Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *outono de 2019*

Estatística nas Ciências da Saúde

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe@spestatistica.pt*
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 400 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

www.spestatistica.pt

Editorial

… no dealbar de nova década SPE…

1. Fiz uma breve retrospetiva sobre as edições do Boletim SPE que se dedicaram a *Estatística e a Medicina*.
Basicamente, temos o Boletim outono de 2012 sobre *Métodos Estatísticos em Medicina*. Antes, houve entradas pela *Bioestatística* em outono de 2007 e em *Análise de Sobrevivência* em 2011. Mais recente, foi a edição que abordou a *Genética*.
A tudo isto acresce a motivação da proposta científica que se realizou em maio na Universidade de Aveiro, conforme noticiamos nesta edição. Uma excelente iniciativa a fazer realçar e a fortalecer a ligação entre a investigação e a sociedade. A oportunidade estava criada!
Nesse contexto e nos objetivos do Boletim SPE "era a hora" para (re)visitar aquele tema central e criar uma nova edição. Foi o que fiz. Consultei especialistas e o seu apoio e dedicação fez nascer o presente Boletim SPE sobre *Estatística nas Ciências da Saúde*.
A todos os co-editores e autores do tema central desta edição do Boletim outono de 2019, é devido um agradecimento pela colaboração prestada, desde logo na escolha do título, bastante (mais) abrangente do que os anteriores e também pela generosidade para, em cada uma das respetivas especialidades, partilharem as suas mais recentes reflexões científicas. Graças a eles fazemos um novo ponto da situação desta especialidade na Ciência Estatística.

2. Muito perto da data em que esta edição vai chegar aos seus leitores, temos o dealbar de uma nova década da Sociedade Portuguesa de Estatística. De facto, foi em 28 de Novembro de 1980 que foi fundada a SPE — Sociedade Portuguesa de Estatística (nome adotado desde 1991, pois originalmente era Sociedade Portuguesa de Estatística e Investigação Operacional) tendo a escritura da sua formação 11 outorgantes. O relato geral dessa vertente pode ser seguido na edição *Memorial da SPE* (2005).
Uma nova década que, a todos os títulos e para o maior sucesso, devemos saber construir como uma década nova onde todos os novos desafios e caminhos científicos sejam, ao máximo, percorridos e aprofundados. Preparemos e vivamos esses novos tempos com a intenção primordial do maior sucesso para a SPE, isto é, para a Estatística em Portugal. Para já e como acontecimento de partida tivemos a realização do XXIV Congresso de que damos o devido realce e onde com muita força "se sentiu juventude". E o próximo ano "já está bem preenchido" como se pode avaliar pelas iniciativas divulgadas nesta edição e que vão culminar com a realização do XXV Congresso SPE.
3. Além disso, no contexto da época aniversaria já referida e, muito em especial, pela nova valência que lhe é acrescentada, esta edição do Boletim SPE simbolicamente é um dealbar.
Este *Boletim outono de 2019* tem o privilégio do início de uma colaboração periódica e mais regular por parte do Instituto Nacional de Estatística – INE.
O INE e a SPE desde sempre que mantêm uma ligação profunda. A SPE muito deve ao apoio recebido do INE nas mais diversas realizações científicas e editoriais. Nesta época e como um passo no sentido de uma maior ligação entre a SPE e a Comunidade a colaboração agora iniciada, a todos os títulos, apenas virá reforçar essa ligação existente e apresentar uma consolidação na divulgação da Ciência Estatística. Para iniciar, damos relevo à notícia na respetiva secção e um breve relato de atividades em *SPE e a Comunidade*. São devidos agradecimentos ao INE por (mais) esta oportunidade criada.
O INE – Instituto Nacional de Estatística inicia em breve uma época de aniversário. É uma razão suficiente para que essa instituição de referência, nacional e internacional, seja eleita como o tema fulcral da próxima edição do Boletim, bem como toda a sua atividade e os seus colaboradores.

Fernando Rosa

O Tema Central do próximo Boletim SPE será *INE - 85 anos de estatísticas a servir o país*.

Mensagem da Presidente

Caros sócios da SPE,

Escrevo este texto após o XXIV Congresso da Sociedade pelo que, naturalmente começo por fazer uma breve referência a este evento que é sempre marcante para a nossa Sociedade. A realização deste Congresso foi fruto de um trabalho extenso e generoso de muitos e por isso quero, em nome da atual Direção, exprimir publicamente a nossa gratidão à Comissão Organizadora pelos esforços que desenvolveu para levar a cabo esta iniciativa, com o sucesso que pudemos testemunhar. Um agradecimento especial é devido também à Comissão Científica. O número de participantes no Congresso foi muito elevado e agradeço a todos o contributo não apenas para o sucesso deste congresso como para a vitalidade da Estatística em Portugal. Neste boletim podem os sócios encontrar mais detalhes relativos ao Congresso.

As atividades da SPE em 2019 não se limitaram à realização do XXIV Congresso. Entre 15 e 17 de Julho realizou-se na Faculdade de Ciências da Universidade de Lisboa, o curso intitulado *ECAS2019 - Statistical Analysis for Space-Time Data*, organizado pela SPE e pela SEIO (*Spanish Society of Statistics and Operational Research*), com o apoio do CEAUL (Centro de Estatística e Aplicações da Universidade de Lisboa), do CIM (Centro Internacional de Matemática), da FCT (Fundação para a Ciência e Tecnologia) e da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, no âmbito do projeto PREFERENTIAL: PTDC/MAT-STA/28243/2017. O Curso contou com 56 participantes, de 14 países, sendo metade dos participantes estudantes. Também neste caso a participação excedeu as expectativas e eu quero agradecer à Comissão Organizadora e à Comissão Científica todo o trabalho que culminou num evento excelente. Um relato detalhado encontra-se nas páginas deste boletim.

Organizámos vários prémios: os Prémios Estatístico Júnior 2019 - este ano patrocinados pelo CMUC (Centro de Matemática da Universidade de Coimbra); o Prémio SPE 2019, que foi atribuído a Jessica Lomba com o trabalho intitulado ***L-moments for automatic threshold selection in extreme value analysis of wave heights from the Gulf of Mexico.*** Atribuímos bolsas de participação no XXIV Congresso SPE a estudantes de Mestrado e Doutoramento. Ainda no âmbito do Congresso, reforçámos os laços com diversas sociedades congéneres - CLAD, DStatG, SGAPEIO, SEIO e APDIO - através da co-organização de sessões temáticas. Participámos, pelo 3º ano consecutivo, na VI Feira da Matemática, que decorreu nos dias 25 e 26 de Outubro, no Museu de História e Ciência e que acolhe a visita de muitas