos de credibilidade foram todos significantes. Para fins de comparação, observe que as médias da TD cresceu de acordo com a dose de cálcio administrada e que os seus respectivos intervalos de credibilidade sucessivos se sobrepõem quase que completamente, indicando que não se pode constatar diferenças médias significativas em intervalos de aplicação sequencial de 0,5 mg de cálcio.

Tabela 1: Medidas descritivas das densidades a posteriori da dose de cálcio

	Cálcio						
	Dose 0.5	Dose 1	Dose 1.5	Dose 2	Dose 2.5	Dose 3	Dose 3.5
1^{o} Quartil	4.34	4.81	5.31	5.58	5.84	6.05	6.20
Mediana	4.44	4.90	5.40	5.68	5.93	6.14	6.29
Média	4.44	4.90	5.40	5.68	5.94	6.14	6.26
3^{o} Quartil	4.53	5.006	5.50	5.78	5.94	6.14	6.29
Inter. de credibilidade	[4.17; 4.72]	[4.63; 5.18]	[5.12;5.66]	[5.41; 5.96]	[5.67; 6.21]	[5.86; 6.42]	[6.01; 6.56]

A figura 1 representa os histogramas dos parâmetros da TD nas diferentes manobras de cálcio. A inspeção visual mostra simetria em todos os níveis de cálcio.

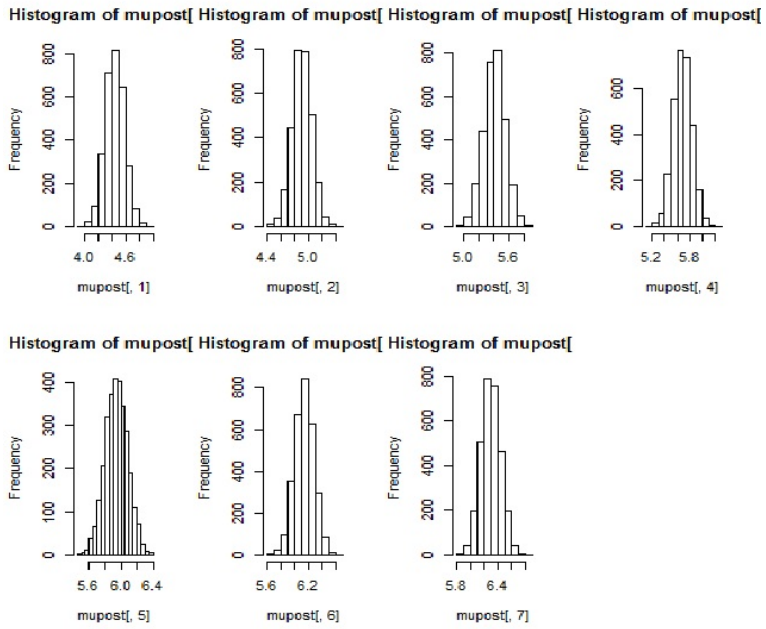


Figura 1: Histograma dos parâmetros da TD para os diferentes níveis de cálcio.

4. CONCLUSÕES

De acordo com o resultado obtido, pode-se inferir que entre os níveis consecutivos de cálcio não se verifica resposta diferenciada da TD ($p > 0{,}05$). Porém para os níveis mais baixos ($\leq 2{,}0mM$), se considerar um salto de 1.0 mM de cálcio há diferenciação significativas na respostas médias da TD, situação não configurada nos níveis acima de 2.0 mM. No contexto do perfil médio da resposta da TD em função dos níveis de cálcio, os resultados indicam uma taxa de crescimento mais elevada nos níveis iniciais seguida de uma estabilização da variação a partir do nível 2.0 mM. Esse indicativo pode ser utilizado na eventualidade de interesse de um modelo preditivo para a TD.

AGRADECIMENTOS

A Pró-reitoria de Pós-graduação da Unesp (PROPG) pelo apoio financeiro.

REFERÊNCIAS

Referencias

[1] BOX, G. E. P; TIAO, G. C., *Bayesian inference in statistical analysis. J. Wiley, New York, NY, EUA, 1973.*

[2] CHIB, S.; GREENBERG, E., Understanding the Metropolis-Hastings Algorithm. *American Statistician, 49 (4), 327-335, 1995.*

[3] Evangelista, M. C., *Construção de curvas polinomiais para a resposta funcional do músculo papilar isolado em diferentes concentrações de cálcio. UNESP, 2016*

[4] GEWEKE, J., *Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bayesian Statistics 4ed., JM Bernado, JO Berger, AP Dawid and AFM Smith). Clarendon Press, Oxford, UK, 1992.*

[5] ANDRADE H. A.; KINAS, P. G. *Inferência em ecologia: comentários e um exemplo. 2008*

[6] JIM, A., *Bayesian Computation with R: Department of Mathematics and Statistics. 3ed., Springer, New York, NY, USA, 2008.*

[7] Moreira, V. O. *Critério ecocardiográficos para definição de graus de disfunção ventricular em ratos portadores de estenose aótica, 2006.*

[8] Paulino, C.D., Amaral Turkman, A., Murteira, B. (2003). *Estatística Bayesiana.* Fundação Calouste Gulbenkian, Lisboa.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

EFECTO DEL LLENADO DE LA VEJIGA EN TRATAMIENTOS GINECOLÓGICOS DE BRAQUITERAPIA DE ALTA TASA

Z Martín-Rodríguez[1], E Ambroa-Rey[2], MJ Ginzo-Villamayor[3] E Gónzalez-Patiño[4], J Mosquera-Seoane[1] y M Pombar-Cameán[1]

[1] Servicio de Radiofísica y Protección Radiológica, Complejo Hospitalario Universitario de Santiago de Compostela,
[2] Servicio de Radiofísica, Hospital General de Cataluña, Barcelona,
[3] Servicio de Consultoría Estadística. Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela
[4] Servicio de Oncología Radioterápica, Complejo Hospitalario Universitario de Santiago de Compostela

RESUMEN

En este estudio se ha investigado el impacto dosimétrico sobre los órganos de riesgo cuando se procede al llenado de la vejiga para el tratamiento intracavitario vaginal con braquiterapia de alta tasa, frente al tradicional tratamiento con la vejiga vacía.

Palabras y frases clave: llenado vejiga, dosimetría, órganos de riesgo.

1. INTRODUCCIÓN

Actualmente, la braquiterapia vaginal post-operatoria (como tratamiento único o en combinación con radioterapia externa) es el procedimiento más frecuente en las Unidades de Braquiterapia. En el cáncer de endometrio, tres importantes estudios aleatorizados fase III [1-3] han comprobado que las recaídas más frecuentes en pacientes no irradiadas se produce en el fondo vaginal.
Sólo unos pocos estudios han sido publicados sobre el efecto del llenado de la vejiga en la distribución de dosis en los órganos de riesgo [4,5]. El objetivo de este trabajo es evaluar las diferencias dosimétricas en los órganos de riesgo entre los estudios con la vejiga llena y vacía para braquiterapia vaginal HDR y así determinar el mejor modo de tratamiento.

2. MÉTODOS Y MATERIALES

A. Pacientes.
Desde Mayo 2012 hasta Octubre 2014, un total de 33 pacientes consecutivas con una media de edad de 65,1 años (41-89) fueron tratadas con braquiterapia vaginal intracavitaria. De las cuales, 30 presentaban cáncer de endometrio (estadio I: 43%) y el resto cáncer de cérvix operado. La mitad fueron tratadas con radioterapia externa y braquiterapia y el resto con braquiterapia como tratamiento exclusivo (30%).

B. Radioterapia.
Todas las pacientes fueron sometidas a braquiterapia vaginal intracavitaria con aplicadores cilíndricos de un sólo canal para una fuente de 192Ir (Nucletron, Veenendaal, The Netherlands). El tamaño del cilindro de tratamiento oscila entre 2,5 a 3,5 cm y su longitud de 2,5 a 4,0 cm. La prescripción de dosis varía entre 530cGy y 750cGy por fracción, para tratamientos entre 3 y 6 fracciones.

C. Planificación basada en tomografía computerizada.
Cada paciente se somete a dos estudios CT consecutivos: primero, se realiza con la vejiga completamente vacía, el segundo se procede a llenar solamente con 48 ml de salino y 2 ml de contraste, total 50 ml (vejiga llena). Ambos estudios se llevan a cabo en la misma localización para preservar la posición del aplicador.
El delineado del PTV y de los órganos de riesgo fue realizado por el mismo oncólogo radioterápico. La planificación fue llevada a cabo por un único radiofísico tanto para la vejiga llena como vacía, utilizando los mismos puntos de prescripción y optimización con un planificador Oncentra® .
Para cada paciente, se tomaron los siguientes datos para cada uno de los órganos de riesgo contemplados: D_{2cc}(%), D_{1cc}(%), $D_{0,1cc}$(%), así como el volumen total y el punto ICRU y los volúmenes $V_{50\%}$(%) y $V_{55\%}$(%) para la vejiga y el recto, respectivamente. Y $D_{0,1cc}$(%) y $V_{20\%}$(%) para la uretra porque están recogidos en la literatura como los volúmenes que presentan una correlación con efectos secundarios tardíos para la vejiga y el recto en cáncer de vejiga recomendados por GEC-ESTRO.

D. Análisis estadístico.
Un test de Wilcoxon no paramétrico se ha llevado a cabo para estudiar la diferencia de las medianas de nuestra muestra pareada. De esta manera, observar la incidencia de del llenado de la vejiga con 50 ml de agua frente a la vejiga llena. Además, se llevaron a cabo correlaciones lineales de Pearson. Dicho cálculo permite determinar la relación entre las variables: los volúmenes dosimétricos anteriormente definidos y los parámetros propios del tratamiento: el diámetro del cilindro, longitud de tratamiento, prescripción y número de fracciones. Se considerará que hay correlación lineal para valores superiores a 0,7 en valor absoluto. Además, realizamos un test ANOVA con un nivel de significación alpha= 0.05, para investigar la posible relación entre las mismas variables. Ambos estudios, se han desarrollado utilizando el paquete computacional para estadística R.

3. RESULTADOS

La Tabla 1 ilustra la media y el rango obtenido en el análisis estadístico descriptivo para los tres órganos de riesgos contemplados: vejiga, recto y uretra para ambos casos. Se observa un incremento notable del volumen de la vejiga cuando los estados lleno y vacío se comparan. El volumen del 50% resulta ser más alto en promedio para el caso de la vejiga vacía en comparación con la llena. Por otro lado, las diferencias de las medias en ambos casos son más relevantes cuando atendemos a D2cc (%) que para volúmenes más pequeños. Estos mismos resultados se observan en el caso del recto y la uretra, también recogidos en la misma tabla.

Vejiga	**Llena (rango)**	**Vacía (rango)**
Volume (cm3)	121,11 (180,14-79,94)	62,63 (122,86-0,00)
$V_{50\%}$(%)	1,82 (10,24-0,00)	3,00 (13,54-0,00)
$D_{0,1cc}$ (%)	60,21 (79,94-39,74)	59,80 (85,41-23,55)
D_{1cc} (%)	50,36 (66,64-32,40)	48,33 (72,14-18,95)
D_{2cc} (%)	45,78 (62,62-29,10)	43,31(67,50-16,83)
Recto	**Llena (rango)**	**Vacía (rango)**
$V_{55\%}$(%)	5,98 (15,02-0,00)	7,28 (26,00-0,01)
$D_{0,1cc}$ (%)	80,85 (102,9-45,50)	80,27 (98,24-47,04)
D_{1cc} (%)	66,34 (84,00-36,80)	66,00 (80,97-37,07)
D_{2cc} (%)	59,27 (78,99-32,41)	58,93 (77,75-32,13)
Uretra	**Llena (rango)**	**Vacía (rango)**
$D_{0,1cc}$ (%)	36,48 (100,01-11,60)	36,39 (100,01-5,94)
$V_{20\%}$(%)	39,29 (100,01- 0,01)	40,35 (100,01-0,01)

Tabla 1: Medias y rangos para los volúmenes dosimétricos estudiados para los tres órganos de riesgo. D_{2cc}, D_{1cc} ,$D_{0,1cc}$ es la dosis máxima para un volumen de 2 cm3 , 1 cm3 , 0,1 cm3 en la vejiga o

recto, respectivamente. $V_{50\%}$, $V_{55\%}$, $V_{20\%}$ es la dosis para el 50%, 55% o 20 % del volumen para la vejiga, el recto o la uretra, respectivamente.

Las Tablas 2 y 3 resumen los resultados obtenidos para el estudio de las correlaciones lineales. Como podemos observar, para todas las comparaciones llevadas a cabo se han obtenido valores que no se acercan al valor considerado de 0,7 para que exista correlación. Igualmente, resultados similares se obtuvieron para el estudio de la ANOVA para las mismas variables. D diámetro, L longitud, P prescripción y N número de sesiones.

Vejiga	**Llena**			
	Coeficiente de correlación			
	D (cm)	**L (cm)**	**P (cGy)**	**N**
$V_{50\%}$(%)	0,09	0,07	0,14	0,13
$D_{0,1cc}$(%)	0,00	-0,19	-0,11	0,08
D_{1cc}(%)	0,12	-0,10	0,03	0,15
D_{2cc}(%)	0,15	-0,07	0,08	0,17
Recto	**Llena**			
	Coeficiente de correlación			
	D (cm)	**L (cm)**	**P (cGy)**	**N**
$V_{55\%}$(%)	0,18	0,07	0,14	0,00
$D_{0,1cc}$(%)	0,17	0,23	0,1	0,27
D_{1cc}(%)	0,28	0,26	0,19	0,34
D_{2cc}(%)	0,31	0,25	0,22	0,35
Uretra	**Llena**			
	Coeficiente de correlación			
	D (cm)	**L (cm)**	**P (cGy)**	**N**
$D_{0,1cc}$(%)	0,13	0,28	0,04	0,36
$V_{20\%}$(%)	0,12	0,23	0,18	0,16

Tabla 2. Coeficientes de correlación para las correlaciones lineales de los volúmenes dosimétricos con el diámetro y longitud del cilindro, prescripción y números de sesiones para la vejiga llena.

Vejiga	**Vacía**			
	Coeficiente de correlación			
	D (cm)	**L (cm)**	**P (cGy)**	**N**
$V_{50\%}$(%)	0,15	0,14	0,12	0,13
$D_{0,1cc}$(%)	0,53	0,47	0,48	0,36
D_{1cc}(%)	0,54	0,42	0,48	0,36
D_{2cc}(%)	0,54	0,41	0,48	0,36
Recto	**Vacía**			
	Coeficiente de correlación			
	D (cm)	**L (cm)**	**P (cGy)**	**N**
$V_{55\%}$(%)	0,15	0,08	0,19	0,04
$D_{0,1cc}$(%)	0,13	0,20	0,08	0,10

D_{1cc}(%)	0,23	0,23	0,18	0,17
D_{2cc}(%)	0,23	0,23	0,18	0,17

Uretra	**Vacía**			
	Coeficiente de correlación			
	D (cm)	**L (cm)**	**P (cGy)**	**N**
$D_{0,1cc}$(%)	0,11	0,25	0,02	0,30
$V_{20\%}$(%)	0,11	0,27	0,1	0,13

Tabla 3. Coeficientes de correlación para las correlaciones lineales de los volúmenes dosimétricos con el diámetro y longitud del cilindro, prescripción y números de sesiones para la vejiga vacía. D diámetro, L longitud, P prescripción y N número de sesiones.

4. CONCLUSIONES

De acuerdo con los resultados obtenidos, en los tratamientos de braquiterapia vaginal con un llenado de la vejiga con tan sólo 50 ml, no se observa ninguna desventaja frente al tratamiento con vejiga vacía, desde el punto de vista dosimétrico sobre los órganos de riesgo valorados.
La estimación de la dosis en tratamientos con vejiga vacía posee una incertidumbre más alta debido a la dificultad para su delineación sin contraste. Por lo que consideramos más conveniente realizar los tratamientos de braquiterapia vaginal de 50 ml. Además, permite una mejor valoración de los histogramas dosis-volumen.

AGRADECIMIENTOS

La investigadora Mª José Ginzo agradece el apoyo de los proxectos 2013-PG064 de la Consellería de Educación e Ordenación Universitaria, al MTM2013-41383-P del Ministerio de Economía y Competitividad y a la Red IAP P7/06 de Belgian Science Policy.

Referencias

[1] Creutzberg CL, van Putten WLJ, Koper PCM, et al, for the PORTEC Study Group. Surgery and postoperative radiotherapy versus surgery alone for patients with stage-1 endometrial carcinoma: multicentre randomized trial. Lancet 2000; 355: 1404–11.

[2] Scholten AN, van Putten WL, Beerman H, et al. Postoperative radiotherapy for Stage 1 endometrial carcinoma: long-term outcome of the randomized PORTEC trial with central pathology review. *Int J Radiat Oncol Biol Phys* 2005; 63: 834–38.1

[3] Keys HM, Roberts JA, Brunetto VL, et al. A phase III trial of surgery with or without adjunctive external pelvic radiation therapy in intermediate risk endometrial adenocarcinoma: a Gynecologic Oncology Group study. *Gynecol Oncol* 2004; 92: 744–51.

[4] Stewart AJ, Cormack RA, Lee H, et al. A prospective clinical trial of bladder filling and three-dimensional dosimetry in high-dose-rate vaginal cuff brachytherapy. *Int J Radiation Oncology Biol Phys* 72(3): 843-848, 2008.

[5] Adli M, Garipagaoglu M, Kocak Z et al. Effect of bladder distention on bladder base dose in gynaecological intracavitary high dose rate brachytherapy. *The British Journal of Radiology*, 82 dose rate brachy (2009), 243–248.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

IMPACT OF WORSENING RENAL FUNCTION IN SURVIVAL IN CRITICAL PATIENTS

A. Baluja[1,2], M. López Ratón[3], P. Otero Castro[1,2], A. Iglesias[1], J. Álvarez Escudero[1,2] y C. Cadarso Suárez[3]

[1]Servicio de Anestesiología y Reanimación. Hospital Clínico Universitario de Santiago de Compostela, España.
[2]Grupo Enfermo Crítico. Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS).
[3]Unidade de Bioestadística, Facultade de Medicina. Universidad de Santiago de Compostela, España.

SUMMARY

Renal function is often associated with poor outcome in intensive care units. Retrospective, cohort study involving patients admitted to an intensive care unit (ICU). Acute and chronic kidney damage, risk factors and mortality were evaluated. Flexible regression modelling showed that the differences in creatinine changes from baseline may be useful for survival assessment in critical patients.

Keywords: renal, creatinine, intensive care, survival.

1. BACKGROUND

Renal function has long been considered one of the main signs associated with poor outcome in intensive care units. Despite significant improvements during the last years in intensive patient support, mortality rates associated with impaired kidney function are still high. In this study we aim to elucidate the role of kidney function changes on the mortality of patients admittted to an intensive care unit.

2. METHODS AND RESULTS

Retrospective observational cohort design with N=16 patients admitted to an intensive care unit. Basic demographic data was collected and kidney function and survival was also assessed by means of changes in creatinine levels. A total number of 434 blood samples were collected for the sum of all patients during all the follow-up in the ICU. Mortality was modelled on the basis of age and other risk factors. Statistical analyses were done with R v.3.2.2 (library `mgcv`).
The median age was 78 years, the median ICU stay was 13.5 days, and the readmission rate was 31% (5 patients). The survival rate was 56% (9 patients). Logistic GAM multivariate modelling showed the differences in creatinine changes from baseline during ICU stay for survivors and non-survivors.

3. CONCLUSIONS

Creatinine changes from baseline may be useful for survival modelling in critical patients.

	exp(Estimate)	Std. Error	p-value
Age	1.01	0.01	0.23
Creatinine (mg/dL)	0.77	0.12	0.03

Table 1: GAM model of mortality risk adjusted by age and serum creatinine.

References

[1] Bellomo R, Ronco C, et al. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. (2004). *Crit Care* 8:R204-R212.

[2] SM Bagshaw, S Uchino, JA Kellum et al. Association between renal replacement therapy in critically ill patients with severe acute kidney injury and mortality. (2013). *J Crit Care* 28(6):1011-8.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

INTERFAZ PARA EL CÁLCULO AUTOMÁTICO DE BLINDAJES EN BRAQUITERAPIA

Z Martín-Rodríguez[1], I Taboada-Torres[1], M Giménez-Insua[1], J Guiu-Souto[1], M García-Sánchez[1], C Otero-Martínez[1], V Luna-Vega[1], M Pombar-Cameán[1] y Mª José Ginzo-Villamayor[2]

[1] Servicio de Radiofísica y Protección Radiológica, Hospital Universitario de Santiago de Compostela
[2] Servicio de Consultoría Estadística. Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela

RESUMEN

En este trabajo se ha desarrollado una interfaz en MATLAB que permite la automatización del cálculo y/o comprobación de blindajes en braquiterapia.

Palabras y frases clave: Automatización, dosis secundaria, puerta búnker.

1. INTRODUCCIÓN

Una de las labores primordiales del radiofísico hospitalario es el adecuado cálculo de blindajes en el diseño de instalaciones radiativas. Decisivo para el cumplimiento del principio ALARA ("As low as posible achievable") [1] para la disminución de las radiaciones ionizante recibidas tanto por personal sanitario como por los miembros del público. El cálculo de blindajes es una labor infrecuente ya que se realiza, ocasionalmente, coincidiendo con la aceptación de nuevos equipos. En la actualidad, este cómputo se lleva a cabo manualmente con la ayuda de los planos de las instalaciones sobre los que se miden las distancias a los diferentes puntos de interés de cálculo desde la fuente radiativa. Tarea en sí, laboriosa y proclive a errores.
Es por ello, que en este trabajo se ha desarrollado una interfaz gráfica para la automatización del cálculo de blindajes. Esta interfaz está provista de una característica innovadora que permite la selección de los diversos puntos de medida sobre el mismo plano de la instalación radiativa, así como de los diversos parámetros requeridos para el cómputo de los blindajes. Además, está dotada de ayudas que asisten al usuario en la elección de los mismos.
En la literatura, se menciona escasamente el cálculo de la radiación dispersa en braquiterapia. Recientemente, se ha desarrollado un estudio en el que se trata de adaptar la metodología de la NCRP 151 [2] para radioterapia externa a braquiterapia [3]. Este trabajo calcula la tasa de dosis en la puerta del laberinto para una serie de diseños de instalaciones radiativas en braquiterapia. Tiene en cuenta tres contribuciones, a saber: la transmisión directa de la pared del laberinto, la transmisión del área opuesta a la puerta que se ve desde la misma y por último, la transmisión del resto de las paredes. El estudio fue corroborado comparando con simulaciones Monte Carlo para cada diseño de instalación radiativa considerado; demostrando excelente acuerdo con los resultados sólo en el caso de diseños más convencionales. De esta manera, el cómputo de la radiación dispersa en la entrada del laberinto se introduce en este estudio.

Este trabajo se organiza de la siguiente manera: en la Sección 2 se realiza una descripción de los métodos seguidos, a continuación, la Sección 3 recoge los resultados obtenidos y finalmente, la Sección 4 expone las conclusiones logradas en el trabajo.

2. MÉTODOS Y MATERIALES

Esta interfaz gráfica ha sido desarrollada utilizando la herramienta Guide de MATLAB (Figura1). Esta herramienta permite la definición gráfica de menús interactivos para la elección de los parámetros empleados en el cálculo y/o comprobación de blindajes. La interfaz posibilita la obtención del espesor de los blindajes, en función de la carga de trabajo [5], la distancia al punto de medida, el límite de dosis, el factor de ocupación y el factor de uso.

Para ello, se ha llevado cabo la implementación de la expresión analítica que ajusta valores de transmisión obtenidos con Monte Carlo para cada uno de los radioisótopos y materiales empleados en las barreras de blindajes [6]. Asimismo, implementa la expresión para el cálculo de la radiación dispersa en la puerta como la contribución de la transmisión directa y del área opuesta a la puerta del laberinto que se ve desde la misma. La tercera de las contribuciones [3], no se incluye ya que para las energías de los isótopos con los que se trabaja en braquiterapia, se ha considerado despreciable.

El uso del interfaz consiste en buscar el plano de la instalación radiativa comenzando por el propio directorio. A su vez, se cargan el resto de las imágenes ya que son susceptibles de ser utilizadas para cálculos posteriores. El plano es un imagen en píxeles y debe ser calibrada para llevar a cabo el resto de las medidas de la distancia al punto de interés de cálculo. Para ello, elegimos dos puntos de distancia conocida sobre el plano. Para el cómputo del factor de transmisión se requiere: la carga de trabajo, la distancia, el límite de dosis anual, el factor de ocupación y el factor de uso. Este último es uno, debido a que las fuentes en braquiterapia no están colimadas [7]. La carga de trabajo viene dada como el producto de la TKRA, el número de tratamientos semanal y el tiempo de radiación para cada uno de ellos. Estos datos son introducidos manualmente por el usuario siguiendo las instrucciones que aparecen en el apartado de información y estado de la interfaz. Esta carga de trabajo es única ya que todas las paredes de una instalación radiativa en braquiterapia son barreras primarias [5]. A continuación, se posiciona la fuente radiativa sobre el plano, a partir de la que miden automáticamente todas las distancias a los puntos de interés de medida. La elección de los puntos de medida es un proceso iterativo. El proceso comienza con la selección de un punto detrás de una barrera primaria en el plano.

Entonces, se escogen tanto el límite anual de dosis, el factor de ocupación y el de uso de los menús disponibles.

Y así, sucesivamente. En segundo lugar, se toman puntos en la puerta para la estimación de la tasa de dosis en la puerta del laberinto. Se tendrán en cuenta dos de los tres términos introducidos en el estudio de Pujades et al. [3], ya que hemos considerado la tercera contribución desestimable para las energías de los radioisótopos tratados en braquiterapia. Las aquí consideradas corresponden a la transmisión directa a través de la pared del laberinto y la contribución a la pared opuesta que se ve desde la puerta. La tercera corresponde a la contribución debida al resto de las paredes. Los factores de dispersión de la pared para el 192Ir y 60Co [3] fueron obtenidos como la interpolación lineal de los valores que aparecen en las tablas 8.a y 8.b del apéndice B del NCRP 151 para energías nominales de 380 y 1250 keV, respectivamente.

En la actualidad, no hay datos publicados en la literatura sobre los factores de dispersión para el resto de los radioisótopos aquí utilizados. Para los mismos, se toma 0,2 que corresponde al tanto por ciento de la tasa de dosis absorbida debido a la radiación incidente dispersa a 1 metro irradiada por un área de 400 cm2 de un maniquí fabricado con material que se asemeja al tejido humano [4]. Tradicionalmente, este valor es el considerado para el cálculo de radiación dispersa por tratarse del caso más desfavorable.

Igualmente, se escogen los valores correspondientes para el límite de dosis, el factor de ocupación y uso de los menús disponibles.

La interfaz proporciona la posibilidad de editar o borrar alguno de los puntos elegidos en el plano, de manera que se renumeran los puntos ya existentes en el plano y se actualiza la matriz que acumula el límite de dosis anual, el factor de ocupación y uso para cada uno de los puntos.

Finalmente, se procede al cálculo de los blindajes para cada uno de los materiales o la tasa de dosis con barrera, dependiendo de si se trata del cálculo de blindajes o la comprobación de los mismos. En el primero de los casos, se calculan los factores de transmisión necesarios, así como los blindajes para cada uno de los materiales. En el segundo, se calculan los factores de transmisión obtenidos así como la tasa de dosis con barrera. Los resultados obtenidos son exportados a Excel dinámicamente y por orden de acuerdo con el número que tienen en el plano, junto con la distancia, el tipo de barrera, el límite de dosis, el factor de ocupación y uso. Asimismo, la plantilla recopila datos relevantes de la fuente radiativa, así como su estimación de uso.

3. RESULTADOS

Los resultados computados automáticamente muestran una excelente concordancia con aquellos calculados manualmente mediante la aplicación directa de la metodología anteriormente descrita.

Figura 1: Interfaz

Cuando se recogen datos de una medida experimental, éstos se ven afectos por falta de precisión en sus valores por varias causas, unas debidas a las imperfecciones del aparato de medida pero otras debidas a la forma de registrar esas medidas.

En este trabajo para determinar con el fin de determinar el error relativo, se ha implementado además un algoritmo que simula un usuario seleccionando puntos en el plano. Para ello se generaron diez mil muestras normales, como si fueran 10000 selecciones con el ratón de media el dato de una posición media dada alrededor de una fuente radiactiva o de cualquiera de los puntos de interés [3]. Las simulaciones se realizaron para los diez mil puntos alrededor de las localizaciones dadas para calcular la dispersión de las mediciones. Para la simulación se eligieron normales con desviación estándar 0.05, 0.025 y 0.250.

Entonces se calcularon las tasas de dosis directas y dispersas en el laberinto entrada para cada instalación utilizando dicho algoritmo y la interfaz y se compararon con lo de la literatura [3]. El algoritmo de simulación fue desarrollado con el paquete estadístico R.

4. CONCLUSIONES

La automatización de cálculos es tarea fundamental del radiofísico hospitalario. En este trabajo, se presenta una herramienta gráfica, compacta y robusta que permite el cálculo y/o comprobación de blindajes de manera rápida y ordenada. Proporciona no sólo un ahorro de tiempo, sino que también una reducción de posibles errores en una labor tan poco rutinaria.

Además con el algoritmo de simulación para el cálculo de errores se verifica que la interfaz funciona adecuademante.

AGRADECIMIENTOS

La investigadora Mª José Ginzo agradece el apoyo de los proxectos 2013-PG064 de la Consellería de Educación e Ordenación Universitaria, al MTM2013-41383-P del Ministerio de Economía y Competitividad y a la Red IAP P7/06 de Belgian Science Policy.

Referencias

[1] J. J. Bevelacqua. Practical and Effective ALARA. Health Physics, 98, pp.39-47, 2010

[2] NCRP Report no. 51, Structural Shielding Design and Evaluation for Megavoltage X- and Gamma-Ray Radiotherapy Facilities, *Recommendations of the National Council on Radiation Protection and Measurements*, December 2005.

[3] M.C. Pujades, D. Granero, J. Vijande, F. Ballester, J. Perez-Calatayud, P. Papagiannis and F.A. Siebert. Air-kerma evaluation at the maze entrance of HDR brachytherapy facility. *J. Radiol. Proct.* 34, pp. 741-753, September 2014

[4] ICRP Publication 33, Protection against ionizing radiation from external sources used in medicine, *Pergamon Press*, vol.9, no.1, 1982

[5] M.C. Pujades-Claumarchirant, J. Perez- Calatayud and F. Ballester, Estudio de la carga de trabajo a aplicar en el cálculo de blindajes en instalaciones de braquiterapia HDR con Ir-192, *Radioprotección*, 19, no. 69, pp. 30-34, Octubre 2011.

[6] P. Papagiannis, D. Baltas, D. Granero, J. Perez-Calatayud, J. Gimeno, F. Ballester, J.L.M.Venselaar, Radiation transmission data for radionuclides and materials relevants to brachytherapy facility shielding, *Medical Physics*, 35, no. 11, November 2008

[7] Safety Reports Series no. 47 Radiation protection in the design of radiotherapy facilities, IAEA, Vienna, September 2006.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

MODELOS DE REGRESIÓN DISTRIBUCIONAL BIVARIANTE. APLICACIONES EN DIABETES

Jenifer Espasandín-Domínguez[1],Carmen Cadarso-Suárez[1], Thomas Kneib[2], Nadja Klein[2], Francisco Gude[3]

[1] Grupo de Bioestadística, Centro de Investigación en Medicina Molecular y Enfermedades Crónicas (CiMUS). Universidad de Santiago de Compostela.
[2] Chair of Statistics, Georg-August-Universität Göttingen, Germany.
[3] Unidad de Epidemiología Clínica. Hospital Clínico Universitario. Santiago de Compostela.

RESUMEN

Hasta la actualidad, en la mayoría de las revisiones bibliográficas, los diferentes marcadores existentes de concentración de azúcar en sangre empleados en la diagnosis y el control de diabetes, se estudiaron de forma independiente. Sin embargo, se conoce que existen correlaciones entre estos marcadores, la cual se espera que dependa de factores clínicos ó bioquímicos. Ignorar la existencia de este tipo de correlaciones, así como la heterocedasticidad de las distribuciones puede ocasionar errores en la clasificación de los pacientes. El objetivo de este trabajo es la identificación de de covariables importantes en la modelización de la media y la varianza de la hemoglobina glicosilada y de la fructosamina, así como en la correlación entre ambas proteinas. Para conseguir dicho objetivo, se utilizarán modelos de regresión distribucional aditiva estructurada de respuesta multivariante.

Palabras y frases clave: diabetes, hemoglobina glicosilada, regresión distribucional multivariante, frutosamina.

1. INTRODUCCIÓN

La *diabetes mellitus* es un trastorno metabólico que se manifiesta por unos niveles de glucosa en sangre (glucemia) por encima de los límites normales. Si no se trata adecuadamente, estos niveles alcanzan valores excesivamente altos, dando lugar a complicaciones como pueden ser retinopatía, nefropatía ó neuropatía. Según las previsiones de la Organización Mundial de la Salud, se espera que en 2030 afecte a 370 millones de personas en el mundo [1].

La hemoglobina glicosilada (*a1c*) se considera actualmente como un criterio para la diagnosis de diabetes. Sin embargo, la correlación entre los niveles de azúcar en sangre y *a1c* no es perfecta. Además de la hemoglobina, la frutosamina en suero (*fru*) también puede ser utilizada como marcador de los niveles de azúcar en sangre, pero a diferencia de la *a1c* que proporciona un índice de glucosa que hace referencia a los 3 últimos meses, la fructosamina proporciona información sobre los niveles de glucosa de las últimas 2-3 semanas. En consecuencia, la formulación de modelos independientes para *a1c* y *fru* puede ocasionar errores en la diagnosis de pacientes. Es más, puesto

que una de ellas representa el proceso de glicación en los glóbulos rojos (RBC) (intracelular) y la otra en plasma (extracelular), cualquier factor que modifique la supervivencia RBC, también modificará los resultados del test de a1c1 empleados para la determinación de los niveles de azúcar en sangre y por lo tanto su relación con el resto de proteínas glicosiladas.

El objetivo de este trabajo es la modelación conjunta de la hemoglobina glicosilada (a1c) y la fructosamina, en función de una serie de factores confusores como pueden ser la glucosa, la velocidad corpuscular media ó la creatinina entre otros.

2. REGRESIÓN DISTRIBUCIONAL MULTIVARIANTE

En la mayoría de los estudios estadísticos realizados hasta la fecha se centran en el estudio de predictores lineales desde el marco clásico de los modelos de regresión lineales generalizados [2]. En este tipo de modelos, se estudia únicamente la media de la variable respuesta respecto de una serie de covariables, ignorando la existencia de una posible dependencia o correlación entre diferentes variables respuesta como podrían ser *fru* y *a1c*. Supondremos que la distribución de *a1c* y *fru* es normal bivariante (entre otras distribuciones alternativas bivariantes entre las que seleccionaremos la más adecuada), emplearemos los modelos de regresión distribucional estruturados aditivos para estudiar esta posible dependencia [3].

Este tipo de modelos suponen una generalización de los modelos generalizados adictivos para la localización, escala y forma (GAMLSS) que permiten considerar distribuciones más complejas sobre la variable respuesta [4]. No obstante, los GAMLSS se restringen a respuestas univariantes mientras que con los modelos de regresión distribucional podremos considerar distribuciones multivariantes para la repuesta. En este caso particular, toman la forma:

$$\begin{cases} \eta^{\mu} = \beta_0^{\mu} + v'\beta^{\mu} + f_1^{\mu}(v_1) + \cdots + f_q^{\mu}(v_q) + f_{spat}^{\mu}(s) \\ \eta^{\sigma^2} = \beta_0^{\sigma^2} + v'\beta^{\sigma^2} + f_1^{\sigma^2}(v_1) + \cdots + f_q^{\sigma^2}(v_q) + f_{spat}^{\sigma^2}(s) \\ \eta^{\rho} = \beta_0^{\rho} + v'\beta^{\rho} + f_1^{\rho}(v_1) + \cdots + f_q^{\rho}(v_q) + f_{spat}^{\rho}(s), \end{cases} \tag{1}$$

donde η es una variable transformada del parámetro de la respuesta original $(fru, a1c)'$; $v'\beta$ denota los efectos paramétricos β de covariables, v; $f_i, i = 1, \ldots, q$ son funciones desconocidas suaves que nos permiten modelar efectos no lineales de las covariables continuas. $f_{spat}(s)$, representa los efectos espaciales. Los superíndices que aparecen en la fórmula se refieren a la ecuación de la media, la varianza y la correlación.

En este trabajo se estudiará tanto la media como la variabilidad de a1c (en %) y fru ($\mu mol/l$), para lo que se dispone de una muestra amplia de la población general adulta como resultado de un estudio previo realizado en A Estrada con pacientes seguidos hasta diciembre de 2013. Se estudiaron variables como la *edad* (en años) y el *sexo* de los pacientes, de factores bioquímicos como la glucosa (*glu*, en mg/dl), la albúmina en suero (*alb*, medida en g/l), el volumen corpuscular medio (vcm, en g/l en los glóbulos rojos) entre otros factores.

AGRADECIMIENTOS

Trabajo parcialmente subvencionado por el proyecto MTM2014-52975-C2-1-R, cofinanciado por el Ministerio de Economía y Competitividad (fondos FEDER incluidos) y una beca predoctoral concedida por la Xunta de Galicia.

Referencias

[1] World Health Organization. (2016). *Global report on diabetes.* Disponible en; $http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua = 1$.

[2] McCullagh P., Nelder J.A. (1989). *Generalized Linear Models.* Chapman-Hall. London.

[3] Klein N., Kneib T., Klasen S. and Lang S. (2014). Bayesian Structured Additive Distributional Regression for Multivariate Responses. *Journal of the Royal Statistical Society: Series C.* 64, 569–591.

[4] R. A. Rigby and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape.*Applied Statistics.* 54, Part 3, 507–554.

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

EFECTO DE LOS IRRIGANTES DENTINARIOS EN LA CEMENTACIÓN ADHESIVA DE POSTES INTRARRADICULARES: ESTUDIO IN VITRO

Sandra García Varela[1], María José Ginzo Villamayor [2], Benjamín Martín Biedma [1], João Carlos Tomás Ramos[3], Pedro Manuel Rivas Lombardero[1]

[1] Universidad de Santiago de Compostela. Departamento de Estomatología
[2] Universidad de Santiago de Compostela. Departamento de Estadística e Investigación Operativa
[3] Universidad de Coimbra. Área de Medicina Dentária

RESUMEN

Los dientes endodonciados, generalmente, han llegado a este status después de haber perdido gran parte de su corona natural. El objetivo principal de la restauración con postes es la sustitución artificial de la corona y la restauración del diente de modo que la reconstrucción coronaria y la raíz formen una unidad homogénea, funcional y estética. Esta unidad debe asegurar una transmisión/absorción correcta de las fuerzas oclusales y una adecuada retención de la restauración. Los postes intrarradiculares deben ser cementados mediante un sistema de adhesivo/cemento para no sobrecargar la dentina radicular. La aplicación de estos materiales requiere el tratamiento de la superficie dentinaria radicular que es en donde se adhieren los adhesivos y el cemento para unir el poste intrarradicular a la raíz dental.

El principal objetivo de este estudio fue la evaluación del efecto del tratamiento de la superficie dentinaria del conducto radicular mediante diferentes irrigantes en la resistencia de la interfase cemento-adhesivo-dentina de los sistemas adhesivos/cemento para la cementación de postes intrarradiculares de fibra de vidrio. Se plantearon 3 hipótesis nulas que consisten en que el tipo de irrigante no influye en la resistencia de la adhesión a la dentina radicular. No existen diferencias en la adhesión a la dentina radicular proporcionada por los diferentes sistemas adhesivos. No existen diferencias significativas en la adhesión obtenida en diferentes zonas del conducto radicular.

Se realizó un estudio in vitro analítico transversal experimental con caso control de asignación aleatoria. Se utilizaron 120 dientes unirradiculares humanos extraídos previo consentimiento del paciente y siguiendo el protocolo revisado y aprobado por el Comité Autonómico de Ética de Investigación de Galicia (CAEIG). Los dientes fueron mantenidos en condiciones hídricas y asépticas hasta el momento del estudio. Se realizaron 12 grupos (n=10) en función de 4 irrigantes de endodoncia (control con agua destilada; clorhexidina 2%; NaOCl 5%; EDTA 17%) y 3 sistemas adhesivos diferentes (ácido ortofosfórico, ParaBond® A + B / ParaCore®; ParaBond ® Non-Rinse Conditioner, ParaBond® A+B / ParaCore®; Clearfil ™ Universal Bond / Clearfil ™ DC Core Plus. Se utilizó el poste Tenax® Fiber White cementado por su porción cilíndrica. Cada diente se seccionó perpendicularmente a su eje longitudinal desde la unión amelocementaria hasta la zona apical del poste. Se obtuvieron 3 secciones o discos de cada diente de 1 mm de espesor. La Figura 1 representa los cortes seriados de cada diente tratado, desde coronal hasta apical, de 1 mm de espesor. La Figura 2 representa los discos obtenidos después del corte seriado.

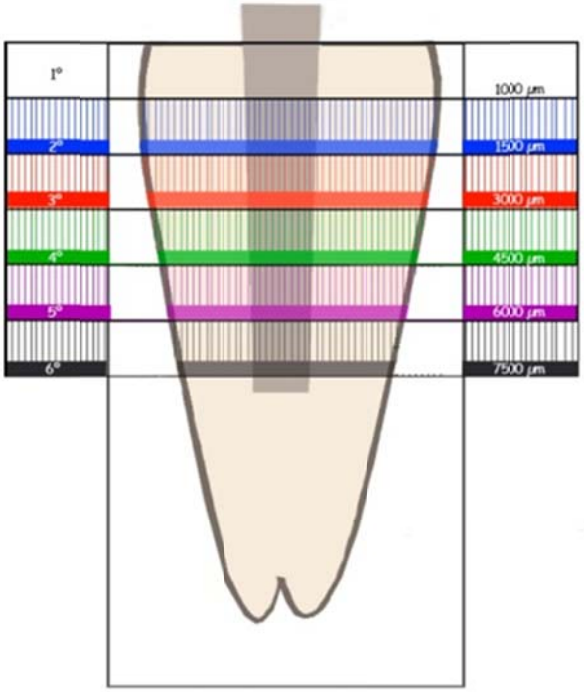


Figura 1: Esquema simplificado de los cortes dentales.

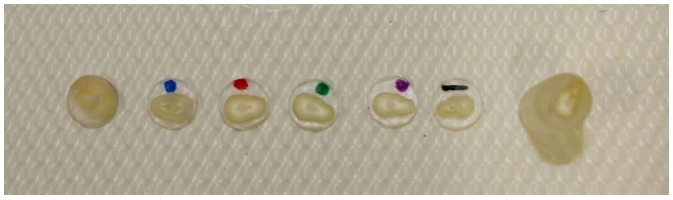


Figura 2: Fragmentos y discos obtenidos del corte seriado.

Cada grupo contiene 30 discos que serán el objeto de este estudio (n=30). Cada disco fue sometido a una fuerza de empuje mediante el test de Push-Out con una velocidad de travesaño de 1 mm/min (veáse la Figura 3, el color naranja corresponde a la dentina; el amarillo al adhesivo/cemento; y el azul al poste.). La resistencia adhesiva se calculó en MPa.

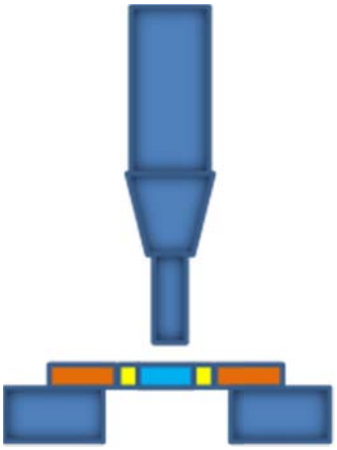


Figura 3: Esquema del posicionado de los discos para el test push-out.

Se realizó un análisis descriptivo de todas las variables incluidas en el estudio, expresando las variables cuantitativas como media±desviación típica.

La comparación de medias para más de 2 grupos se realizó con el test ANOVA. Para aplicar el ANOVA se han de verificar los siguientes supuestos: Normalidad en cada nivel de factor (para ello se ha aplicado el test de Shapiro-Wilks); Homocedasticidad (igual varianza en cada nivel de factor) (en este caso se ha aplicado el Test de Levene) e independencia de las observaciones, que lo son por la naturaleza del experimento. En aquellos casos que no se ha verificado alguno de los tres supuestos del ANOVA se ha usado la prueba no paramétrica de Kruskal-Wallis, que es la alternativa no paramétrica al modelo ANOVA de una vía para comparar más de dos grupos independientes. Cunando rechazamos la hipótesis nula de igual de medias mediante cualquiera de las dos pruebas anteriores, necesitamos aplicar un test de comparaciones múltiples como el de Tukey. El análisis estadístico se ha realizado con R, http://www.r-project.org/. La versión utilizada fue la 3.1.3.

El NaOCl 5%, en relación con los otros irrigantes, disminuyó significativamente la resistencia de la interfaz adhesiva. El uso de NaOCl 5% Clearfil ™ Universal Bond provocó una reducción significativa de la resistencia de la interfaz adhesiva (véase la Tabla 1).

	Ortofosfórico (grabado ácido)	**Parabond (autograbante 2 pasos)**	**Clearfil (autograbante 1 paso)**	**Total**
AGUA DESTILADA	16,76±4,43	16,26±5,88	17,06±4,33	16,69
EDTA 17%	17,07±5,03	14,95±5,14	16,66±4,84	16,23
NaOCl5%	14,61±3,81	14,09±4,64	12,92±3,79	13,87
CHX 2%	17,30±5,95	16,48±3,04	16,60±4,58	16,79
Total	16,43	15,44	15,81	15,89

Tabla1: Media y desviación típica de los valores del test Push-Out en MPa de cada grupo experimental (n=30).

En la Tabla 1, en la última fila tiene la media obtenida para cada uno de los tipos de adhesivo. El análisis en función de los adhesivos revela que al aplicar la prueba Anova la media de los sistemas adhesivos es la misma. Se deduce que no existen diferencias significativas de resistencia en la adhesión en función del tipo de adhesivo seleccionado.

El análisis en función del tipo de irrigante revela el *p*-valor obtenido al aplicar la prueba Anova es menor que 0,01, por lo que podemos decir que las medias de los tipos seleccionados no es la misma. A modo descriptivo, en la Tabla 1, en la última columna (Total) se tiene la media para cada uno de los tipos de irrigante. La media no es igual para todos, por lo tanto, tenemos que hacer comparaciones dos a dos para ver si algunos de los grupos son iguales. Para ello, empleamos el test de Tuckey. Al hacer comparaciones dos a dos resulta que las medias son iguales (p valor > 0,01) al comparar Agua Destilada con EDTA 17%; CHX 2% con EDTA 17% y CHX 2% con Agua Destilada. Cualquiera de los irrigantes con el NaOCl5% tiene diferente media, de hecho la media del NaOCl5% es bastante inferior a las otras tres.

El análisis en función de los tipos de irrigantes y adhesivos se realiza mediante un Anova de dos vías para ver la interacción tipo:irrigante y se obtiene un *p*-valor de 0,71, es decir, el término de interacción no es significativo: no hay interacción entre las dos variables. En este caso lo que está indicado es realizar un análisis de una vía entre las combinaciones de tratamientos para encontrar la mejor combinación de los mismos.

Palabras y frases clave: Irrigantes, poste de fibra, adhesión, dentina, push-out.

AGRADECIMIENTOS

La investigadora Mª José Ginzo agradece el apoyo de los proxectos 2013-PG064 de la Consellería de Educación e Ordenación Universitaria, al MTM2013-41383-P del Ministerio de Economía y Competitividad y a la Red IAP P7/06 de Belgian Science Policy.

Referencias

[1] Kadam A., Pujar M., Patil C. (2013). Evaluation of push-out bond strength of two fiber-reinforced composite posts systems using two luting cements in vitro. *J Conserv Dent* 16(5), 444-8.

[2] Varela S.G., Rábade L.B., Lombardero P.R., Sixto J.M., Bahillo J.D., Park S.A. (2003). In vitro study of endodontic post cementation protocols that use resin cements. *J Prosthet Dent* 89(2),146-53.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

ESTIMAÇÃO DA DEGRADAÇÃO DE BIOPOLÍMEROS PARA USO ODONTOLÓGICO ATRAVÉS DE MODELOS DE CRESCIMENTO

Yaroslava Robles-Bykbaev[1], Salvador Naya[2], Javier Tarrío-Saavedra[2]

[1] Facultade de Informática. Departamento de Matemáticas. Universidade da Coruña.

[2] Escola Politécnica Superior. Departamento de Matemáticas. Universidade da Coruña

RESUMO

O presente trabalho visa retalatar a modelagem estatística da degradação de diversas formulações de biopolímeros basados nos ácidos poliláctico e glicólico, usados como andaimes em odontologia. Os andaimes dentais são dispositivos cuja funcão é ajudar o tecido ósseo a regenerarse após da extracção do dente, antes de colocar um implante. Os polímeros e copolímeros baseados no ácido poliláctico e no glicólico são conhecidos e utilizados neste tipo de aplicações dentais principalmente porque são biodegradáveis e biocompatíveis. É importante salientar que antes da aplicação destes materiais, é preciso estimar seu nível de degradação en função do tempo para alcançar um equilíbrio entre o crescimiento do tecido ósseo e perda da massa dos andaimes. Por tanto, a perda de massa tem sido definida como a característica da degradação dos biopolímeros, aplicando-se diversos modelos de regressão não lineares chamados modelos de crescimiento para definir a sua dependencia com respeito do tempo; estimar o seu nível de degradação a cada instante e obter os parâmetros que permitan caracterizar e comparar os biomateriais. Em particular têm sido aplicados os modelos de regressão baseados na função de Gompertz, na função logística e na função de Richards, e tem sido escolhido o modelo mais adecuado según os críterios da bondade de ajuste e análise da significação dos parâmetros. Além disso, o ajuste dos modelos semi-paramétricos smoothing spline cúbicos e lowess permitem avaliar o desenvolvimento dos modelos paramétricos além de fornecer uma solução inicial para obter estimativas mais confiáveis dos seus parâmetros.

Palavras-chave: Biopolímeros, regressão não linear, modelos de crescimiento, Gomperz, função logística, smoothing splines.

1. MATERIAIS E OBTENÇÃO DOS DADOS

As mostras dos 7 biopolímeros baseados no ácido poliláctico (PLA) e glicólico têm sido analissados, com uma massa inicial de 0.3g cada uma. Elas foram diluídas em solução salina PBS (1 ml), foram mantidas a 37ºC para reproducir as condições do corpo humano. A perda de massa en função do tempo tem sido tomada como a resposta, e quantificada em intervalos de 3 dias. Na Tabela 1 apresenta-se uma definição mais completa dos biopolímeros estudados.

Tabela 1: Várias formulações de biopolímeros à base de PLA e ácido glicólico.

Etiqueta	Polímero	Mw* (kg/mol)
PDLG 5002	50/50 DL-lactide/glycolide copolymer	17
PDLG 5004	50/50 DL-lactide/glycolide copolymer	44
PDLG 5010	50/50 DL-lactide/glycolide copolymer	153
PDLG 7502	75/25 DL-lactide/glycolide copolymer	17
PDLG 7507	75/25 DL-lactide/glycolide copolymer	95
PDL 02A	Poly(DL-lactide)	17
PDL 02	Poly(DL-lactide)	17
PDL 02	70/30 L-lactide/DL-lactide copolymer	17

* Relativo a padrões de poliestireno (PS). Valores indicativos baseados na correlação de Mark Houwink.

2. MODELOS DE CRESCIMENTO

A necessidade de modelar o crescimento biológico popularizou o uso de funções monótonas crescentes não lineares, caracterizadas por um crescimento exponencial da variábel medida e um ponto de inflexão a partir do qual a velocidade do seu crescimento diminui gradualmente atè que esta variábel atinja uma zona de saturação [1]. A seguir, apresentam-se tres das funções mais conhecidas:

Função de Gompertz integrada:

$$y(t) = A\exp\left[-\exp\left(\frac{\mu\ \exp(1)}{A}\ (\lambda\ -\ t) + 1\right)\right]$$

Função de Richards:

$$y(t) = A\left[1 + \nu\exp\left(1 + \nu + \frac{\mu}{A}(1+\nu)^{1+1/\nu}\ (\lambda\ -\ t)\right)\right]^{-1/\nu}$$

Função logística:

$$y(t) = \frac{A}{1 + \exp\left(\frac{4\mu\ (\lambda\ -\ t)}{A} + 2\right)}$$

Onde A é a assimptota superior da função (neste caso particular, a quantidade de perda de massa do polímero), μ é a taxa máxima da mudança da curva (neste caso a máxima velocidade de perda de massa), λ é o tempo de atraso no início do crescimento exponencial, ν è o parâmetro de forma e t é o tempo de degradação para um dos 7 biopolímeros.

Além disso, foram aplicados modelos de regressão semi-paramétricos (model-free) [2] baseados na aplicação de smoothing splines cúbicos. A aplicação deste tipo de modelo permite obter estimativas fiáveis dos parâmetros acima descritos, a partir das curvas ajustadas.

3. RESULTADOS

Nesta seção apresenta-se o ajustamento do modelo de regressão não linear com a função de Gompertz para sete tipos de biopolímeros definidos pela percentagem de ácido láctico (PLA), glicólico (PEG) e caprolactona (PLPLG), para determinar a perda de massa de cada um deles en função do tempo. O pacote `grofit` [1] do software estatístico R permite obter o ajuste do modelo paramétrico partindo de uma serie de valores inciais (solução inicial) dos parâmetros que têm sido estimados pelo ajuste de um modelo lowess [1]. Este procedimiento permite resolver os problemas do proceso de otimização e obter estimativas confiáveis dos parâmetros. Os modelos de regressão obtidos, foram avaliados por meio de diversas técnicas de bondade de ajuste, por exemplo o coeficiente de de determinação (R^2) definido como $R^2 = \frac{\left[\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{y})\right]^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$ [3], o Criterio de Informação de Akaike (AIC), a soma dos quadrados dos

resíduos, a soma quadrática residual de previção (PRSS), e a validação cruzada generalizada (GCV). O ajuste paramétrico não linear amostra-se na Figura 1 e caracteriza-se por um R^2 = 0.94, RSS = 0.0014, AIC = 185.78, enquanto que o ajuste com smoothing splines cúbicos é definido por um PRESS = 0.0006, GCV = 0.0002 e AIC = -119.6. Na Figura 1 é apresentado que o ajuste não paramétrico tende a sobreajustar os dados reais, de fato, también ajusta o error experimental correspondiente a perda da massa negativa no inicio do experimento. No en tanto, através do ajuste da função de Gompertz explica-se a maior parte das mudanças na perda da massa relativa ao tempo, evitando a modelagem de uma parte do error experimental próprio deste tipo de test de laboratorio. Os parámetros do modelo permitem definir e caracterizar a degradação do biopolímero quantitativamente, e permite a sua comparação com outros materiais alternativos.

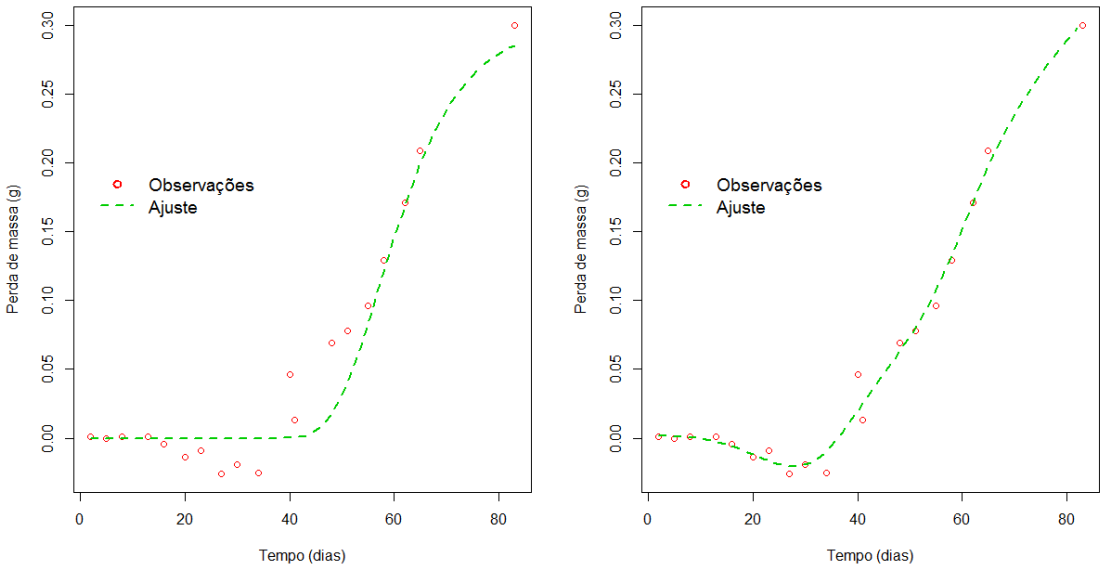


Figura 1: Á esquerda amostra-se o ajuste não linear da função Gompertz para modelar a perda da massa do polímero PDL02 en função do tempo, enquanto que no gráfico da direita é apresentado um ajuste smoothing spline cúbico, ambos aplicados aos dados de perda de massa do polílmero PDL02 em função do tempo.

AGRADECIMENTOS

Este trabalho foi financiado pela Secretaría Nacional de Educación Superior, Ciencia, Tecnología, e Innovacióno (SENESCYT) do Governo do Ecuador e do projeto MTM2014-52876-R (fundos FEDER incluídos).

REFERÊNCIAS

[1] Kahm, M., Hasenbrink, G., Lichtenberg-Fraté, H., Ludwig, J., & Kschischo, M. (2010). grofit: fitting biological growth curves with R. *Journal of Statistical Software*, 33(7), 1-21.

[2] De Boor, C., De Boor, C., De Boor, C., & De Boor, C. (1978). A practical guide to splines (Vol. 27, p. 325). New York: Springer-Verlag.

[3] Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of statistical software*, *27*(5), 1-32.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

INFORME SOBRE A XESTIÓN E O ESTADO ECONÓMICO DO COOPERATIVISMO AGRARIO EN GALICIA

Higinio Mougán-Bouzón[1], Mª José Cabaleiro Casal[2], Mª José Ginzo Villamayor[3], Ruth Rodríguez Ferreiros[1], David Vila López[1], Ricardo Francos Gendra[1], Sandra Rey Louzán[1]

[1] Asociación Galega de Cooperativas Agroalimentarias (AGACA)
[2] Universidade de Vigo (UVigo)
[3] Universidade de Santiago de Compostela (USC)

RESUMO

O cooperativismo agroalimentario en Galicia é un factor fundamental na vertebración do desenvolvemento rural, tanto no ámbito territorial como no social e económico. A súa fortaleza reside na capacidade que posúe para converterse en vector de cohesión social e económica, xerando emprego de calidade directo e indirecto, e en resorte de defensa e adaptación aos cambios no sector agroalimentario e no contorno xeral. O movemento cooperativo en Galicia leva moitos anos xerando riqueza na economía e na sociedade galegas. As cooperativas son empresas de persoas que melloran a viabilidade das explotacións agrarias e doutros negocios, son motores de desenvolvemento económico, crean e distribúen equitativamente riqueza, contribúen á igualdade, teñen efectos beneficiosos sobre o contorno social e evitan a deslocalización. Ademais, o modelo cooperativo é un xeito comercialmente eficaz e eficiente de realizar actividades comerciais. As cooperativas teñen a vantaxe de ser capaces de traballar con eficacia tanto a escala nunha moi pequena como nun ámbito moi grande. As sociedades cooperativas ofrecen produtos e servizos cun valor engadido que os diferenza no mercado, e que os posiciona como a verdadeira alternativa para aquelas persoas que buscan adquirir bens realmente sostibles, solidarios e de calidade. As cooperativas agroalimentarias enfróntanse hoxe a grandes retos e desafíos, para o cal é preciso traballar a favor da profesionalización do sector, da cualificación técnica e da modernización da súa xestión.

Ter unha visión en profundidade da realidade, a través de estudos coma o INFORME SOBRE A XESTIÓN E O ESTADO ECONÓMICO DO COOPERATIVISMO AGRARIO EN GALICIA contribúe a sentar as estratexias para a súa adaptación aos cambios tanto sociais como económicos que se están a producir. As cifras do cooperativismo agroalimentario en Galicia:

200 cooperativas agroalimentarias con actividade económica destacada

7.000 postos de traballos directos, o que representa case o 13% das persoas ocupadas en Galicia no sector primario[1].

Preto de 1.800 millóns de euros facturados, polas cooperativas agroalimentarias e polas sociedades participadas, o que supón o 62% do valor da Produción Final Agraria Galega[2]

Aproximadamente 40.000 persoas socias de cooperativas agroalimentarias (se ben, unha parte ten pouca actividade).

[1] O sector primario engloba agricultura, gandaría, caza, silvicultura, pesca e acuicultura. Enquisa poboación activa. INE. 2013.
[2] Anuario Estatística 2006-2012. Consellería do Medio Rural.

No contorno de 7.000 postos de traballo directos, nas cooperativas agroalimentarias e nas sociedades participadas.

Neste traballo preséntase a análise xeral do cooperativismo agrario en Galicia dende 1995 ata 2013.

Palabras e frases chave: cooperativismo agroalimentario, estatística de cooperativas, mostraxe estratificada.

1. INTRODUCIÓN

O obxectivo xeral do Informe é esclarecer a verdadeira contribución do cooperativismo agroalimentario ao desenvolvemento económico e social en Galicia, incorporando unha análise socioeconómica que permite ter unha visión global da situación actual dun sector estratéxico para a economía galega. Así mesmo, entre os obxectivos específicos máis importantes destacan:

- Que os axentes cooperativos teñan un coñecemento acorde coa realidade cooperativa.
- Que os propios cooperativistas do agro saiban e coñezan en qué medida contribúen cos seus esforzos a conformar un espazo socioeconómico no que o seu papel é de gran importancia.
- A reflexión e sensibilización por parte da sociedade en xeral, sobre unha fórmula empresarial da que aínda existen moitos prexuízos.

O estudo trata de ilustrar o que significan as empresas cooperativas agroalimentarias galegas e, en especial, o importante labor que xogan dentro do medio rural, así como as súas preocupacións, os seus retos e desafíos máis urxentes e as estratexias necesarias para acadar con éxito os obxectivos marcados.

Para indagar nas características socioeconómicas das cooperativas agroalimentarias galegas elabórase unha enquisa, que se vén facendo dende o ano 1995 ata actualidade. Faise unha mostraxe estratificada, definíndose os estratos en función do sector ao que pertencen as cooperativas: Estrato 1: cooperativas do sector da Gandería; Estrato 2: cooperativas do sector da Horta-Flor e Estrato 3: cooperativas do sector do Viño.
O Informe estrutúrase nas seguintes partes:

- Introdución sobre a relevancia do cooperativismo agroalimentario no contorno internacional, nacional e autonómico
- Aspectos socio-laborais nas cooperativas agroalimentarias
- Análise económica-financeira e comercial
- Resumo-conclusións

Centrándonos na vertente estatística, imos achegar datos sobre os apartados 2 e 3.

2. ASPECTOS SOCIO-LABORAIS NAS COOPERATIVAS AGROALIMENTARIAS

As cooperativas agroalimentarias teñen demostrada a súa capacidade para promover o desenvolvemento das áreas nas que se asentan, de maneira sostible e con visión de futuro, favorecendo a redistribución dos recursos e o equilibrio territorial e medioambiental. Para isto, precisan dunha dimensión óptima, que se pode conseguir a través dos procesos de concentración empresarial que se veñen levando nestes últimos anos. A redución de custes e as economías de escala permiten dar resposta a un mercado que esixe calidade, trazabilidade, seguridade alimentaria e diferenciación dos produtos.
As cooperativas estudadas constitúen unha evidencia fiel da traxectoria xeral do cooperativismo agroalimentario en Galicia, habida conta dos criterios de selección das súas variables.
A Táboa 1 contén unha primeira aproximación dos datos máis significativos da mostra.

	1995	2000	2005	2010	2013
Nº de cooperativas	44	59	54	52	48
Promedio persoas socias	192	226	265	253	266
Promedio cadro persoal	6,6	8,8	13	12,77	13,15
Promedio facturación (millóns €)	2,7	3,99	5,29	5,08	5,8

Táboa 1: Dimensión mostra 1995-2013

O número medio de persoas contratadas case se duplica desde 1995 ata 2013. A media por cooperativa pasa de 6,6 a algo máis de 13 persoas (véxase a Figura 1). Pero, tan importante coma este dato é salientar que a porcentaxe de contratos fixos en 2013 é do 88%, sete veces a de contratos temporais. En 1995, a porcentaxe de contratos indefinidos era do 55%.

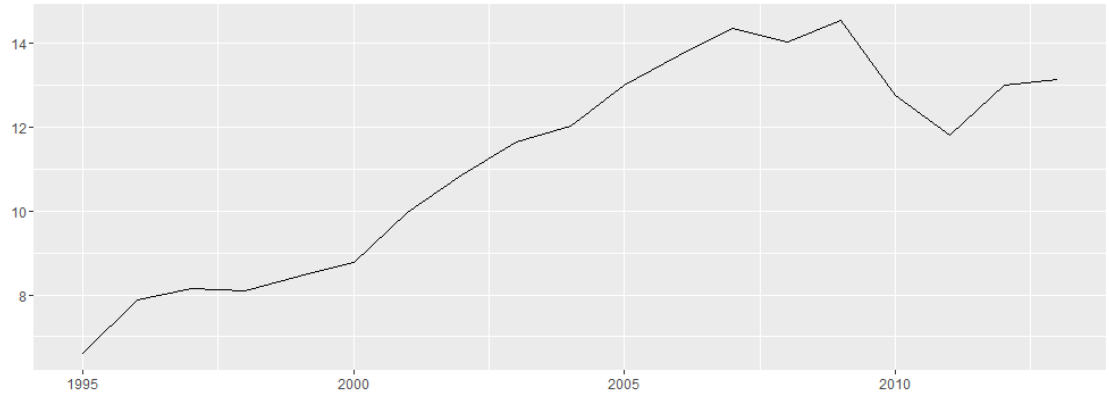


Figura 1: Evolución da cifra media de contratación (cadro de persoal). 1995-2013

Polo que respecta ao número medio de persoas das cooperativas da mostra, desde 1995 ata 2013 dáse un aumento do 39%, acadando os 266 cooperativistas (véxase a Figura 2).

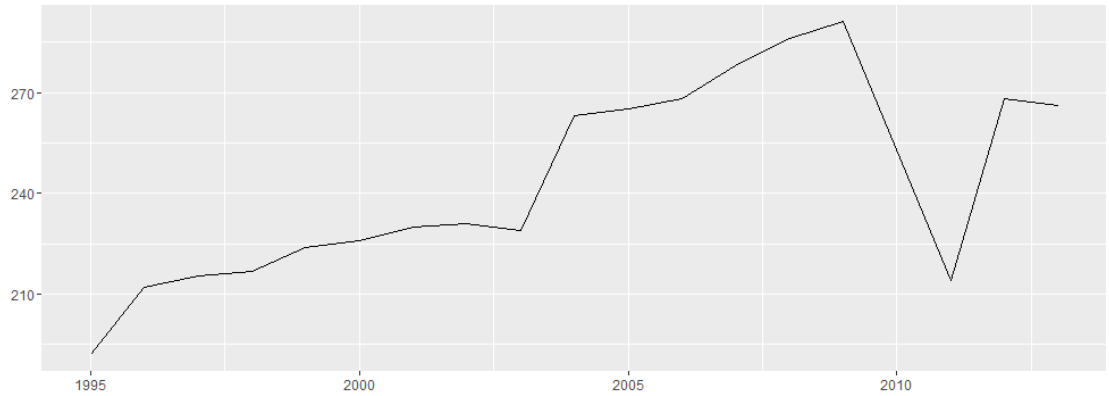


Figura 2: Evolución da cifra media de persoas socias. 1995-2013

Polo que respecta á cifra media de facturación das empresas da mostra, en 2013 acádanse os 5,8 millóns de euros, o que indica un incremento do 115% con respecto a 1995. (Véxase a Figura 3).

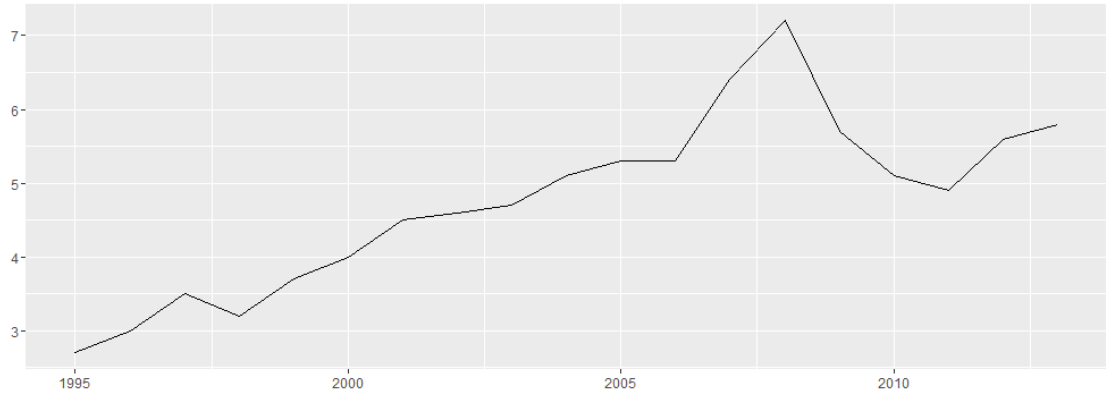


Figura 3: Evolución da cifra media de facturación. 1995-2013

3. ANÁLISE ECONÓMICA XERAL DO COOPERATIVISMO AGROALIMENTARIO EN GALICIA

Para a análise económica selecciónanse os ratios da Táboa 2. En canto á evolución da liquidez xeral obsérvase que as cooperativas foron tendo maior capacidade para facerlle fronte ás súas débedas a curto prazo co paso dos anos. Porén, nos últimos anos considérase demasiado elevado porque pode incidir negativamente na rendibilidade.

O fondo de manobra indica que as cooperativas teñen un correcto equilibrio financeiro. Cando é alto, indica que hai un risco menor de falta de liquidez (como se pode observar, sobre todo, en 2010).

O coeficiente de garantía mide a garantía que a cooperativa ofrece aos seus acredores. Obsérvase que foi mellorando na evolución temporal.

A evolución do índice de seguridade de custes fixos indica que as cooperativas poden incrementar moi pouco os custes fixos sen entrar en perdas, sobre todo, nos últimos anos, debido a que a crise socioeconómica tamén incide nestas sociedades.

	1995	**2000**	**2005**	**2010**	**2013**
Liquidez xeral	1,37	1,94	1,80	3,66	3,21
Fondo de manobra	0,38	0,32	0,28	1,77	0,27
Coeficiente de garantía	1,97	2,31	2,29	4,03	3,20
Índice de seguridade de custes fixos	0,43[3]	0,26	0,17	0,11	0,10

Táboa 2: Ratios para análise económica

AGRADECEMENTOS

Os autores deste traballo agradecen a participación das cooperativas agroalimentarias que achegan os seus datos ao estudo, xa que sen elas non sería posible a súa elaboración.
Vaia o noso agradecemento para a Consellería de Economía, Emprego e Industria, Xunta de Galicia, polo interese e apoio mostrados ano tras ano, sempre comprometida co cooperativismo agroalimentario de Galicia.
María José Ginzo agradece o apoio dos proxectos 2013-PG064 da Consellería de Educación e Ordenación Universitaria, MTM2013-41383-P do Ministerio de Economía e Competitividad, e a Rede IAP P7/06 de Belgian Science Policy
.

REFERENCIAS

[1] ARDÁN: Galicia 2014. Directorio 13.500 empresas, Vigo: IGAPE, Consorcio Zona Franca de Vigo, 2014.

[2] INSTITUTO GALEGO DE ESTATÍSTICA, www.ige.eu

[3] INSTITUTO NACIONAL DE ESTADÍSTICA, www.ine.es

[4] Jordán Rodríguez, M., (2010). Informe de síntese. 2010, Santiago de Compostela: Consellería de Traballo. Dirección Xeral de Relacións Laborais.

[3] Dato de 1995 non dispoñible

II Encuentro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de junio, 1 y 2 de julio de 2016

PREBED: Software de evaluación odontológica

Lozano, D.[1], Martínez, I.[1], Valera, I. [2]; Limeres, J.[2] , Roca, J.[1], Diz P.[2] y Cadarso C.[3]

[1] Departamento de Estadística e I.O. – Universidad de Vigo
[2] Departamento de Estomatología - Universidad de Santiago de Compostela
[3] Unidad de Bioestadística - Dpto. de Estadística e I.O. – Universidad de Santiago de Compostela

RESUMEN

En la atención odontológica de Pacientes Especiales es muy importante la determinación del tipo de comportamiento ante una exploración bucal antes de llevarla a cabo. En la actualidad, en la Unidad de Pacientes Especiales se realiza un cuestionario al acompañante del paciente previo a la exploración. En este trabajo presentamos PREBED, una aplicación web basada en códigos R que predice el tipo de comportamiento del paciente ante una exploración odontológica a través del conocimiento del tipo de patología que presenta, su edad y diversas variables que recogen sus rutinas diarias utilizando modelos de regresión logística. El programa también predice el tipo de anestesia a utilizar, en caso de ser necesaria, para las distintas patologías.

Palabras y frases clave: predicción, modelo logístico, clasificación, GLM, GAM.

1. INTRODUCCIÓN

En la atención odontológica a Pacientes Especiales (PE) es crucial poder conocer de antemano el comportamiento ante una exploración bucal puesto que, dependiendo de la patología, pueden llegar a producirse sucesos desagradables e incluso violentos en la consulta. De modo habitual estos pacientes son derivados a Unidades de Pacientes Especiales (UPE) donde se les practica la exploración con las medidas restrictivas necesarias después de recoger información acerca de su patología y de sus hábitos diarios con un cuestionario establecido. Con el fin de facilitar este proceso de clasificación se ha trabajado en diversos modelos de regresión según la patología del paciente. Sin embargo, puesto que estos modelos son poco prácticos si no están soportado por software adecuado hemos implementado PREBED, un software basado en código R para la clasificación de los PE que es accesible vía web y es compatible con Windows, Mac OS y Unix.

2. DESCRIPCIÓN DEL SOFTWARE

Los PE se clasifican en base a si necesitan o no necesitan restricción física (RF/NoRF) para poder realizar la exploración bucal. Se han utilizado tres modelos de regresión diferentes según la patología del paciente (Síndrome de Down (SD), Parálisis Cerebral (PC) y Desórdenes de Espectro Autista (DEA)). El software procesa las respuestas a las preguntas acerca de sus rutinas diarias (véase Tabla 1) y da una clasificación RF/NoRF.

Preguntas	Respuestas		
Recibe medicación psicotrópica	Sí	No	
Tiene hábitos de mordisqueo	Sí	No	
Tolera los ruídos	Sí	No	
Está escolarizado	No	Sí, en un centro especializado	Sí, en un centro normal
Colabora cuando le cortan las uñas	Sí	Sí, pero…	No

Colabora cuando le asean	Sí	Sí, pero…	No
Colabora cuando le cepillan los dientes	Sí	Sí, pero…	No

Tabla 1: Preguntas que se utilizan en los diferentes modelos.

PREBED no es ejecutable como tal, si no que es más bien un código que es ejecutable dentro de otra aplicación. Está diseñado para funcionar en un entorno R (www.r-project.org) dentro de un entorno web.

2.1. Base de datos

La aplicación requiere de una base de datos derivada de la atención en la UPE de 180 pacientes. Esta contiene la información recogida en las preguntas de la Tabla 1 además de la patología del paciente (SD, PC, DEA), la edad y su comportamiento ante la exploración bucal (RF(/NoRF).

Esta base de datos se utilizó para desarrollar y probar los distintos modelos lineales generalizados (GLM) y aditivos generalizados (GAM) para clasificar la comportamiento en la exploración.

El diagrama de flujo se puede observar en la Figura 1.

2.2. Programa

El programa permite predecir la clasificación tanto del comportamiento ante la exploración como del tipo de anestesia a utilizar con el paciente. Ambas clasificaciones utilizan modelos diferentes dependiendo de la patología.

Para iniciar el programa simplemente es necesario tener conexión a internet y acceder a la web de PREBED. Al inicio del programa aparecerá una ventana de entrada de datos y tras introducirlos se mostrará la clasificación correspondiente como puede verse en la Figura 2.

3. VALIDACIÓN

Los 180 datos de la base de datos inicial se han particionado aleatoriamente en 100 datos para la obtención del modelo y 80 datos para su validación. Se han obtenido valores de sensibilidad por encima del 80% en la clasificación de los pacientes dentro de cada una de las tres patologías.

Referencias

[1] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] Monica Lopez-Raton, Maria Xose Rodriguez-Alvarez, Carmen Cadarso Suarez, Francisco Gude Sampedro (2014). OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. *Journal of Statistical Software*, 61(8), 1-36. URL http://www.jstatsoft.org/v61/i08/..

[3] Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* (B) 73(1):3-36

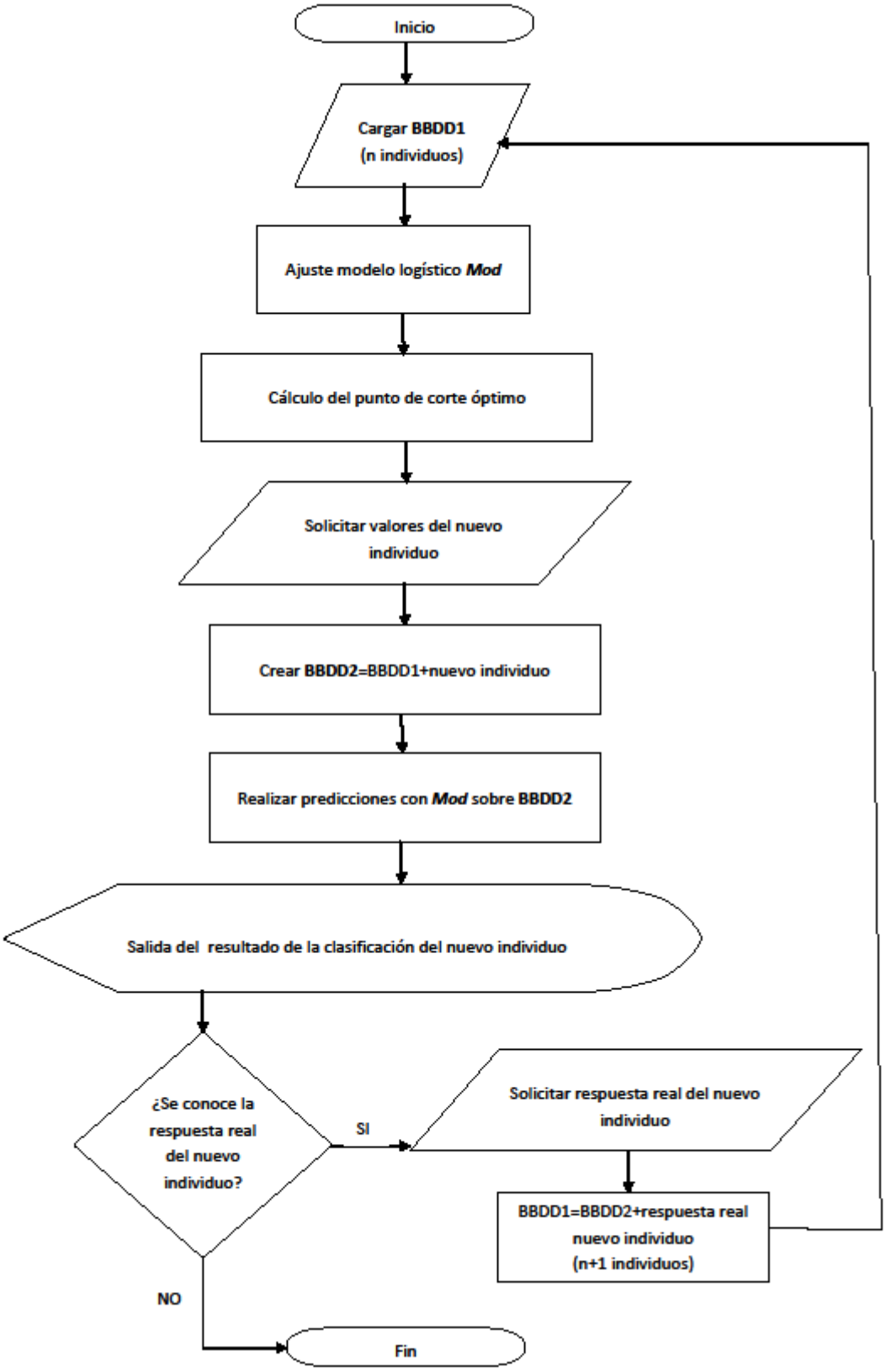


Figura 1: Diagrama de flujo de PREBED.

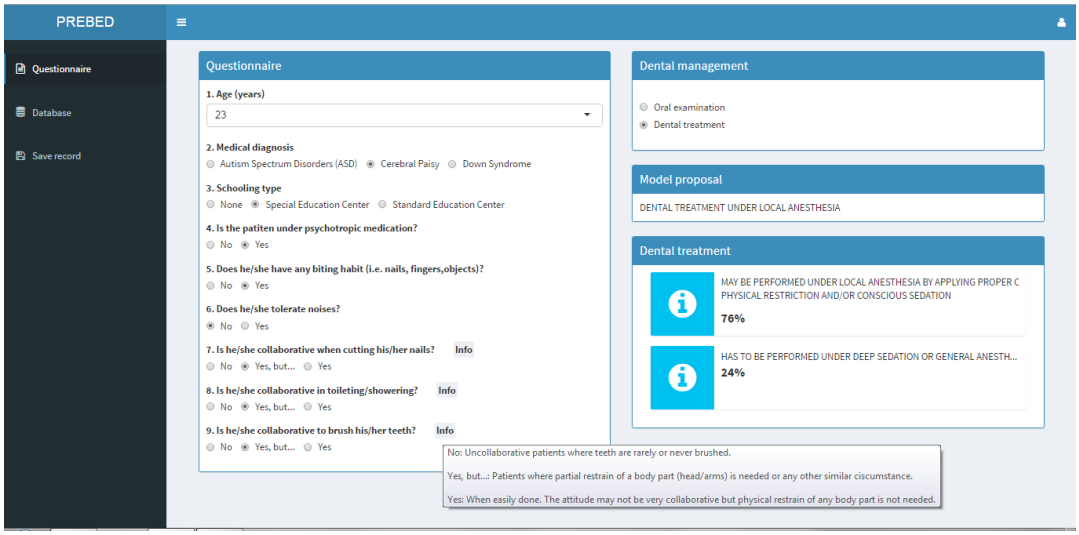


Figura 2: Pantalla de inicio de PREBED.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

A general class of exponentiated sinh Cauchy regression models

Ramires, T.G 1[1] and Ortega, E.M.M. 2[2]

[1] LCE - University of São Paulo/ESALQ. e-mail: *thiagogentil@gmail.com*
[2] LCE - University of São Paulo/ESALQ. e-mail: *edwin@usp.br*

ABSTRACT

A new general class of exponentiated sinh Cauchy regression models for location, scale and shape parameters is introduced. It can be applied to censored data and be used more effectively in survival analysis. the usefulness of the model is demonstrated through the an analyze of real data set. Finally, the proposed algorithm the estimation has been coded and implemented in the GAMLLS script and is was made available.

Palabras e frases chave: Exponentiated sinh Cauchy regression model; GAMLSS; Bimodality.

1. The exponentiated sinh Cauchy regression model

In many practical applications, the lifetimes are affected by explanatory variables. Parametric models, considering the class of location models, for estimating univariate survival functions and for the censored data regression problems are widely used. The disadvantage to consider the class of locations models is that the variance, skewness, bimodality, kurtosis and other parameters are not modelled explicitly in terms of the explanatory variables. As an alternative, the generalized additive models for location, scale and shape (GAMLSS) [2] wherein the systematic part of the model is expanded to allow not only the location but all the parameters of the conditional distribution of Y to be modelled as parametric functions of explanatory variables. Here, we introduce the exponentiated sinh Cauchy (Exp-SC) regression model in the GAMLSS family, allowing model the location, scale, bimodality and skewness parameters with the explanatory variables. Based on the Exp-ESC distribution, a linear regression model linking the response variable y_i and the explanatory variable can be defined by

$$y_i = \mu_i + \sigma_i\, z_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the random error $z_i = (y_i - \mu_i)/\sigma_i$ follows the density function

$$f(z;\nu,\tau) = \tau\, g_{\mathcal{SC}}(z)\, G_{\mathcal{SC}}(z)^{\tau-1} = \frac{\tau\nu}{\pi}\, \frac{\cosh(z)}{\nu^2 \sinh^2(z) + 1} \left\{ \frac{1}{2} + \frac{1}{\pi} \arctan\left[\nu \sinh(z)\right] \right\}^{\tau-1}. \tag{2}$$

Plots of the density function (2) for selected parameter values are shown in Figure 1.
Let $\boldsymbol{\theta}^T = (\mu, \sigma, \nu, \tau)$ the vector of parameters modelled here by using additive models. We can define the vector of parameters $\boldsymbol{\theta}$ using a appropriated link function as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\sigma} \\ \boldsymbol{\nu} \\ \boldsymbol{\tau} \end{bmatrix} = \begin{bmatrix} g_1(\mathbf{X}_1\boldsymbol{\beta}_1) \\ g_2(\mathbf{X}_2\boldsymbol{\beta}_2) \\ g_3(\mathbf{X}_3\boldsymbol{\beta}_3) \\ g_4(\mathbf{X}_4\boldsymbol{\beta}_4) \end{bmatrix} \tag{3}$$

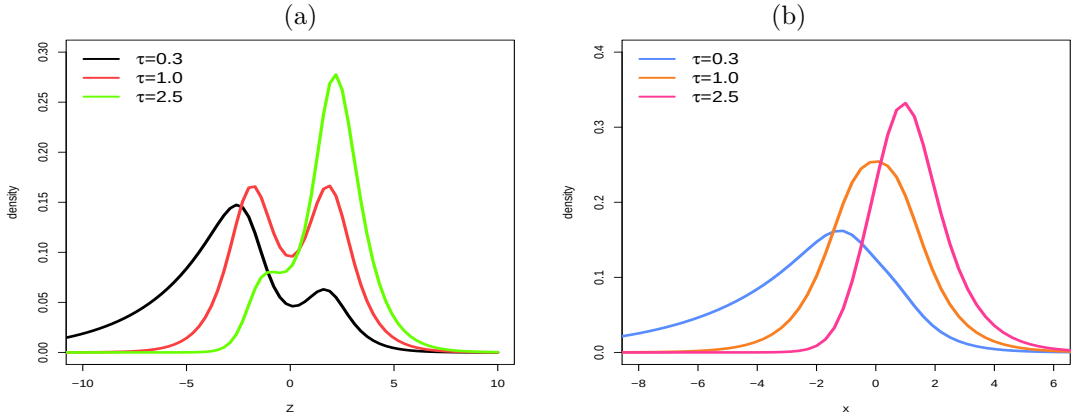


Figure 1: Plots of density function (2) for different values of τ considering: (a) $\nu = 0.3$; (b) $\nu = 0.8$

where $g_k(.)$, for $k = 1, 2, 3, 4$, represents the link functions contains values of the explanatory variables, $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \ldots, \beta_{p_k k})^T$ is a parameter vector of length $(p_k + 1)$ and $\mathbf{X}_k$ is a known design matrix of order $n \times (p_k + 1)$. The total number of parameters to be estimated is defined by $p = p_1 + p_2 + p_3 + p_4 + 4$.

Consider a sample of n-independent observations, where each random response is defined by $y_i = \min[\log(t_i), \log(c_i)]$. We assume non-informative censoring and that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which y_i is the log-lifetime or log-censoring, respectively. The total log-likelihood function for the model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})^T$, from model (1), has the form $l(\boldsymbol{\theta}) = \sum_{i \in F} \log f(y_i; \boldsymbol{\theta}_i) + \sum_{i \in C} \log S(y_i; \boldsymbol{\theta}_i)$. The log-likelihood function for $\boldsymbol{\theta}$ reduces to

$$l(\boldsymbol{\theta}) = -\sum_{i \in F} \log\left[1 + \nu_i^2 \sinh^2(z_i)\right] + \sum_{i \in F} \log\cosh(z_i) + (\tau_i - 1) \sum_{i \in F} \log\left\{\frac{1}{2} + \frac{1}{\pi}\arctan[\nu_i \sinh(z_i)]\right\}$$
$$+ \sum_{i \in F} \log(\tau_i \nu_i) - \sum_{i \in F} \log(\sigma_i \pi) + \sum_{i \in C} \log\left(1 - \left\{\frac{1}{2} + \frac{1}{\pi}\arctan\left[\nu_i \sinh\left(z_i\right)\right]\right\}^{\tau_i}\right). \quad (4)$$

The numerical maximization of the log-likelihood function (4) can be performed in the GAMLSS package in R. The maximization algorithms used are the RS and CG procedures. All methods and algorithms are described by [2] and [3] available in the documentation of the GAMLSS package.

Under standard regularity conditions, the asymptotic distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is $N_p(\mathbf{0}, I(\boldsymbol{\theta})^{-1})$, where $I(\boldsymbol{\theta})$ is the expected information matrix. This asymptotic behavior is valid if $I(\boldsymbol{\theta})$ is replaced by $\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})$, i.e., the observed information matrix evaluated at $\widehat{\boldsymbol{\theta}}$ given by $\ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T}$. The multivariate normal $N_p(\mathbf{0}, \ddot{\mathbf{L}}(\widehat{\boldsymbol{\theta}})^{-1})$ distribution can be used to construct approximate confidence intervals for the individual parameters.

2. APPLICATION

Consider the data set relating to biometric measurements in the shrimp of *farfantepenaeus brasiliensis* species. These data were obtained from three regions of the Rio Grande do Norte state in Brazil, in which the objective was to relate the weight of shrimp in each region. The importance of characterizing the weight of shrimp per region is because the marketing is done by weight, see [1] for more details.

Here, we consider the full sample ($n = 120$), where the response variable t_i represents the ith shrimps weight in grams and the three groups of region variable are defined by the dummy variables as: *Baia formosa*($x_{i1} = 0$ and $x_{i2} = 0$), *Diogo Lopes*($x_{i1} = 1$ and $x_{i2} = 0$) and *Touros*($x_{i1} = 0$ and $x_{i2} = 1$). Let the random variable $y_i = \log(t_i)$ follows the Exp-SC distribution and, with a preliminary analysis, we can be observed that the explanatoy variable *region* affects the location, scale, bimodality and asymmetry parameters. This fact can be easily observed in Figure 2.

Now, we present results by fitting the model

$$y_i = \mu_i + \sigma_i z_i,$$

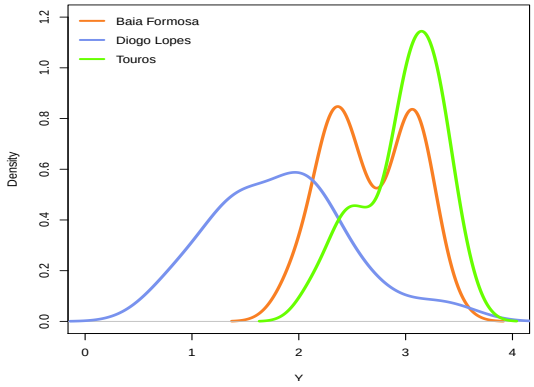


Figure 2: The true density of Y considering the different regions.

where z_i has density function (2) and the model parameters are defined by

$$\mu_i = \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2}, \qquad \sigma_i = \exp(\beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}),$$
$$\nu_i = \exp(\beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2}) \quad \textit{and} \quad \tau_i = \exp(\beta_{04} + \beta_{14}x_{i1} + \beta_{24}x_{i2}).$$

Table 1 provides the MLEs, standard errors and p-values calculated from the Exp-SC GAMLSS regression model. Still in Table 1, we conclude that the explanatory variable *region* should be used to model the location, scale, bimodality and skewness parameters at the 5% level. Therefore, we can conclude that for each region, the weight of shrimp has different forms (bimodal and unimodal), different location scales and asymmetry, and then, they can not be fitted only with a location-scale model.

Table 1: MLEs of the parameters from the fitted Exp-SC GAMLSS regression model to the shrimp data.

Parameter	Estimate	SE	p-value	Parameter	Estimate	SE	p-value
β_{01}	2.721	0.034	<0.001	β_{03}	-2.616	0.613	<0.001
β_{11}	-1.163	0.398	0.004	β_{13}	2.059	0.777	0.009
β_{21}	0.594	0.091	<0.001	β_{23}	2.425	0.754	0.001
β_{02}	-2.235	0.175	<0.001	β_{04}	-0.189	0.232	0.416
β_{12}	1.223	0.387	0.002	β_{14}	0.655	0.713	0.360
β_{22}	-0.057	0.495	0.908	β_{24}	-1.165	0.595	0.052

In order to assess if the model is appropriate, the empirical cdf and estimated cdf of the Exp-SC regression model are plotted in Figure 3 for different regions levels. We conclude that the Exp-ESC regression model provides a good fit to these data.

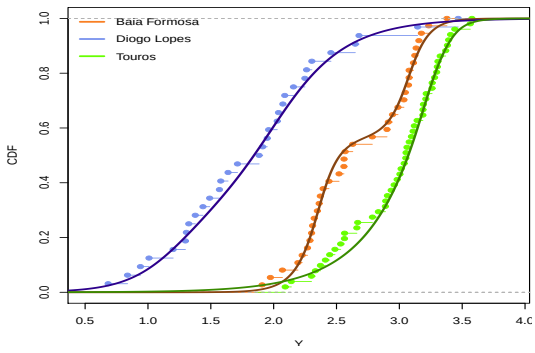


Figure 3: Estimated cumulative Exp-SC regression models and empirical cdf for shrimp data.

To detect possible outlying observations as well as departures from the assumptions of Exp-SC regression model, we present in Figure 4, the graphs of normal quantile residual and normal Q-Q plot. We can observe that the normalized quantile residuals of this model are approximately normally distributed indicating a good model fit.

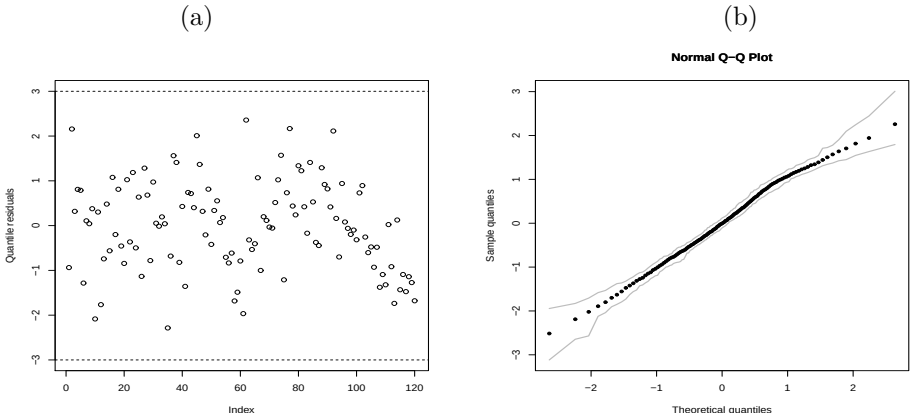


Figure 4: Residuals plots from the Exp-SC regression model fitted to shrimp data.

3. CONCLUSÕES

we propose a new general class of exponentiated sinh Cauchy (Exp-SC) regression models, where mean, dispersion, skewness and bimodal parameters vary across observations through regression structures. The new class of regression models is very suitable for modeling censored and uncensored lifetime data. The proposed model serves as an important extension to several existing regression models and could be a valuable addition to the literature. We use the GAMLSS script in the R package to obtain the maximum likelihood estimates and perform asymptotic tests for the parameters based on the asymptotic distribution of these estimates. Moreover, the usefulness of the model is also demonstrated through the an analysis of real data set. Finally, the new model has been coded and implemented in the GAMLLS script of the software R [4].

References

[1] Pinheiro, A. P. (2010). Caracterização genética e biométrica das populações de camarão rosa Farfantepenaeus brasiliensis de três localidades da costa do Rio Grande do Norte.

[2] Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–554.

[3] Stasinopoulos, D.M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1–46.

[4] Team, R.C. (2000). R Language Definition.

II Encontro Galaico-Português de Biometría
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

DADOS DE ESPETROMETRIA GAMA PORTÁTIL E TIPOLOGIA DE ESPAÇOS INTERIORES

Marco Lima[1], Jorge Sanjurjo-Sánchez[2] e Carlos Alves[3]

[1] LandS/Lab2PT - Laboratório de Paisagens, Património e Território (FCT UID/AUR/04509/2013; FEDER COMPETE POCI-01-0145-FEDER-007528)
[2] Instituto Universitario de Xeoloxía "Isidro Parga Pondal", Universidade da Coruña, A Coruña, Espanha
[3] LandS/Lab2PT - Laboratório de Paisagens, Património e Território (FCT UID/AUR/04509/2013; FEDER COMPETE POCI-01-0145-FEDER-007528) e Escola de Ciências, Universidade do Minho, Braga, Portugal

RESUMO

Neste trabalho são discutidos resultados de análise multivariada de estimativas obtidas por espectrometria gama portátil em espaços interiores de construções na cidade de Braga, com diferentes proporções de granito, para avaliar uma proposta de tipologia dos mesmos.

Palavras e frases chave: granito, materiais, perigosidade, ambiente construído, análise multivariada

1. INTRODUÇÃO

Os riscos radiológicos incluem a radiação gama externa proveniente de várias fontes incluindo materiais de construção como os materiais geológicos. As principais fontes de radiação gama nos materiais de construção são o isótopo ^{40}K (potássio) e isótopos das séries de decaimento do urânio e do tório presentes em alguns minerais. Como referido pelo Conselho da União Europeia [1], algumas rochas são consideradas merecedoras de especial atenção por apresentarem, frequentemente, valores elevados (em termos naturais) destes isótopos, como é o caso das rochas graníticas. A avaliação da perigosidade em termos de radiação gama externa é feita sobretudo para locais interiores e precisa de considerar os teores destes isótopos radioativos nos materiais mas também (ver, por exemplo, Markkanen [2]) a quantidade desses materiais presentes, a distribuição espacial dos mesmos e as caraterísticas geométricas do espaço considerado (nomeadamente a distância à fonte), assim como o tempo de exposição. Em geral, os espaços interiores (nomeadamente nas construções históricas) mostram uma grande variabilidade não só em dimensões mas também em termos do tipo e quantidade de material aplicado e da distribuição do mesmo no espaço considerado. A proposta do Conselho da União Europeia [1] refere a necessidade de realizar estudos específicos para materiais de construção que apresentem valores superiores ao valor de referência de um índice calculado para um determinado cenário de aplicação de materiais (que pode não corresponder a vários espaços históricos e atuais). Pelo contrário, outros documentos, como um de 1999 da Comissão Europeia [3], indicam diferentes valores de referência em função da extensão da aplicação dos materiais. Este é um tema que poderá vir a ter implicações negativas para uma componente importante da atividade económica da região da Galiza e do norte de Portugal. A espetrometria gama portátil está entre os métodos radiométricos de terreno indicados em IAEA (International Atomic Energy Agency) [4] e consiste numa sonda transportada por um operador que permite obter espetros com contagens de fotões em determinados valores de energia, sendo que é possível relacionar os valores de energia com isótopos radioativos específicos. Estes sistemas têm mostrado potencial para o mapeamento de parâmetros relacionados com a radiação gama mesmo em ambientes urbanos complexos [5]. Ainda que a preocupação deste estudo seja a radiação gama externa, as estimativas obtidas por espetrometria gama portátil podem ter interesse para o estudo de outros riscos radiológicos como os relacionados com as emissões de Rn (rádon), uma vez que

este gás resulta do decaimento radioativo do urânio e, se todas as outras condições forem iguais, um teor superior em urânio implicará maior perigosidade em termos de Rn. A análise multivariada tem sido utilizada na interpretação de dados de campo de radiação gama há bastante tempo (há pelo menos um exemplo de 1973 por Schwarzer & Adams [6]). Num estudo anterior sobre medições de campo de radiação gama em terrenos de diferentes tipos de rochas utilizamos a análise grupal para comparar os resultados em granitos e rochas metassedimentares [7]. Todavia, não encontramos nenhuma publicação com um objetivo como o aqui considerado. Neste trabalho, com base em dados de espetrometria gama portátil, tentamos avaliar a utilização da análise grupal hierárquica para a discussão de propostas de categorias ordinais das construções em termos de isótopos radioativos que contribuem para a dose de radiação gama externa resultante dos materiais de construção.

2. OBJECTOS E MÉTODOS DE ESTUDO

Os espetros de radiação gama foram adquiridos com um espetrómetro portátil *GF Instruments* GRS-2000, transportado por um operador com uma sonda BGO ($Bi_4Ge_3O_{12}$) de 512 canais (medindo energias até 3 MeV –Megaelétron Voltios). Cada espetro corresponde a um período de aquisição de 180 segundos. A partir destes espetros são obtidas estimativas de teores de K (potássio), eU (equivalente de urânio) e eTh (equivalente de tório). O teor em potássio (K, %) é estimado partir do pico do ^{40}K. As séries de decaimento do urânio e do tório têm vários isótopos radioativos e, assumindo equilíbrio secular, são obtidas estimativas de equivalente de urânio (eU, ppm) e equivalente de tório (eTh, ppm) a partir dos picos do ^{214}Bi (bismuto) e ^{208}Tl (tálio), respetivamente. As estimativas apresentadas são expressas em diferentes escalas uma vez que o potássio é um elemento maior na crusta e apresenta nos materiais geológicos teores muito superiores aos teores de urânio ou tório (que são elementos traço). Com as estimativas obtidas, foram calculados valores de taxa de dose absorvida (Dt) com base nos fatores indicados em [4] e os fatores indicados em [2] para um "quarto" padrão. A parte histórica da cidade de Braga (NW de Portugal) está localizada em terrenos graníticos e o granito é um dos principais materiais nas construções históricas (em quantidades variáveis). Os espetros gama considerados foram adquiridos em espaços interiores de diversos locais desta cidade que serão divididos em 4 tipos (A, B, C e D). A ordem alfabética corresponde a uma, estimada, influência decrescente dos materiais graníticos, de acordo com a descrição seguinte:
- Tipo A, cinco espetros gama em espaços interiores de construções históricas que apresentam uma pequena dimensão e onde o granito é o material claramente dominante;
- Tipo B, cinco espetros gama em espaços interiores de construções históricas onde o granito é o material dominante e que apresentam uma dimensão superior às dimensões do tipo anterior;
- Tipo C, três espetros em espaços interiores de construções históricas onde há elementos em granito mas onde há outros materiais que, em termos grosseiros, são tão ou mais importantes;
- Tipo D, um espetro num espaço interior de uma construção contemporânea (último quarto do século XX) onde não há materiais graníticos à vista.

A numeração dentro de cada tipo segue a ordem de medição dentro de cada tipo (mas não necessariamente a ordem global), servindo para a identificação das medições na discussão dos resultados.

Considerando que os parâmetros são expressos em escalas diferentes, e para evitar o efeito de variações de valores mais elevadas, foram utilizados, nas análises multivariadas, os valores normalizados (média igual a zero e desvio-padrão igual a 1) das estimativas de K, eU e eTh. As análises estatísticas e os gráficos aqui apresentados foram preparados com o *Statistica* 11 (Statsoft).

3. RESULTADOS E DISCUSSÃO

Na Figura 1a são projetados os valores normalizados das estimativas de K, eU e eTh, assim como os valores normalizados da taxa de dose absorvida (Dt) calculada a partir destas. O gráfico apresentado é do tipo caixa-de-bigodes com as diferentes medições individualizadas e onde os valores acima e abaixo das extremidades dos "bigodes" serão considerados potenciais "outliers". O gráfico sugere uma certa tendência para valores mais elevados, nos diferentes parâmetros, em espetros dos espaços de tipo A, tendência que fica mais evidente nos valores de taxa de dose absorvida. Também é visível que há medições abaixo do limite do "bigode" inferior: B5 para todos os parâmetros e também D para eTh e taxa de dose absorvida. Se os resultados para D seriam expectáveis, de acordo com a tipologia apresentada, o resultado B5 merece atenção. Este poderá, simplesmente, resultar de problemas de medição mas, uma vez que há medições antes e após B5 que não aparentam ter problemas, vamos considerar uma hipótese alternativa mais interessante

do ponto de vista da investigação da perigosidade radiológica. O espaço onde foi adquirido o espetro B5 corresponde a um depósito de publicações periódicas, com estruturas que podem atenuar a radiação (vários armários com as referidas publicações). Utilizando os valores normalizados de K, eU e eTh foram experimentadas análises hierárquicas aglomerativas considerando várias regras de agregação. Todas as análises hierárquicas consideradas juntaram os espetros D e B5 num primeiro "cluster" separado de todos os outros. As análises pelo método de Ward (Figura 1b) e pelo UPGMA juntam, seguidamente, as medições em locais de tipo A num "cluster" separado das outras medições. Algumas medidas de distância com estes procedimentos (por exemplo a distância euclidiana) juntam também num "cluster" as medições B1, B2 e B4 e, noutro cluster, as medições B3 e as medições nos espaços de tipo C. Outras regras podem produzir resultados variados que tem em comum juntarem A3 e A4 (após a separação de B5 e D) e colocarem as outras medições em espaços de tipo A num cluster com as restantes medições. Os valores normalizados de K, eU e eTh foram estudados, também, por análise de componentes principais. Em termos de associação das medições com as componentes principais calculadas e considerando (Figura 1c) as duas primeiras (CP1 e CP2), podemos observar que a CP1 separa claramente B5 e a medição no espaço tipo D e que as medições em espaços do tipo A são separadas das medições nos espaços dos outros tipos. A CP2 tende a separar as medições em espaços de tipo B das medições em espaços de tipo C (mas B3 aparece entre as medições dos espaços de tipo C), divide as medições nos espaços tipo A em dois grupos semelhantes aos "clusters" definidos e separa B5 de D. Em termos de associação com as variáveis (Figura 1d), K, eU e eTh têm pesos semelhantes (e negativos) em CP1. A taxa de dose absorvida (Dt), calculada de acordo com o procedimento acima indicado, foi projetada como variável suplementar (não utilizada no cálculo das componentes principais) e também tem um comportamento semelhante em termos de CP1. A CP2 mostra pesos quase simétricos para eU e eTh, sugerindo a influência da relação entre os dois.

O limitado número de casos considerados aconselha uma elevada prudência nas ilações a retirar. Todavia, os resultados obtidos sugerem que análise grupal poderá ser uma ferramenta importante na definição de tipologias de construção para uma cartografia da perigosidade relacionada com radiação gama externa. As análises pelo método de Ward e pelo UPGMA mostraram os resultados mais próximos da tipologia inicialmente proposta com base na avaliação qualitativa da importância da presença dos materiais graníticos e aparentam ser convergente com os resultados da análise de componentes principais. Os resultados obtidos sugerem, também, que o efeito na dose de radiação gama externa de aspetos como a presença de elementos que diminuem a radiação (armários, papel) poderá merecer um estudo mais atento, e mesmo estimativas do fator de atenuação considerando propriedades físicas como a densidade ou número atómico (Z). Poderá existir um outro aspeto, potencialmente preocupante, nesta questão: a redução das estimativas de campo dos teores em urânio poderá não corresponder a uma redução das emissões de Rn, o que prejudicará a utilidade desta técnica na avaliação de riscos radiológicos.

AGRADECIMENTOS

O Lab2PT - Laboratório de Paisagens, Património e Território (UID/AUR/04509/2013) é financiado pela Fundação para a Ciência e Tecnologia (Portugal), com verbas da República Portuguesa e da União Europeia (FEDER, Programa Operacional Factores de Competitividade – COMPETE 2020 – Programa Operacional Competitividade e Internacionalização, POCI, POCI-01-0145-FEDER-007528). Gostaríamos de agradecer também à Mara Dória e ao Tomaz Assunção pela colaboração no trabalho de Campo.

REFERÊNCIAS

[1] DIRETIVA 2013/59/EURATOM (2014) Normas de segurança de base relativas à proteção contra os perigos resultantes da exposição a radiações ionizantes, e que revoga as Diretivas 89/618/Euratom, 90/641/Euratom, 96/29/Euratom, 97/43/Euratom e 2003/122/Euratom. Jornal oficial da União Europeia. L13/1-73.

[2] Markkanen, M. (1995). *Radiation Dose Assessments for Materials with Elevated Natural Radioactivity*. Report STUK-B-STO 32 Radiation and Nuclear Safety Authority – STU. Finnish Centre for Radiation and Nuclear Safety, Helsínquia.

[3] Comissão Europeia (1999). *Radiation Protection 112 - Radiological protection principles concerning the natural radioactivity in building materials*. Luxemburgo.

[4] IAEA - International Atomic Energy Agency (2003*) Guidelines for radioelement mapping using gamma ray spectrometry data. IAEA-TECDOC-1363*. Viena.

[5] Cresswell, A. J., Sanderson, D. C. W., Harrold, M., Kirley, B., Mitchell, C., Weir, A. (2013). Demonstration of lightweight gamma spectrometry systems in urban environments. *Journal of Environmental Radioactivity*, 124, 22–28.

[6] Schwarzer, T. F., & Adams, J. A. S. (1973). Rock and Soil Discrimination by Low Altitude Airborne Gamma-Ray Spectrometry in Payne County, Oklahoma. *Economic Geology*, 68(8), 1297–1312.

[7] Dias, F., Lima, M., Sanjurjo-Sánchez, J., Alves, C. (2016). Analysis of spectra from portable handheld gamma-ray spectrometry for terrain comparative assessment. *Journal of Environmental Radioactivity*, 154, 93–100.

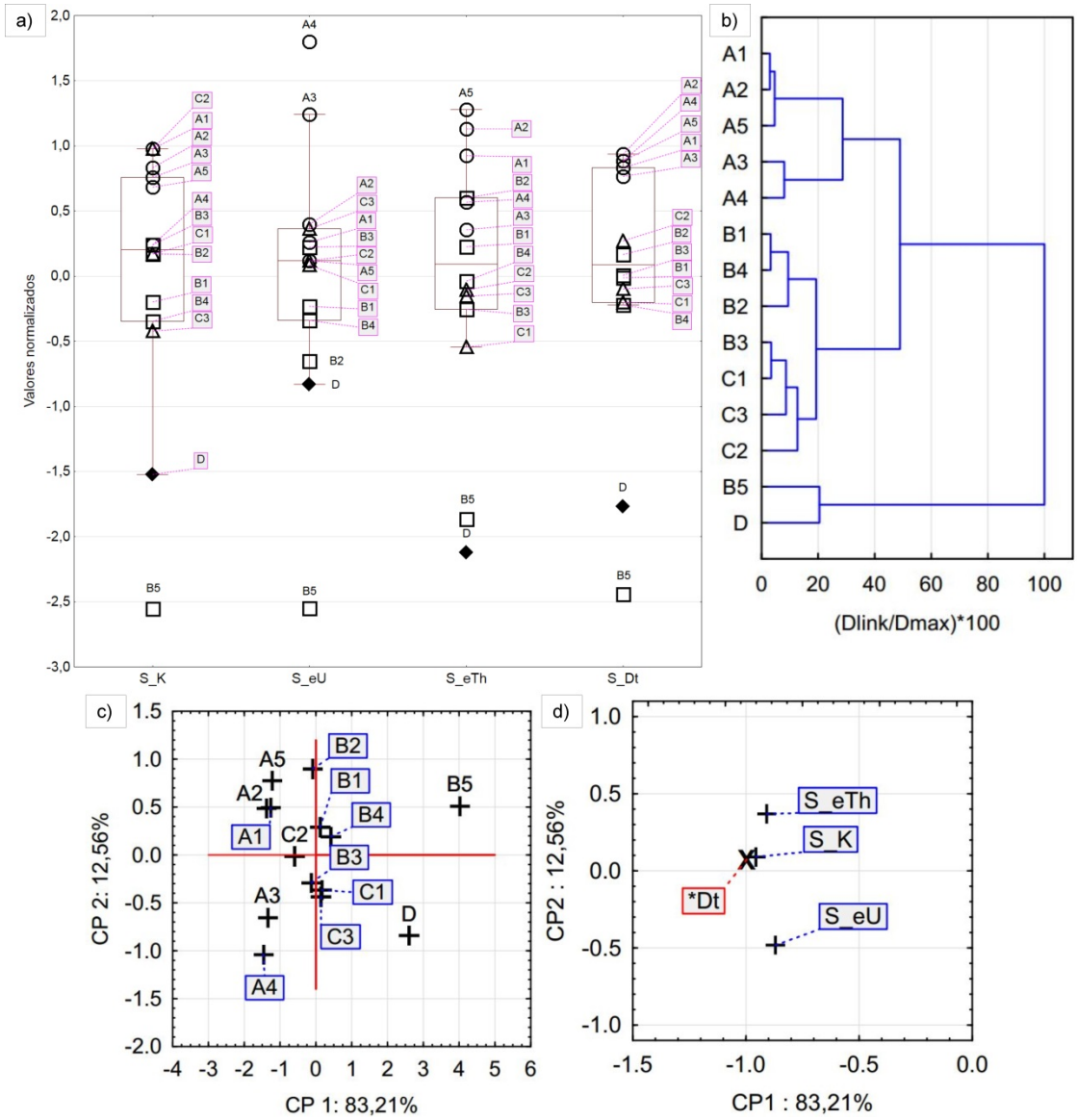


Figura 1: Estudos estatísticos dos resultados (valores normalizados) obtidos por espetrometria gama portátil: a) Diagrama de caixa-de-bigodes assinalando os resultados de K (S_K), eU (S_eU) e eTh (S_eTh), assim como da taxa de dose absorvida calculada a partir dos mesmos (S_Dt); b) Dendrograma (Método de Ward) considerando os resultados de K, eU e eTh; c) Análise de Componentes Principais dos locais estudados (resultados de K, eU e eTh); d) Análise de Componentes Principais das variáveis consideradas com projeção da taxa de dose absorvida (Dt) como variável suplementar (não utilizada na Análise de Componentes Principais).

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

DINÂMICA POPULACIONAL MENDELIANA

Maria de Fátima Brilhante[1], Dinis Pestana[2] Paulo Semblano[3] e Fernando Sequeira[4]

[1] Universidade dos Açores, Departamento de Matemática e Centro de Estatística e Aplicações da Universidade de Lisboa, fbrilhante@uac.pt
[2] Centro de Estatística e Aplicações da Universidade de Lisboa e Instituto de Investigação Científica Bento da Rocha Cabral, dinis.pestana@fc.ul.pt
[3] Caixa Geral de Depósitos e Centro de Estatística e Aplicações da Universidade de Lisboa, PauloSemblano@cgd.pt
[4] Universidade de Lisboa, FCUL/DEIO e Centro de Estatística e Aplicações da Universidade de Lisboa, fjsequeira@fc.ul.pt

RESUMO

As misturas de uniforme e de Beta(2,1) ou Beta(1,2) — estas associadas respetivamente às densidades do máximo e do mínimo de duas uniformes independentes — têm relevância na análise de processos de Mendel, em que o experimentador faz (porventura involuntariamente) batota, reportando o melhor dos resultados quando repete a experiência por o primeiro resultado obtido não ser conveniente para as suas expetativas. Estudamos a dinâmica populacional numa extensão da equação de Verhulst em que quer o fator de incremento quer o fator retroativo clássicos são substituídos por expressões analíticas associadas às referidas misturas.

Palavras e frases chave: Modelo de Verhulst, valores de prova mendelianos, misturas, variáveis beta.

1. INTRODUÇÃO

Fibonacci (c. 1170 – c. 1250) no seu *Liber Abaci* propôs um modelo populacional $\{F_n\}$, $F_0 = 0$, $F_1 = 1$, $F_{n+2} = F_n + F_{n+1}$. O cálculo do efetivo populacional em cada geração F_n pode ser feito de forma simples usando a fórmula de Binet $F_n = \frac{(1+\sqrt{5})^{n+1} - (1-\sqrt{5})^{n+1}}{2^{n+1}\sqrt{5}}$.
Rescrevendo $F_{n+2} = F_n + F_{n+1}$ sob a forma $F_{n+2} - F_{n+1} = F_{n+1} - F_{n-1}$ e aproximando esta equação às diferenças pela equação diferencial $\frac{\mathrm{d}}{\mathrm{d}t} N(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} N(t)$ conclui-se que o modelo de Fibonacci modela um crescimento geométrico de razão $\left(1+\sqrt{5}\right)/2$, o famoso "número de ouro" que é o limite de F_{n+1}/F_n. (Mesmo para valores tão modrados quanto $n = 11$, por exemplo, $F_{12} = 144 \approx F_{11}\left(1+\sqrt{5}\right)/2 = 144{,}05$.) Assim, no modelo de Fibonacci há sempre crescimento, o que contraria o conhecimento empírico da possibilidade de extinção de famílias e de espécies, veja-se por exemplo Lotka (1925).
Usando aleatoriedade (processos de ramificação, processos de nascimento e morte) ou modelos deterministas tendo em linha de conta o efeito das populações no consumo de recursos naturais é possível desenvolver modelos que contemplem crescimento sustentável, e em que a extinção é um acontecimento possível.

Verhulst (1838, 1845, 1847) abordou a questão admitindo que a dimensão da população $N(t)$ no instante t é "regular"no sentido em que admite uma representação $\mathrm{d}N(t)/\mathrm{d}t = \sum_{k=0}^{\infty} A_k[N(t)]^k$ (com as hipóteses plausíveis $A_0 = 0$ porque uma população extinta não tem descendência, $A_1 > 0$ porque é necessário assegurar a possibilidade de reprodução, e $A_2 < 0$ uma retroação que regula o crescimento devido à limitação de recursos naturais), usou a aproximação de segunda ordem

$$\frac{\mathrm{dN(t)}}{\mathrm{d}t} \approx r\,N(t)\left[1 - \frac{N(t)}{K}\right],$$

com a interpretação usual r=taxa mendeliana de crescimento instantâneo, K um limite de equilíbrio imposto pelo desenvolvimento sustentável. Assim, a dimensão da população é aproximada pela função logística $N(t) = \dfrac{K\,N_0}{N_0 + (K - N_0)\,\mathrm{e}^{-rt}}$.
Quer o modelo de Fibonacci vieram a ter grande protagonismo em muitas áreas da Matemática. Nomeadamente, a *parábola logística* associada à discretização da equaccão de Verhulst teve um papel singular na teoria das bifurcações de Feigenbaum e dinâmica caótica. Daí resultou um grande investimento em extensões da equação de Verhulst, veja-se por exemplo Tsoularis (2001).
Observando em particular que a parábola logística é proporcional à densidade Beta(2,2), tornou-se interessante o estudo de equações diferenciais associadas a betas, e em particular proporcionais a densidades de máximos ($2x$) e mínimos ($2(1-x)$) ou a produtos ($-\ln x$) de duas variáveis uniformes independentes, veja-se Aleixo *et al.* (2011), Pestana *et al.* (2011), Brilhante *et al.* (2011, 2012).
Observe-se que

1. associado ao núcleo (*kernel*) uniforme, $\mathrm{d}N(t)/\mathrm{d}t = r$ tem a solução linear $N(t) = Crt$;
2. associado ao núcleo $Beta(2,1)$, $\mathrm{d}N(t)/\mathrm{d}t = rN(t)$ tem a solução exponencial $N(t) = C\exp(rt)$;
3. associado ao núcleo $Beta(1,2)$, $\mathrm{d}N(t)/\mathrm{d}t = r\,[1 - N(t)/K]$ tem uma solução exponencial $N(t) = K - C\exp(rt)$;
4. associado ao núcleo Beta(2,2), como já vimos, tem-se a solução logística.

2. EXTENSÕES MENDELIANAS DO MODELO DE VERHULST

O passo seguinte é investigar extensões da equaccão de Verhulst em que o equilíbrio seja governado por núcleos mendelianos, isto é

$$\mathrm{d}N(t)/\mathrm{d}t = r\left[1 - \frac{m}{2} + mN(t)\right]\left[1 - \frac{p}{2} + pN(t)\right] \tag{1}$$

onde $[1 - \frac{m}{2} + mN(t)]$, $m \in [0,2]$ regula o incremento populacional e $[1 - \frac{p}{2} + pN(t)]$, $p \in [-2,0]$ é uma retroação. (A equação de Verhulst é o caso especial $m = 2$, $p = -2$.)
Observe-se que o primeiro fator está associado a uma mistura de uniforme (Beta(1,1)) e de máximo de duas uniformes independentes (Beta (2,1)), e o segundo fator está associado a uma mistura de uniforme e de mínimo de duas uniformes independentes (Beta(1,2)). Estas misturas foram estudadas por Gomes *et al.* (2009), e Pires e Branco (2010) usaram-nas para investigar a controvérsia de Mendel–Fisher, sobre a possibilidade de haver o uso de valores de prova-p problemáticos, obtidos com repetição de testes quando o resultado não é o desejado pelo investigador. É por essa razão que apelidamos de "mendelianos.os procedimentos em que uma fração dos resultados reportados corresponde a uma escolha do mais favorável aos intuitos do experimentador quando não obtém o que esperava, não havendo em geral uma explicitação de que tal foi feito, nem em que proporção.
Dividindo ambos os membros de (1) por $(1 + m(N(t) - \frac{1}{2}))(1 + p(N(t) - \frac{1}{2}))$ obtém-se

$$\frac{1}{(1 + m(N(t) - \frac{1}{2}))(1 + p(N(t) - \frac{1}{2}))}\frac{\mathrm{d}N(t)}{\mathrm{d}t} = r,$$

e integrando em ordem a t

$$\int \frac{1}{(1+m(N(t)-\frac{1}{2}))(1+p(N(t)-\frac{1}{2}))}\frac{\mathrm{d}N(t)}{\mathrm{d}t}\,\mathrm{d}t = \int r\,\mathrm{d}t.$$

Simplifique-se o primeiro membro da equação mudando para a variável de integracão N:

$$\int \frac{1}{(1+m(N(t)-\frac{1}{2}))(1+p(N(t)-\frac{1}{2}))}\,\mathrm{d}N(t) = \int r\,\mathrm{d}t$$

e integre-se usando frações parciais:

$$\frac{1}{(1+m(N(t)-\frac{1}{2}))(1+p(N(t)-\frac{1}{2}))} = \frac{A}{(1+m(N(t)-\frac{1}{2}))} + \frac{B}{(1+p(N(t)-\frac{1}{2}))}.$$

Consequentemente

$$Ap + Bm = 0 \quad \Longleftrightarrow \quad A = -\frac{Bm}{p}, \tag{2}$$

$$A - \frac{Ap}{2} + B - \frac{Bm}{2} = 1. \tag{3}$$

Substituindo A por $-Bm/p$ em (3), tem-se

$$-\frac{Bm}{p} + \frac{Bm}{2} + B - \frac{Bm}{2} = 1 \quad \Longleftrightarrow \quad B(1-\frac{m}{p}) = 1 \quad \Longleftrightarrow \quad B = \frac{p}{p-m}.$$

Então, de (2), $A = m/(m-p)$. Assim

$$\int \frac{\frac{m}{m-p}}{(1+m(N(t)-\frac{1}{2}))} - \frac{\frac{p}{m-p}}{(1+p(N(t)-\frac{1}{2}))}\,\mathrm{d}N(t) = \int r\,\mathrm{d}t$$

$$\Longleftrightarrow \quad \int \frac{m}{(1+m(N(t)-\frac{1}{2}))} - \frac{p}{(1+p(N(t)-\frac{1}{2}))}\,\mathrm{d}N(t) = (m-p)\int r\,\mathrm{d}t$$

$$\Longleftrightarrow \quad \ln\left|\frac{2mN(t)-m+2}{2pN(t)-p+2}\right| = (m-p)rt + C$$

e consequentemente

$$\frac{2mN(t)-m+2}{2pN(t)-p+2} = \mathrm{e}^{C}\,\mathrm{e}^{(m-p)rt} = C_0\,\mathrm{e}^{(m-p)rt}. \tag{4}$$

Portanto

$$2mN(t) - m + 2 = (2pN(t) - p + 2)C_0\mathrm{e}^{(m-p)rt}$$

$$\Longleftrightarrow \quad 2mN(t) = 2p\,N(t)\,C_0\,\mathrm{e}^{(m-p)rt} - p\,C_0\,\mathrm{e}^{(m-p)rt} + 2C_0\,\mathrm{e}^{(m-p)rt} + m - 2$$

$$\Longleftrightarrow \quad 2mN(t) - 2p\,N(t)\,C_0\,\mathrm{e}^{(m-p)rt} = (2-p)\,C_0\,\mathrm{e}^{(m-p)rt} + m - 2$$

$$\Longleftrightarrow \quad (2m - 2p\,C_0\mathrm{e}^{(m-p)rt})\,N(t) = (2-p)\,C_0\,\mathrm{e}^{(m-p)rt} + m - 2,$$

e finalmente

$$N(t) = \frac{(2-p)\,C_0\,\mathrm{e}^{(m-p)rt} + m - 2}{2m - 2p\,C_0\,\mathrm{e}^{(m-p)rt}} \tag{5}$$

(que para $m = 2,\ p = -2$ é o crescimento logístico).

AGRADECIMENTOS

FCT Financiado pela FCT - Fundação para a Ciência e a Tecnologia, Portugal, através do projecto UID/MAT/00006/2013.

Referencias

[1] Aleixo, S., Rocha, J. L., and Pestana, D. (2011). Probabilistic Methods in Dynamical Analysis: Population Growths Associated to Models Beta (p,q) with Allee Effect. In Peixoto, M. M; Pinto, A.A.; Rand, D. A. J., editors, *Dynamics, Games and Science, in Honour of Maurício Peixoto and David Rand*, vol II, 79–95, New York, 2011. Springer Verlag.

[2] Brilhante, M. F., Gomes, M. I., and Pestana, D. (2011). BetaBoop Brings in Chaos. *Chaotic Modeling and Simulation* 1, 39–50.

[3] Brilhante, M. F., Gomes, M. I., and Pestana, D. (2012). Extensions of Verhulst Model in Population Dynamics and Extremes. In Ch. Skiadas, editor, *Proceedings, 5th Chaotic Modeling and Simulation International Conference* 115–122.

[4] Gomes, M. I., Pestana, D., Sequeira, F., Mendonça, S., and S. Velosa, S. (2009). Uniformity of offsprings from uniform and non-uniform parents. In V. Luzar-Stiffler, I. Jarec, and Z. Bekic, editors, *Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces*, 243–248, IEEE, ISBN 978-953-7138-15-8.

[5] Lotka, A. J. (1925). *Elements of Physical Biology*, Williams and Wilkins Co, Baltimore (reprinted under the title *Elements of Mathematical Biology*, Dover, New York, 1956).

[6] Pestana, D., Aleixo, S., and Rocha, J. L. (2011) Regular variation, paretian distributions, and the interplay of light and heavy tails in the fractality of asymptotic models. In C. H. Skiadas, I. Dimotikalis and C. Skiadas, editors, *Chaos Theory: Modeling, Simulation and Applications* 309–316, World Scientific, Singapore.

[7] Pires, A. M., and Branco, J. A. (2010) A statistical model to explain the Mendel–Fisher controversy. *Statistical Science* 25, 545–565.

[8] Tsoularis, A. (2001) Analysis of logistic growth models. *Res. Lett. Inf. Math. Sci.* 2, 23–46.

[9] Verhulst, P.-F. (1838) Notice sur la loi que la population poursuit dans son accroissement. *Corresp. Math. Physics* 10, 113–121.

[10] Verhulst, P.-F. (1845). La loi de láccroissement de la population, *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles* 18, 1–42. (Partially reproduced in David, H. A., and Edwards, A. W. F. *Annotated Readings in the History of Statistics* Springer Verlag, New York, 2001.

[11] Verhulst, P.-F. (1847). Deuxième mémoire sur la loi dáccroissement de la population. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique* 20, 1–32.

II Encontro Galaico-Portugués de Biometría
Santiago de Compostela, 30 de xuño, 1 e 2 de xullo de 2016

MODELOS EXTREMAIS EM DINÂMICA DE POPULAÇÕES

Maria de Fátima Brilhante[1], Maria Ivette Gomes[2] e Dinis Pestana[3]

[1] Universidade dos Açores, Departamento de Matemática e Centro de Estatística e Aplicações da Universidade de Lisboa, fbrilhante@uac.pt
[2] Academia das Ciências de Lisboa, Centro de Estatística e Aplicações da Universidade de Lisboa e Instituto de Investigação Científica Bento da Rocha Cabral, migomes@fc.ul.pt
[3] Centro de Estatística e Aplicações da Universidade de Lisboa e Instituto de Investigação Científica Bento da Rocha Cabral, dinis.pestana@fc.ul.pt

RESUMO

A equação de Verhulst descreve uma dinâmica sustentável. Exploram-se extensões dessa equação cujos fatores estão associados à densidade de máximos, mínimos ou produtos de uniformes independentes, mostrando que retroações mais fracas do que o tradicional $[1 - N(t)/K]$ levam ao clássico modelo populacional de Gompertz, associado ao modelo extremal de Gumbel, podendo o resultado ser generalizado para obter soluções associadas aos modelos extremais de Fréchet e de Weibull. Isso leva-nos a uma reinterpretação da solução logística da equação de Verhulst como equíbrio sutentável associado a leis limites estáveis de máximos numa sucessão geometricamente emagrecida de variáveis independentes e identicamente distribuídas, e a uma extensão levando aos outros modelos geo-estáveis de extremos.

Palabras e frases chave: Equação de Verhulst, máximos, mínimos e produtos de uniformes independentes, leis estáveis extremais, leis estáveis geo-extremais.

1. INTRODUÇÃO

Verhulst (1845, 1847) revolucionou os estudos populacionais admitindo que a dimensão da população $N(t)$ no instante t é "regular"no sentido em que admite uma representação

$$\mathrm{d}N(t)/\mathrm{d}t = \sum_{k=0}^{\infty} A_k [N(t)]^k$$

usando a aproximação de segunda ordem

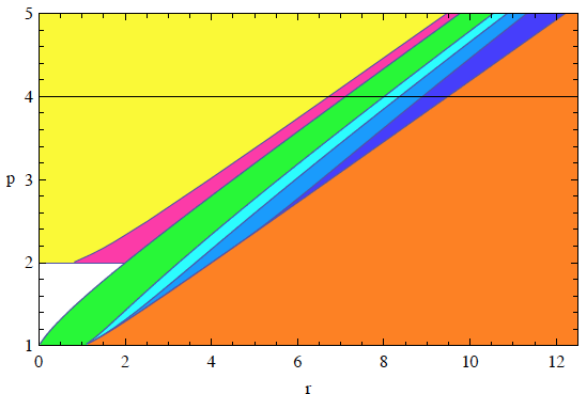
$$\mathrm{d}N(t)/\mathrm{d}t \approx r\, N(t) \left[1 - \frac{N(t)}{K}\right],$$

com a interpretação usual r=taxa mendeliana de crescimento instantâneo, K um limite de equilíbrio imposto pelo desenvolvimento sustentável. Assim, a dimensão da população é aproximada pela função logística

$$N(t) = \frac{K\, N_0}{N_0 + (K - N_0)\, \mathrm{e}^{-rt}},$$

que corresponde a uma dinâmica sustentável, em que o fator de expansão $N(t)$ e a retroação $1 - N(t)/K$ devida a fatores ambientais e limitação de recursos se equilibram.

A *parábola logística* associada à discretização da equação de Verhulst teve um papel singular na teoria das bifurcações de Feigenbaum e dinâmica caótica. Daí resultou um grande investimento em extensões da equação de Verhulst, veja-se por exemplo Tsoularis (2001).
Observando em particular que a parábola logística é proporcional à densidade Beta(2,2), tornou-se interessante o estudo de equações diferenciais associadas a betas. Aleixo *et al.* (2011) fizeram um estudo dinâmico exaustivo de modelos associados a outras Beta(p,2), em que a retroação, de ordem 1 ($q = 2$), não equilibra necessariamente o fator expansivo p, mostrando que para cada valor deste parâmetro se podem distinguir regiões topologicamente distintas,



nomeadamente

- região de extinção súbita, seguida de regiões de estabilidade com entropia topológica nula;
- região de duplicação de período (bifurcações) com entropia topológica constante;
- região caótica, onde a entropia topológica tem o valor máximo constante $\ln 2$;
- uma região de extinção decorrente do efeito de Allee;
- uma região de extinção diferida.

A parábola logística $x(1-x),\ x \in (0,1)$ é proporcional ao produto da densidade da Beta(2,1), $2x,\ x \in (0,1)$ pela densidade $2(1-x),\ x \in (0,1)$ da Beta(1,2), que modelam o máximo e o mínimo de duas variáveis uniformes independentes. Isso levou Pestana *et al.* (2011) a investigar o uso de um fator retroativo mais moderado, associado antes à densidade $-\ln x,\ x \in (0,1)$ do produto de duas uniformes independentes.
A solução de

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = r\,N(t)\,[-\ln\,N(t)]$$

encontra-se facilmente, e é também um clássico da dinâmica populacional: é a função de Gompertz, que em estatística é mais conhecida como densidade de Gumbel, associada a um índice de valores extremos $\xi = 0$, sendo $\xi \in \mathbb{R}$ o parâmetro fundamental em estatística de extremos.
É assim natural estudar

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = r\,N(t)\,[-\ln\,N(t)]^{q-1}$$

e mais geralmente ainda

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = r\,N^{p-1}(t)\,[-\ln\,N(t)]^{q-1}\,,$$

um estudo que foi iniciado em Brilhante *et al.* (2011, 2012). Note-se que a extensão

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = r\,[N(t)][-\ln(N(t))]^{1+\xi},$$

tem como soluções normalizadas a distribuição de Fréchet (para máximos) quando $\xi > 0$, e a distribuição max-Weibull quando $\xi < 0$.
De facto, tem ainda mais interesse o estudo de núcleos associados à densidade

$$\frac{x^{p-1}(1-x)^{q-a}(-\ln(1-x))^{\pi-1}(-\ln x)^{\rho-1}}{\int_0^1 x^{p-1}(1-x)^{q-a}(-\ln(1-x))^{\pi-1}(-\ln x)^{\rho-1}}, \quad x \in (0,1),$$

a que Pestana *et al.* (2011) chamaram densidade BetaBoop. Por outras palavras, interessa proceder ao estudo sistemático de

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = r\, N^{p-1}(t)\,[1-N(t)]^{q-1}\,[-\ln(1-N(t))]^{\pi-1}\,[-\ln\, N(t)]^{\rho-1}$$

em que os fatores $N^{p-1}(t)$ e $[-\ln(1-N(t))]^{\pi-1}$ modelam a expansão da população enquanto os fatores $[1-N(t)]^{q-1}$ e $[-\ln\, N(t)]^{\rho-1}$ são retroações que equilibram o crescimento populacional.
Os diversos equilíbrios são conhecidos apenas quando alguns dos expoentes são 1. Em particular quando $p = q = 2, \pi = \rho = 1$ temos a solução logística, e mais geralmente há combinações interessantes de p e q que com $\pi = \rho = 1$ levam a soluções log-logística ou sua simétrica. Por outro lado, $p = \rho = 2, q = \pi = 1$ leva como já se referiu à solução Gumbel, e outras combinações de p e ρ com $q = \pi = 1$ levam a soluções Fréchet e Weibull. Para pormenores, veja-se Brilhante *et al.* (2012).
Assim, extensões naturais da equação de Vershulst têm como solução equilíbrios dinâmicos que são mais adequados do que o modelo clássico quando se pretende modelar fenómenos de crescimento incontrolado, como o número de células cancerosas num tumor invasivo.

3. CONCLUSÕES

O peso relativo do fator expansivo e do fator retroativo nas extensões da equação diferencial de Verhulst levam-nos a obter modelos populacionais associados às leis extremais.
Quando a retroação $1-N(t)$ é moderada, usando-se antes $-\ln N(t)$, a solução é a classe de Gumbel, em vez do clássico equilíbrio logístico — sendo este afinal também uma lei estável extremal, mas no esquema max-geo estável estudado por Rachev and Resnick (1991) — isto é, a classe das funções de distribuição para as quais existem constantes $a_\theta > 0$ e $b_\theta \in \mathbb{R}$ tais que

$$F(a_\theta x + b_\theta) = \frac{\theta\, F(x)}{1-(1-\theta)\, F(x)}.$$

De facto, o emagrecimento geométrico leva a que se considere o máximo das variáveis remanescentes no esquema de rarefação de Rényi (1956), que necessariamente é estocasticamente inferior ao máximo de todas as variáveis. Quer Kovalenko (1965) quer Kozubowski (1994) apresentaram resultados definitivos no que se refere a leis limites estáveis, respetivamente no caso de variáveis positivas e no caso geral, que mostram que quer a log-logística quer a sua simetrização, que obtivemos quando usámos

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = r\,([N(t)])^{1+\xi}[1-N(t)]^{1-\xi}, \quad \xi \neq 1$$

como generalização da equação de Verhulst, são também leis max-geo estáveis.
Por outro lado usando potências de $-\ln N(t)$ obtemos soluções associadas às clássicas leis estáveis de Fréchet e de Weibull, tendo assim um largo espetro de possíveis equilíbrios dinâmicos, que podem ser de grande utilidade no estudo de desenvolvimentos explosivos, por exemplo de células em tecidos neoplásicos.
Note-se que se em vez de substituir $N(t)\,[1-N(t)]$ por $N^{p-1}(t)\,[-\ln N(t)]^{q-1}$ usarmos a substituição $[-\ln(1-N(t)]^{p-1}\,[1-N(t)]^{q-1}$, em vez de soluções associadas às clássicas leis de máximos obtemos soluções associadas às leis de mínimos.

AGRADECIMENTOS

FCT Financiado pela FCT - Fundação para a Ciência e a Tecnologia, Portugal, através do projecto UID/MAT/00006/2013.

Referencias

[1] Aleixo, S., Rocha, J. L., and Pestana, D. (2011). Probabilistic Methods in Dynamical Analysis: Population Growths Associated to Models Beta (p,q) with Allee Effect. In Peixoto, M. M; Pinto, A.A.; Rand, D. A. J., editors, *Dynamics, Games and Science, in Honour of Maurício Peixoto and David Rand*, vol II, 79–95, New York, 2011. Springer Verlag.

[2] Brilhante, M. F., Gomes, M. I., and Pestana, D. (2011). BetaBoop Brings in Chaos. *Chaotic Modeling and Simulation* 1, 39–50.

[3] Brilhante, M. F., Gomes, M. I., and Pestana, D. (2012). Extensions of Verhulst Model in Population Dynamics and Extremes. In Ch. Skiadas, editor, *Proceedings, 5th Chaotic Modeling and Simulation International Conference* 115–122.

[4] Kovalenko, I. N. (1965). On a class of limit distributions for rarefied flows of homogeneous events, *Lit. Mat. Sbornik* 5, 569–573, 1965. (Translation: On the class of limit distributions for thinning streams of homogeneous events, *Selected Transl. Math. Statist. and Prob.* (1971) 9 75– 81 Providence, Rhode Island.)

[5] Kozubowski, T. J. (1994) Representation and properties of geometric stable laws, *Approximation, Probability, and Related Fields*, 321–337, PlenumNew York.

[6] Pestana, D., Aleixo, S., and Rocha, J. L. (2011) Regular variation, paretian distributions, and the interplay of light and heavy tails in the fractality of asymptotic models. In C. H. Skiadas, I. Dimotikalis and C. Skiadas, editors, *Chaos Theory: Modeling, Simulation and Applications* 309–316, World Scientific, Singapore.

[7] Rachev, S. T. and Resnick, S.(1991). Max-geometric infinite divisibility and stability, *Communications in Statistics — Stochastic Models* 7, 191-218.

[8] Rényi, A (1956) A characterization of the Poisson process. *MTA Mat. Kut. Int. Kozl.* 1, 519–527. (English translation in *Selected Papers of Alfred Rényi, Vol. 1, 1948–1956*, P. Turán, editor, 622–279 Akadémiai Kiadó, Budapest, with a note by D. Szász on posterior developments untill 1976).

[9] Tsoularis, A. (2001) Analysis of logistic growth models. *Res. Lett. Inf. Math. Sci.* 2, 23–46.

[10] Verhulst, P.-F. (1845). La loi de láccroissement de la population, *Nouveaux Mémoires de l' Académie Royale des Sciences et Belles-Lettres de Bruxelles* 18, 1–42. (Partially reproduced in David, H. A., and Edwards, A. W. F. *Annotated Readings in the History of Statistics* Springer Verlag, New York, 2001.

[11] Verhulst, P.-F. (1847). Deuxième mémoire sur la loi dáccroissement de la population. *Mémoires de l' Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique* 20, 1–32.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

Screening and validation in PERSSILAA – the personalized ICT service for independent living and active ageing

Joana Fernandes[1], Antónia Turkman[1], Marília Antunes[1], Feridun Turkman[1], Stephanie Jansen-Kosterink[2], Javier Solana[3], Enrique Gómez[3], Amelia Rauter[1] and Miriam Vollenbroek[2]

[1]Universidade de Lisboa, Faculdade de Ciências
[2]Roessingh Research and Development, University of Twente
[3]Universidad Politécnica de Madrid

ABSTRACT

PERSSILAA is an ICT service directed to older adults (65+). It was designed with the purpose of promoting independent living and active ageing by offering screening tools to identify the frailty status and provide intervention and/or counselling in the physical, cognitive and nutritional domains to screened participants[1,2] .

The screening procedure is performed in two phases: an online self-assessment questionnaire (first screening) and a questionnaire administered by a trained person (second screening). Both phases' questionnaires are built using well known validated questionnaires in each of the three domains. Individuals are classified as frail, pre-frail or robust according to a set of cut-off points in the scales of the questionnaires.
Cluster analysis was used to study the discriminative power of the first screening procedure and probabilistic models were used to compute probabilities of membership in each class. A comparative study between the results of the first and the second screening allowed us to revise the set of initial cut-off points used on the first screening.

Posterior probabilities of membership were found to be an effective tool to identify subjects who are screened out by the first screening but still could benefit from the program. Also, the statistical modelling approach proved to perform well and that it can be used as a guide for evaluation of functional decline during follow-up.

Keywords and key sentences: PERSSILAA, screening, frailty, active ageing.

ACKNOWLEDGMENT

PERSSILAA is a EU FP7 project (FP7-ICT-2013-10; project number 610359).

REFERENCES

References

[1] *PERSSILAA Description of Work.*

[2] *https://perssilaa.com/*

II Encontro Galaico-Português de Biometría
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

Statistical techniques for sensorial characterization of wines from three different grape varieties of Vinho Verde Region.

Catarina Marques[1], Alice Vilela[2], Elisete Correia[3]

[1] Student of Enology Master, UTAD, 5001 801 Vila Real, Portugal
[2] CQ-VR, Dep. of Biology and Environment, UTAD, 5001 801 Vila Real, Portugal
[3] CM-UTAD, Dep. of Mathematica, UTAD, 5001 801 Vila Real, Portugal

ABSTRACT

Vinho Verde is a table wine produced in the Demarcated Region of Vinho Verde in northwestern Portugal. It is distinguished by its freshness, lightness and its unique fruity aromas, low alcohol and notorious typicality.

The region of Vinho Verde contains a large number of grape vines indigenous varieties, these have an aroma, flavor and peculiar color, defining their origin. Among the most important are the Alvarinho, Arinto and Loureiro [1] .

For the study of the organoleptic characteristics of these three grape varieties, sensory analysis is critical [2]. Therefore the main goal of this work was to develop a methodology of sensory analysis to study the characteristics of Alvarinho, Arinto and Loureiro monovarietal wines and recruit, select, train and qualify an internal tasting panel of Aveleda S.A. Company in order to get possible playback results when carrying out sensory tests. For this purpose, 17 wines were evaluated (4 from Arinto variety, 6 from Alvarinho and 7 from Loureiro) and its sensory descriptor scored with an ordinal scale from 1 to 5.

To characterize the sensory profile of the wines, given the nature of the variables (ordinal variables) and the dimension of the sample, nonparametric techniques such as Categorical Principal Components Analysis (CATPCA) and non-parametric MANOVA were used [3].

At the end of this study we found that there are common descriptors to more than one grape-variety. The attributes common to all grape-varieties are: yellow, clean, floral aroma, citrus aroma, apple flavor, sweet, sour, bitter, persistence, body and sense of balance. In the case of aromas, the three varieties have in common floral aroma, citrus and apple aromas. Alvarinho and Loureiro share the aromas of tropical fruit and vegetable. There were also found in these last two varieties of the flavor vegetable notes and tropical fruits.

Keywords: Vinho Verde wine descriptors, sensory evaluation, CATPCA, nonparametric MANOVA.

REFERENCES

[1] Comissão de Viticultura da Região dos Vinhos Verdes (2002). Comissão de Viticultura da Região dos Vinhos Verdes, Departamento de Marketing, Porto.

[2] Stone, H. & Sidel, J. (2004). Sensory evaluation practices. 3rd Edition. Academic, Elsevier, New York.

[3] Hair, J.; Anderson, R.E.; Tatham, R.L.; Black, W. (2009). Multivariate Data Analysis. 687 p. Prentice Hall, New Jersey.

II Encontro Galaico-Português de Biometria
Santiago de Compostela, 30 de Junho, 1 e 2 de Julho de 2016

ESCOLHA DE TRATAMENTOS OTIMIZADOS NA CONSTRUÇÃO DE FATORIAIS FRACIONÁRIOS

Claudiney Nunes de Lima[1]

[1]Campus Alto Paraopeba, Universidade Federal de São João del-Rei, Caixa Postal: 131, CEP: 36420-000, Ouro Branco, MG, Brasil, e-mail: *pleudiney@yahoo.com.br*

RESUMO

Este trabalho trata da implementação de algoritmo de simples troca do tipo *exchange* (trocas). Foi empregado um critério de eficiência para a escolha de tratamentos de fatoriais fracionários. O delineamento encontrado foi comparado a uma alternativa usual da literatura. O delineamento D-ótimo foi mais eficiente que o delineamento composto central. O algoritmo resultante é flexível e recomenda-se que sejam utilizados por pesquisadores práticos para construir alternativas aos fatoriais fracionários comumente encontrados na literatura. Isso vale tanto para estágios iniciais da pesquisa (escolha de tratamentos) quanto para fases mais adiantadas de otimização usando modelos de superfície de resposta.

Palavras-chave: algoritmo de troca; delineamento ótimo; delineamento composto central; fatoriais fracionários.

1. INTRODUÇÃO

Limitações diversas, como disponibilidade de material experimental, de recursos financeiros, de laboratórios e de equipamentos disponíveis, dentre outros, podem exigir ensaios com uma quantidade reduzida de parcelas ou unidades experimentais. Essa e outras justificativas também podem levar à redução do número de tratamentos. O uso de apenas uma parte dos tratamentos de um fatorial completo traz a vantagem de economia de recursos e de tempo, mas apresenta a desvantagem de que efeitos principais e interações podem estar confundidos, isto é, não podem ser estimados separadamente.

Do ponto de vista estatístico, a escolha dos tratamentos pode tornar um delineamento mais eficiente do que outro. Essa eficiência pode ser mensurada por meio de alguns critérios de otimalidade que podem ser implementados em algoritmos de busca. Dessa forma, os planos experimentais resultantes dessa busca são designados como ótimos. O problema da escolha otimizada de tratamentos consiste em selecionar, de um conjunto N tratamentos candidatos, um subconjunto com apenas $v < N$ tratamentos a serem avaliados, de tal maneira que o delineamento resultante seja o melhor possível em algum sentido. Em geral, para modelos lineares fixos, isto se reduz a estudar as propriedades da matriz de informação de Fisher (Chaloner e Verdinelli, 1995). Vários algoritmos computacionais para a construção de delineamentos com v tratamentos estão disponíveis na literatura, mas uma das classes mais importantes nessa direção são os algoritmos de troca (EA, do inglês *exchange algorithm*), dentre os quais podem ser citados: EA, atribuído a Fedorov (1972); EA, devido a Wynn (1970); EA, chamado de DETMAX , elaborado por Mitchell (1974); EA, atribuído a Cook & Nachtsheim (1980); e, por fim, EA, atribuído a Atkinson & Donev (1989).

Neste sentido, os objetivos deste trabalho podem ser resumidos em:

- implementar algoritmos de simples troca para a escolha de tratamentos para a construção de um fatorial fracionário;

- discutir a obtenção de delineamentos ótimos em termos de eficiência estatística para ensaios com estrutura fatorial em que o interesse seja estimar os efeitos principais e interações de baixa ordem do modelo fatorial (modelo de superfície de resposta).

2. METODOLOGIA

A pressuposição básica deste trabalho é que, para a análise dos dados, serão adotados modelos polinomiais do tipo Gauss-Markov Normais, em que $\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, em que $\boldsymbol{y}$ é o vetor das observações (n×1); X é uma matriz cujas colunas são expansões dos níveis de cada fator para acomodar o polinômio a ser ajustado; $\boldsymbol{\beta}$ é o vetor de parâmetros do modelo e $\boldsymbol{\varepsilon}$ é o vetor dos erros aleatórios correspondentes às observações (n×1).
O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é $\hat{\boldsymbol{\beta}} = (X^{'}X)^{-1}X^{'}\boldsymbol{y}$ e sua matriz de covariâncias é dada por: $\text{var}(\hat{\boldsymbol{\beta}}) = (X^{'}X)^{-1}\sigma^2$. Na prática, σ^2 também deve ser estimado e uma estimativa é obtida pelo quadrado médio do resíduo da análise de variância dos resultados do experimento. Note-se que σ^2 não depende de X e, portanto, pode-se estudar apenas as propriedades da matriz $X^{'}X$. Dentre os critérios de otimalidade disponíveis na literatura para tal otimização, os de maior interesse neste trabalho são os critérios A e D.

2.1 Passos do algoritmo *exchange*

1) **Sorteio**
Definidos o número de fatores do ensaio fatorial para s níveis, o número de parâmetros do modelo e o número de unidades experimentais, o algoritmo, então, constrói a matriz base ξ_N, com todos os pontos candidatos e, então, sorteia um delineamento inicial qualquer ξ_1 contendo os tratamentos a serem alocados às unidades experimentais.
2) **Matriz do delineamento**
Definido o delineamento ξ_1, o algoritmo, então, encontra a sua respectiva matriz de delineamento X_1.
3) **Matriz de informação**
A matriz de informação $(X_1^{'}X_1)$ é encontrada a partir da matriz de delineamento e armazenada.
4) **Critério**
Calcula-se o determinante de $(X_1^{'}X_1)$.
5) **Troca**
Na matriz do delineamento X_1 de ξ_1 é efetivada a troca de uma de suas linhas com uma linha da matriz dos pontos candidatos ξ_N.
6) **Construção do delineamento** ξ_2
Após o intercâmbio, obtém-se um novo delineamento. A partir daí, segue-se o mesmo passo efetivado em ξ_1.
7) **Construção da matriz do delineamento** ξ_2

$$X_2$$

8) **Construção da matriz de informação para o delineamento** ξ_2

$$(X_2^{'}X_2)$$

9) **Decisão**
Compararam-se os valores obtidos pelo critério de otimização que foram encontrados nos delineamentos ξ_1 e ξ_2. Assim, a decisão é tomada comparando-se o critério obtido pelo ξ_1 versus ξ_2.
O delineamento escolhido será aquele que possuir maior determinante para a matriz de informação. Ao se maximizar o determinante, obtém-se um elipsóide de menor volume para os coeficientes do modelo.
Comparando-se os dois delineamentos por meio do determinante das respectivas matrizes de informação, o algoritmo encaminha seguinte decisão:

- Se det $(X_1^{'}X_1)$ $>$ det $(X_2^{'}X_2)$, o melhor delineamento é ξ_1. Assim, o algoritmo o mantém.

- Se det $(X_1'X_1) < \det(X_2'X_2)$, neste caso o delineamento ξ_2 é melhor do que ξ_1. Logo, é armazenado.

10) **Ciclo**

Encontrando-se o melhor delineamento, uma nova troca é efetuada, repetindo-se, assim, todo o processo.

11) **Fim**

O algoritmo repete todo o procedimento para um número pré-determinado de sorteios iniciais. Terminado o processo, o melhor delineamento encontrado pelo algoritmo é, então, apresentado e armazenado.

3. RESULTADOS

O delineamento otimizado (segundo o critério D), encontrado pelo algoritmo de busca, para 15 tratamentos escolhidos, entre os 27 tratamentos de um fatorial completo 3^3 com os fatores A, B e C, considerando o ajuste do modelo de segunda ordem, está apresentado na Tabela 1.

Delineamento D-ótimo			DCC com $\alpha = 1$		
Fatores			Fatores		
A	B	C	A	B	C
-1	-1	-1	-1	-1	-1
1	-1	-1	1	-1	-1
1	1	-1	-1	1	-1
-1	1	-1	1	1	-1
-1	-1	1	1	-1	1
-1	1	1	-1	-1	1
-1	0	0	-1	1	1
0	1	0	1	1	1
1	-1	0	0	0	0
0	-1	-1	-1	0	0
1	-1	1	1	0	0
1	0	-1	0	-1	0
0	0	1	0	1	0
1	0	1	0	0	-1
1	1	1	0	0	1

Tabela 1: Delineamento D-ótimo com 15 tratamentos selecionados para efeitos principais, quadráticos e interações de dois fatores e o delineamento composto central (DCC: $2^k + 2k + 1$) com ponto axial sendo $\alpha = 1$

Apesar de os dois delineamentos possuírem 11 tratamentos em comum, o delineamento D-ótimo foi mais eficiente, $EF = \dfrac{24192 \times 10^4}{18432 \times 10^4} = 1,3125$, ou seja, aproximadamente 31% mais eficiente do que o delineamento composto central. Além disso, a estrutura dos aliases dos dois delineamentos é um pouco complexa, devido ao fato de ambos possuírem confundimento parcial dos efeitos, sendo, assim, complicado estabelecer o conjunto de aliases de cada um.

Observando-se apenas as inversas das matrizes de informação (matriz de dispersão) dos delineamentos ótimo e composto central, respectivamente, tem-se:

$$\left(X_{\text{ótimo}}'X_{\text{ótimo}}\right)^{-1} = \frac{\sigma_e^2}{100} \times$$

	β_0	β_1	β_2	β_3	β_{11}	β_{22}	β_{33}	β_{12}	β_{13}	β_{23}
β_0	**67**	0	−1	−2	−28	−22	−28	2	0	−2
β_1		**9**	1	0	−1	2	−2	1	0	0
β_2			**9**	0	1	0	1	−1	0	1
β_3				**9**	0	3	0	0	−2	1
β_{11}					**45**	−3	−8	1	0	4
β_{22}						**36**	−3	1	−1	−1
β_{33}				*Sim*			**45**	−5	0	−2
β_{12}								**12**	0	0
β_{13}									**10**	0
β_{23}										**12**

em que "*Sim.*" indica que a matriz é simétrica.

$$\left(X'_{\mathrm{DCC}}X_{\mathrm{DCC}}\right)^{-1} = \frac{\sigma_e^2}{100}\times \begin{array}{c} \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_{11} \\ \beta_{22} \\ \beta_{33} \\ \beta_{12} \\ \beta_{13} \\ \beta_{23} \end{array} \begin{array}{c} \begin{array}{cccccccccc} \beta_0 & \beta_1 & \beta_2 & \beta_3 & \beta_{11} & \beta_{22} & \beta_{33} & \beta_{12} & \beta_{13} & \beta_{23} \end{array} \\ \begin{pmatrix} \mathbf{29} & 0 & 0 & 0 & -11 & -11 & -11 & 0 & 0 & 0 \\ & \mathbf{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & \mathbf{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & \mathbf{10} & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & \mathbf{39} & -11 & -11 & 0 & 0 & 0 \\ & & & & & \mathbf{39} & -11 & 0 & 0 & 0 \\ & & & Sim & & & \mathbf{39} & 0 & 0 & 0 \\ & & & & & & & \mathbf{13} & 0 & 0 \\ & & & & & & & & \mathbf{13} & 0 \\ & & & & & & & & & \mathbf{13} \end{pmatrix} \end{array}$$

Note-se que cada linha ou coluna representa os coeficientes do modelo de regressão, sendo que os elementos da matriz de dispersão são as covariâncias dos coeficientes, de forma que os elementos na diagonal principal representam as variâncias e os que se situam fora dela são as covariâncias. Apesar de a estrutura de covariância do delineamento ótimo apresentar mais covariâncias do que a do DCC, as variâncias dos seus coeficientes são, na maioria, inferiores. O fato de que a matriz de covariâncias do DCC está mais próxima da ortogonalidade do que do delineamento otimizado, não impede obter combinações lineares no delineamento otimizado mais precisas. Por exemplo: caso o pesquisador esteja interessado na combinação $\hat{\beta}_1 - \hat{\beta}_2$, as variâncias do seu estimador nos delineamentos otimizado e DCC são, respectivamente, iguais a ($\mathbf{16} = 9 + 9 - 2 \times 1$ e $\mathbf{20} = 10 + 10 - 2 \times 0$) $\times \dfrac{\sigma^2}{1000}$.

4. CONCLUSÃO

Para a escolha de tratamentos em fatoriais fracionários, em geral pode-se obter delineamentos otimizados mais eficientes do que os delineamentos já conhecidos usando algoritmos do tipo *exchange*. Isso se aplica tanto na fase inicial de estudos com muitos fatores quanto para delineamentos de superfície de resposta, como os fatoriais com 3 níveis.

5. REFERÊNCIAS

[1] ATKINSON, A. C.; DONEV, A. N. (1989). *The construction of exact D-optimum experimental designs with application to blocking response surface designs.* Biometrika. n.76, p.515-526.

[2] BOX, G. E. P.; DRAPER, N. R. (1987). *Empirical model building and response surfaces.* New York: J. Wiley, 669p.

[3] BOX, G. E. P.; HUNTER, J. S.; HUNTER W. G. (2005) *Statistics for experimenters: design, innovation, and discovery.* 2. ed. New Jersey: John Wiley & Sons, 639p.

[4] CHALONER, K.; VERDINELLI, I. (1995). *Bayesian experimental design: a review.* Stat. Sci., v.10, n.3, p.237-304.

[5] COOK, R. D.; NACHTSHEIM, C. J. (1980) *A comparison of algorithms for constructing exact D-optimal designs.* Techonometrics. v.22, p.315-324.

[6] FEDOROV, V. V. (1972). *Theory of optimal experiments.* New York: Academic Press.

[7] MITCHELL, T. J.(1974). *An algorithm for the construction of D-optimal designs.* Technometrics. v.20, p.203-210.

[8] R DEVELOPMENT CORE TEAM. *R:* A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2009. Disponível em: <http://www.R-project.org>. Acesso em: 2009.

[9] WYNN, H. P. (1970) *The sequential generation of D-optimal experimental designs.* Ann. Math. Statistical. v.41, n.5, p.1655-1664.

II Galician-Portuguese Meeting of Biometry,
Santiago de Compostela, 30 June – 2 July 2016

SAMPLE SIZE REQUIRED FOR ESTIMATING THE MEAN PARASITE ABUNDANCE

Svitlana Shvydka[1], Volodimir Sarabeev[2], Vicente D. Estruch[3], Carmen Cadarso-Suárez[4]

[1] Department of Mathematics, Zaporizhzhia National University, Ukraine
[2] Department of Biology, Zaporizhzhia National University, Ukraine
[3] Institut d'Investigació per a la Gestió Integrada de Zones Costaneres, Departament de Matemàtica Aplicada, Universitat Politècnica de València, Spain
[4] Biostatistics Group, Center for Research in Molecular Medicine and Chronic Diseases (CIMUS), University of Santiago de Compostela, Spain

ABSTRACT

The present study considers two different approaches, analytic and stochastic, to estimate an adequate sample size to be able to arrive at ethically and scientifically valid mean parasite abundance values. Assuming the negative binomial distribution (NBD) to modelling overdispersed parasite data, the sample size calculation was carried out applying mathematic formula with determined precision levels (D=0.5 and D=0.8) and the estimated dispersion parameter of the NBD. Bootstrap techniques were also applied to appraise the accuracy of sample size formula under a variety of precisions. Two species of monogenean parasites *Ligophorus cephali* and *L. mediterraneus* from *Mugil cephalus* across the Azov-Black Seas localities were subjected to the analysis. The optimum sample size for the studied species of *Ligophorus* was defined as 30-40 fish individuals that approximately correspond to the precision level of 0.8.

Keywords and key sentences: Fish, parasite, mean abundance, optimum sample size, precision, negative binomial distribution, bootstrap.

1. INTRODUCTION

The aim of this work, is to estimate an optimum sample size for mean parasite abundance values. Using the NBD distribution, the optimum sample size and the confidence interval (*CI*) for the mean abundance were both obtained based on a formula using predetermined precision levels for each of the species examined. Bootstrap techniques were also applied to appraise the accuracy of sample size formula under a variety of precisions. Simulation studies were conducted to assess the actual mean abundance and *CI*s for the suggested sample sizes described in the first stage. The accuracy of the stochastic approach was evaluated through the mean estimates, biases, and *CI*s corresponding to the control of their expected widths.

2. MATERIALS AND METHODS

This study is based on 224 dissected individuals of flathead mullet *Mugil cephalus* L. that includes 8 samples for 205 fishes from three localities, the Kerch Strait, the Sivash Lake and the Balaklava Bay, in the Azov-Black Seas sampled in 2001-2013 [1] and one extra sample with 19 fish individuals from the Sivash Lake executed in 2014. The number of fish typically reached 30 specimens per sample. Only two-years old fishes and older within the size range of 24-65 cm (total length) were used in the analyses. A total of 9 samples across all localities, years and seasons are studied here. For each parasite species, only

samples with more than 6 infected hosts were considered, to avoid inadequate estimation of mean abundance due to very low prevalence [2].

Collected fishes were measured and surveyed for parasites within the day of capture or after freezing. Gills were carefully examined under a stereomicroscope for parasites. All monogeneans parasites were identified and counted. Taxonomic identification was attempted to the lowest possible level. The present study considers two species of *Ligophorus* from *M. cephalus* around the Azov-Black Sea localities, *L. cephali* Rubtsova, Balbuena, Sarabeev, Blasco-Costa & Euzet, 2006 and *L. mediterraneus* Sarabeev, Balbuena & Euzet, 2005.

The general formula for determining optimum sampling size (n) have been presented by the equation:

$$n = \left(\frac{Z_{\alpha/2}}{D} \right)^2 \frac{s^2}{m^2}, \quad (1)$$

in which $Z_{\alpha/2}$ is the standard normal deviate such that $P\left(Z > Z_{\alpha/2}\right) = \alpha/2$; D is a level of precision, used to define half-width of the *CI* as a fixed proportion of the mean ($CI/2 = Dm$); s^2 is the sample variance; and m is the sample mean [3]. For a 95% *CI*, $\alpha = 0{,}05$, then $Z_{\alpha/2}$ equals approximately 2. With this approach to calculating optimal sample size the mean is estimated with a *CI* [4]:

$$(m - D \cdot m; m + D \cdot m). \quad (2)$$

In the present study the optimum sample size was determined for two different precision levels: *D*=0.5 and *D*=0.8. These levels are reasonable for practical applications and are acceptable for most sampling research [5, 6].

For modelling overdispersed parasite data, the negative binomial distribution (NBD) is more appropriate than Poisson model [7]. The NBD is characterized by the dispersion parameter *k* and the mean *m*. The relationship between the mean and variance is $s^2 = m + m^2/k$. In this case, the following formula can be used to determine optimum sample sizes [4]:

$$n = \left(\frac{Z_{\alpha/2}}{D} \right)^2 \left(\frac{1}{m} + \frac{1}{k} \right). \quad (3)$$

For each parasite species negative binomial parameter *k* was estimated by using the maximum likelihood method [8]. The **histDist()** function in the **gamlss** package for the **R** statistical data analysis software (version 3.2.3) and Akaike information criterion (AIC) were used to test goodness-of-fit of the NBD for empirical data.

In order to obtain an estimation of the mean abundance, as well as its *CI* and bias, based on samples of different size, the Monte Carlo simulation with bootstrap procedures were applied. By using non-parametric bootstrapping, *n* elements of each host-parasite dataset were randomly selected 1000 times from original dataset and the programmed algorithm perform a bootstrap with 5000 iterations for each time, based on samples with number of elements previously calculated by formula (3). Bias significance was evaluated through *t*-test. The same bootstrap technique was applied to samples with 10, 20, 30, 40, 50, 60 and 70 elements. The differences between the estimates of the mean abundance obtained based on different sample sizes were examined using analysis of variance (ANOVA) followed by Tukey's pairwise comparison test to identify significant differences after a logarithmic transformation of the data. A significance level of 0.05 was considered in all test procedures.

3. RESULTS AND DISCUSSION

The dispersion pattern of both helminth species could be characterized as a highly aggregated distribution with variance that is substantially larger than mean. Obtained values of *k* were lower than 1 also indicating on a highly aggregated distribution of these species in the host (Table 1).

The results of calculation the minimum sample size needed to reach the predetermined precision are presented in Table 2. In all cases, the sample size increased when higher precision level of *D* was selected and with increasing aggregation. The mean abundance values obtained by the Monte Carlo simulation based on bootstrapping technique were close to empirical values of the mean and no significant biases were found in the estimates. The median values of the mean were always under-estimate the empirical value for low sample size (Figure 1). The 95% *CI*s of the mean abundance values

constrained with fixed precision *D* were compared with that obtained for the bootstrap means (Table 2). The width of the *CI* decreases as the sample size increase.

species	mean	variance	*k*	AIC
L. cephali	15.65	1908.92	0.27	1258.54
L. mediterraneus	5.55	221.25	0.23	882.33

Table 1: Summary statistics for two parasite species (*k*: dispersion parameter of the NBD, AIC: Akaike information criterion).

species	*n* (*D*=0.5)/ *n* (*D*=0.8)	Bootstrap mean	Bias	Bootstrap 95% *CI*	95% *CI* based on (2) *D*=0.5/ *D*=0.8
L. cephali	60/ 24	15.56/ 15.70	-0.09/ 0.05	7.4–26.0/ 4.64–34.27	7.83–23.48/ 3.13-28.17
L. mediterraneus	73/ 29	5.53/ 5.65	-0.02/ 0,1	3.02–8.17/ 1.83–11.81	2.78–8.33/ 1.11–9.99

Table 2: Optimal sample sizes (*n*) estimated for two precision level (*D*=0.5 and *D*=0.8) using exponent *k* of the NBD (3) supplied with bootstrap simulated mean, bias and 95% *CI* and compared with 95% *CI* based on (2). Percentile 95% *CI* for bootstrap was defined using the values that mark the upper and lower 2.5% of the bootstrap distribution.

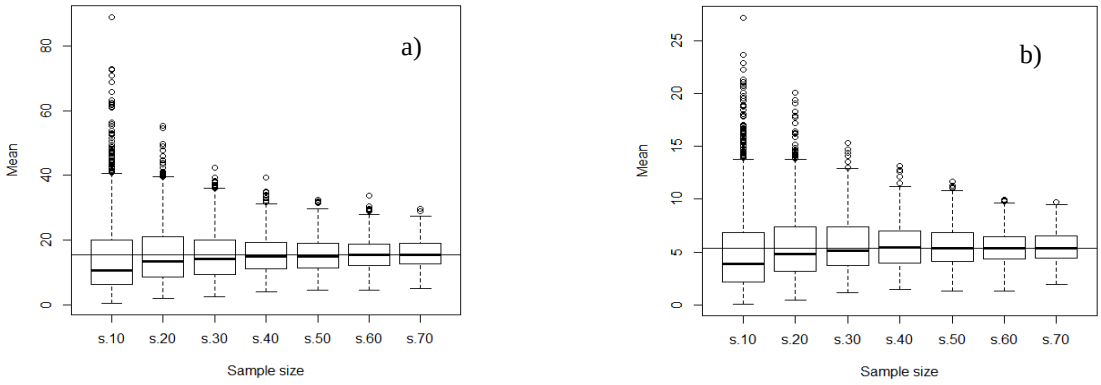


Figure 1: Distribution of mean abundance estimates for all host-parasite pairs, obtained by simulating different sample sizes (solid line – empirical mean abundance). *L. cephali* (a) and *L. mediterraneus* (b).

Pairwise statistical comparison between mean abundance values across sample sizes using the Tukey confidence intervals is represented in Figure 2. This test showed that no statistical difference was found between means based on samples beginning with 40 elements for *L. cephali* and 30 elements for *L. mediterraneus*.

If the purpose of sampling procedure is to get a general idea about parasite abundance in fishes, sample size required for the mean abundance estimation could be reduced by lowering the level of precision to *D*=0.8. If this level of precision is used, the samples with 24-29 individuals could be recommended for field survey of parasites. The similar result was obtained from the simulation. Based on bootstrap procedures the present study shows that low sample sizes with less than 30-40 individuals lead to under-estimated the empirical value of mean abundance with a highly skewed to the right distribution of the mean estimates. This observation together with the results of the Tukey's pairwise comparison test proves that the more rationale sampling decision is 30-40 fish individuals and coincides with computations of Marques and Cabral [9]. Thus, the level of precision *D*=0.8 can be selected as reasonable and acceptable precision for optimal sample size calculation in case of parasite populations with a highly aggregated distribution in a host.

Based on this study, we would stress the importance of both analytical and stochastic approaches for sample size calculation. Formula allows to estimate the sample size with respect to variations in the precision level, the mean and the parameter *k* of NBD. Applying the formula with parameter *k* of NBD

requires to estimate *k* accurately. However, in order to assure the stability of the *k* parameter and increase the confidence in its suitability for sample size determination, the estimates of *k* should be calculated based on large empirical dataset because the parameter *k* depends from empirical sample size [7].

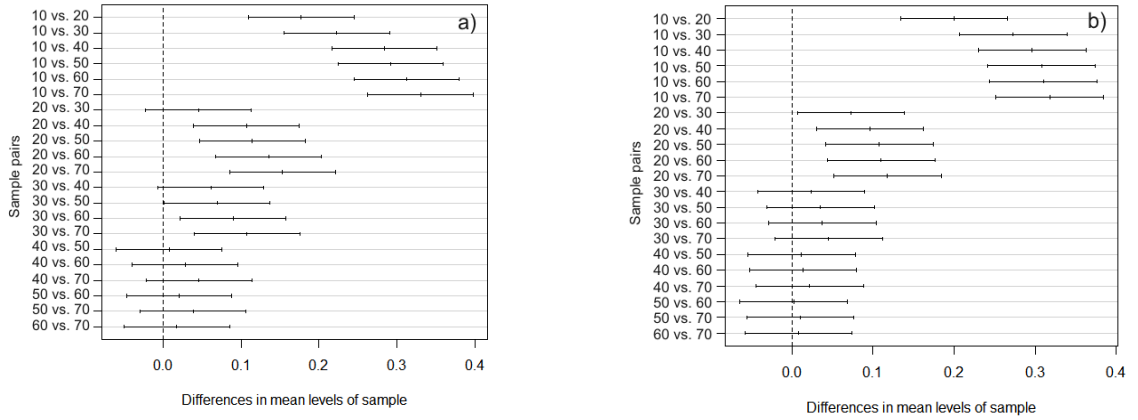


Figure 2: The results of the Tukey's pairwise comparison test for the data of *L. cephali* (a) and *L. mediterraneus* (b) across different sample sizes. All possible pairs of samples are shown along the y-axis.

Simulation-based approach also gives reasonable sample size estimation. The primary advantage of bootstrapping is that no assumptions are made about the distribution of initial dataset. The accuracy of estimates obtained by bootstrapping depends on the number of observations in the original sample and the number of resamples. Large samples are likely to be more representative than small samples.

ACKNOWLEDGMENTS

This study was supported by the project MTM2014-52975-C2-1-R:"Inference in Structured Additive Regression (STAR) Models with Extensions to Multivariate Responses. Applications in Biomedicine" (Cofinanced by the Ministry of Economy and Competitiveness, SPAIN; and the European Regional Development Fund, FEDER), and by the MEDEA Program, Erasmus Mundus Action II.

REFERENCES

[1] Sarabeev, V.L. (2015). Helminth species richness of introduced and native grey mullets (Teleostei: Mugilidae). *Parasitology International* 64, 6-17.

[2] Poulin, R. (2013). Explaining variability in parasite aggregation levels among host samples. *Parasitology* 140, 541-546.

[3] Karandinos, M.G. (1976). Optimum sample size and comments on some published formulae. *Bull. Entomol. Soc. Am.* 22, 417-421.

[4] Pedigo, L.P., Buntin, G.D. (1994). *Handbook of Sampling Methods for Arthropods in Agriculture*. CRC Press, Boca Raton.

[5] Mouillot, D., Culioli, J.-M., Lepretre, A., Tomasini, J.-A. (1999). Dispersion statistics and sample size estimates for three fish species (*Symphodus ocellatus, Serranus scriba* and *Diplodus annularis*) in the Lavezzi Islands Marine Reserve (South Corsica, Mediterranean Sea). *Marine Ecology* 20(1), 19-34.

[6] Opit, G.P., Throne, J.E., Flinn, P.W. (2009). Sampling plan for the Psocids *Liposcelis entomophila* and *Liposcelis decolour* (Psocoptera: Liposcelididae) in Steel Bins Containing Wheat. *J. Econ. Entomol.* 102(4), 1714-1722.

[7] Wilson, K., Bjørnstad, O.N., Dobson, A.P., Merler, S., Poglayen, G., Randolph, S.E., Read, A.F., Skorping, A. (2002). Heterogeneities in macroparasite infections: patterns and processes. In: Hudson, P.J., Rizzoli, A., Grenfell, B.T., Heesterbeek, H., Dobson, A.P. (eds.): *The Ecology of Wildlife Diseases.* Oxford University Press, pp. 6-44.

[8] Young, L.J., Young, J.H. (1998). *Statistical ecology: a population perspective*. Kluwer Academic Publishers, Norwell, MA.

[9] Marques, J.F., Cabral, H.N. (2007). Effects of sample size on fish parasite prevalence, mean abundance and mean intensity estimates. *J. Appl. Ichthyol*. 23, 158-162.

II Galician-Portuguese Meeting of Biometry
Santiago de Compostela, 30 June – 2 July 2016

SPATIAL MODELLING OF THE ABUNDANCE OF *Culex pipiens* IN PORTUGAL - A BAYESIAN APPROACH

P. de Zea Bermudez[1(a)], Marília Antunes[1(a)], Maria da Conceição Proença[2(b)], Maria Teresa Rebelo[3(b)], Hugo Osório[4(b)] and Maria João Alves[4(b)]

[1]CEAUL and DEIO-FCUL, Universidade de Lisboa
[2]LOLS and DF-FCUL, Universidade de Lisboa
[3]CESAM/FCUL and DBA-FCUL, Universidade de Lisboa
[4]Centro de Estudos de Vectores e Doenças Infecciosas and ARSs

ABSTRACT

Arthropods are the transmission vectors for the arbovirus, of which the *Culex pipiens* is the most common species. In order to assess which species are most commonly observed in Portugal and their distribution throughout the country, the National Health Institute, along with the regional and national health authorities, developed a vectors surveillance program. Mosquitoes are captured using traps, and ecological and meteorological variables, as well as geographical coordinates are recorded at the site of capture. The geographical coordinates allow to obtain georeferenced information on variables that may affect the presence and abundance of mosquitoes, such as the proximity of water masses. The data were recorded between May and October, 2006-2012. The observations are neither spatially nor temporally regularly distributed and do not cover the whole country.

One of the features of the data is the large amount of zero counts and also the over-dispersion, which motivates the use of zero inflated models. The abundance of mosquitoes seems to be spatially dependent and varies according to some covariates, such as altitude and the type of still water surfaces that exist in the vicinity of some of the locations. As such, the expected number of mosquitoes at a site is modelled as function of several covariates that describe the geographical and meteorological characteristics, plus a latent random field to account for spatial dependence. The traditional use of GFs is generally inadequate from a practical point of view because the covariance matrices are dense, especially when the number of locations is high. The fitting of the models becomes computationally feasible if Gaussian Markov Random Fields (GRMFs) are used, due to the assumption of conditional independence which enables the precision matrices to be sparse (Rue and Held, 2005). The approximation of the GF through a GMRF exists for certain GFs with Matérn covariance function, and can be derived by means of a Stochastic Partial Differential Equation (SPDE)(Lindgren *et al.*, 2011). This approach is implemented in INLA ($http://www.r-inla.org/$) in a Bayesian framework. Several Bayesian models are fitted to the data using INLA. The goodness of fit of the models, as well as their predictive performance, is analyzed. Abundance maps are presented for several meteorological and temporal scenarios.

Keywords and key sentences: Gaussian Fields, Gaussian Markov Random Fields, Stochastic Partial Differential Equation, Bayesian Models, Zero Inflated Models, INLA.

ACKNOWLEDGMENT

(a) - Funded by FCT - Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013
(b) project PTDC/SAU-SAP/119199/2010
The authors thank the REVIVE team for the field and the lab works.

References

[1] Lindgren F., Rue H., Lindström J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Royal Statistical Society (B)* 73, 423–498.

[2] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields - Theory and Applications.* Chapman and Hall/CRC, Boca Raton, Florida.

Autores

AUTORES:

Afreixo, V. 194
AL Bautista, E. 239
Almendra, R. 243
Alpizar-Jara, R. 165
Altair Bertonha, A. 160
Álvarez Escudero, J. 310
Álvarez, R. 234
Álvarez, X. 58
Álvarez-Salgado, X.A. 216
Alves, C. 245, 338
Alves, M.J. 3615
Amaral Turkman, A. 153
Amaral, M.T.R. 302
Ambrizzi, T. 243
Ambroa-Rey, E. 306
Ameijeiras-Alonso, J. 149
Ana Borges, A. 295
André, C.D.S. 80
Andrés Baselga, A. 215
Antônio C, ...302
Antônio Neco, A. 290
Antunes, M. 153, 350, 361
Aparecida Trinca, L. 96
Araújo, J.A. 178
Argentina Leite, A. 182
Armesto, J. 58
Arostegui, I. 142
Balboa Barreiro, V. 288
Baluja, A. 310
Barreiro-Martínez, T. 123
Basanta, M. 234
Baselga, A. 164, 214
Basile, L. 53
Baturin, O. 269
Benitez-Estévez, A.J. 123
Blanco Vázquez, M. 112
Blanco Vilariño, A. 238
Blanco, F.J. 296
Borges, A. 229
Boubeta, M. 62
Braga, A.C. 298
Braumann, C.A. 23, 120
Brilhante, M.F. 342, 346
Cabaleiro Casal, M.J. 326
Cacheiro, P. 137
Cadarso-Suárez, C. ... 41, 101, 123, 127, 131, 137, 189, 310, 316, 330, 357
Calvo Ocampo, M.E. 201, 205
Cameletti, M. 27
Cao, R. 49
Capelo, J.L. 178
Carracedo, A. 137
Carvalho, S. 181
Casal, M. 234
Casas-Méndez, B. 123
Castro, A. 286
Castro, L. 229
Castro-Insua, A. 215
Cermeño, P. 112
Cetra, M. 210
Cicogna 302
Cicogna, A.C. 37
Cocchi, D. 21
Conesa, D. 22, 30
Correia, E. 351
Cortés Pereira, E. 296
Costa, M. 63
Cotos-Yáñez, T.R. 265
Cousido Rocha, M. 147
Crujeiras, R.M. 67, 149, 249
D Estruch, V. 357
D Hart, J. 147
Dal Molin Ribeiro, M.H. 92
de Sousa, B. 127, 131
de Uña Álvarez, J. 147
de Zea Bermúdez, P. 361
Dias, F. 245
DiMarzio, N. 109
Diz, P. 330
DR Frade, D. 239
Drakeley, C. 167, 294
Duarte, E. 127, 131
Escobar, A. 75
Escudeiro, S. 194
Espasandín-Domínguez, J. 123, 189, 316
Estévez, G. 275
Etxeberria, J. 173
Evangelista 302
F de Andrade, D. 146
Faes, D. 41
Febrero Bande, M. 53
Feijoo, G. 67
Fernandes, J. 350
Fernandez de Carvalho, W. 88
Fernández López, C. 296
Fernández Moreno, M. 296
Fernández, M.A. 45
Fernández-Reiriz, M.J. 216
Ferreiro Velasco, ME. 288

Flores, M. 75
Flournoy, N. 186
Francisco-Fernández, M. 279
Francos Gendra, R. 326
Freijeiro, A. 164
Freitas, A. 194
Fuentes-Santos, I. 116, 216
G Figueiras, F. 112
G Ramires, T. 239
García Soidán, JL. 133
García Varela, S. 319
García-Duro, J. 234
García-Sánchez, M. 312
Giménez-Insua, M. 312 312
Ginzo-Villamayor, M.J. ..253, 306, 312, 319, 326
Goicoa, T. 173
Gomes, M.I. 346
Gómez Duaso, J. 201
Gómez, E. 350
Gómez-Rodríguez, C. 164, 215
Gonçalves, A. 178
Gonçalves, A.M. 269
Gonçalves, F. 243
Gonçalves, L. 30
González Andújar, J.L. 28
González-Manteiga, W. 116
Gónzalez-Patiño, E. 306
Gouveia, S. 33
Gouveia-Reis, D. 257
Gude, F. 123, 316
Guimarães Araújo, M. 84
Guiu-Souto, J. 312
Guler, G. 41
Horna, L. 75
Iglesias Patiño, C.L. 201, 205
Iglesias Pérez, M.C. 133
Iglesias, A. 310
Iglesias-Allones, J.L. 253
Igrejas, G 178
Inês Sousa, I. 295
Ipek Guler, I. 225
J Righetto, A. 239
Jácome, M.A. 49
Jansen-Kosterink, S. 350
Jorge, P.A. 109
Kimura, E. 221
Klein, N. 101, 131, 316
Kneib, T. 20, 101, 127, 131, 316
Kuroda,C. 221
Labarta, U. 216
Lado Baleato, O. 238
Laetitia Teixeira, L. 225
Larriba, Y. 45
Lázaro, J. 142
Lema-Suárez, I. 275
Liliana Coelho, L. 295
Lima, M. 245
Limeres, J. 330
Lombardía Cortiña, M.J. 62
López Ratón, M. 310
López Vizcaíno, M.E. 201
López Vizcaíno, M.E. 205
López-Abente, G. 173
López-Cheda, A. 49
Lorenzo-Toja, Y. 67
Lozano, D. 330
Luís Castro, L. 295
Luna-Vega, V. 312
M Brites, N.M. 120
Magalhães Moreira, T.M. 146
Malvar Pintos, A. 169, 282
Marco Costa, ...269
Marco Lima, M. 338
Marey-Perez, M.F. 253
Marey-Pérez, M.F. 62, 116
Margalho, L. 261
Maria C. 302
Marinho, C. 178
Marín-Navarro, D. 67
Mariño, X. 30
Marques, C. 351
Marques, T.A. 109
Martín Biedma, B. 319
Martín Vila, M. 201, 205
Martínez Villanueva, N. 273
Martínez, A. 53
Martínez, I. 330
Martín-Rodríguez, Z. 312
Martín-Rodríguez, Z. 306
Matabuena Rodríguez, M. 155
Matarazzo, F. 84
Mateo-Abad, M. 142
Mateu, J. 116
Meijide, H. 286
Meilán-Vila, A. 249
Meira Machado, L.F. 18
Mena de Cea, A. 286
Mendonça, D. 41, 225
Menezes, R. 105, 261
Milani, H. 92
Molenberghs, G. 22
Moler, J. 186
Monteiro, A. 105

Monteiro, M. 63, 71
Montoto Marqués, A. 288
Morales, D. 62
Moreira, M.T. 67
Moretti, D.J. 109
Mosquera-Seoane, J. 306
Mougán-Bouzón, H. 326
Mourão, M.F. 298
Mouriño, H. 109
Natário, I. 198
Naveira Barbeito, G. 169, 282
Naya, S. 75, 323
Neves Pereira, C.O. 88
Neves, M.M. 257
Nunes de Lima, C. 353
Oliveira, I. 178
Ordóñez, C. 265, 273
Oreiro Villar, N. 296
Ortega, E.M.M. 334
Osório, H. 361
Otero Castro, P. 310
Otero-Martínez, C. 312
Oviedo de la Fuente, M. 53
Padovani, C.R. 37
Paredes Garcia, E. 96
Paulino, C.D. 294
Pazos Pata, M. 101
Peddada, S.D. 45
Pereira Borges, J.W. 146
Pereira, I. 71
Pereira, J.C. 210
Pereira, J.M. 30
Pereira, O. 84, 92, 160, 221
Pereira, P. 175
Peres da Silva Suguiura, P. 88
Peres, M. 221
Pérez Ríos, M. 169
Pérez-Álvarez, R. 265
Pértega Díaz, S. 286, 288
Pértega Díaz, S. 296 9
Pestana, D. 342, 346
Picos, J. 58
Pimentel, M. 275
Pinto, C. 33
Plo, F. 186
Poeta, P. 178
Pombar-Cameán, M. 306, 312
Poveda, E. 286
Prata Gomes, D. 257
Previdelli, I. 84, 88, 92, 160, 221
Proença, M.C. 361
Pueyo, S. 112
Puig, P. 26
Quintela-del-Río, A. 279
R Nakamura, L. 239
R. Padovani, C. 302
Ramires, T.G. 334
Ramos, S. 153
Raquel Menezes, R. 290
Rauter, A. 350
Rebelo, M.T. 361
Rego Pérez, I. 296
Relaño, S. 296
Rey Louzán, S. 326
Reyes, O. 234
Reyes-Santías, F. 189
Ribeiro Junior, P.J. 239
Ribeiro, C. 175
Ridout, M. 210
Riebler, A. 173
Rivas Lombardero, P.M. 319
Robles-Bykbaev, Y. 323
Roca, J. 330
Roca-Pardiñas, J. 137, 265, 273
Rocha, A.P. 182
Rocha, D. 33
Rocha, F.M.M. 80
Rodrigues Cassy, S. 198
Rodrigues, A. 41, 225
Rodrigues, V. 127, 131
Rodríguez Ferreiros, R. 326
Rodríguez Osorio, I. 286
Rodríguez Sotillo, A. 288
Rodríguez-Casal, A. 149
Rueda, C. 45
Sáez, R. 75
Sahuquillo, E. 275
Salvador de la Barrera, S. 288
Sanjurjo-Sánchez, J. 245, 338
Santana, P. 243
Santiago Pérez, M.I. 205, 282
Santos, H.M. 178
Santos, T. 178
Sarabeev, V. 357
Scotto, M. 33
Scotto, M.G. 71
Semblano, P. 342
Seoane Pillado, M.T. 286, 288
Sepúlveda, N. 167, 294
Sequeira, F. 342
Sestelo, M. 273
Shaffer, J.W. 109
Shvydka, S. 357
Silva Alonso, T. 133

Silva, C. 181
Silva, G.L. 243
Silva, M.E. 105, 182
Silveira Calviño, S. 201
Singer, J.M. 80
Solana, J. 350
Soto Hermida, A. 296
Sousa, I. 225, 229, 261
Sousa, L. 181
Stresman, G. 167
Suarez, C. 225
Susana Faria, S. 290
T R Amaral, M.T.R. 37
Taboada-Torres, I. 312
Tarrío-Saavedra, J. 75, 323
Tavares, J.N. 33
Teixeira, I. 112
Teixeira, L. 41
Thaden, H. 101
Thomas Kneib, T. 123
Thomas, L. 109
Tomás Ramos, J.C. 319
Tormena, M. 84
Torner, N. 53
Turkman, A. 350
Turkman, F. 350
Ugarte, M.D. 173
Urkaregi, A. 142
Vasconcelos, J. 243
Vázquez Mosquera, M.E. 296
Vázquez-Rowe, I. 67
Vila López, D. 326
Vilar Fernández, J.A. 30
Vileigas, D.F. 37
Vilela, A. 351
Vollenbroek, M. 350
White, M. 167
Zerbini, T. 80

COLABORADORES

//ABANCA

agaca
Asociación Galega de Cooperativas Agroalimentarias

BALIDEA

BIOstatnet

CESGA
Centro de Supercomputación de Galicia

CiMUS
Center for Research in Molecular Medicine and Chronic Diseases

CONCELLO DE SANTIAGO

Deputación DA CORUÑA

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Dirección Xeral de Saúde Pública

Estrella Galicia

USC
FACULTADE DE MATEMÁTICAS

IBERIA

IGE
INSTITUTO GALEGO DE ESTATÍSTICA

INE
Instituto Nacional de Estadística

modestya

gmodes

Pesqueras de Bon, S.A.

renfe

SIDOR
Universidad de Vigo

UHLIG
Solutions for tubes | fittings | boilers

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cátedra Camoens UC3M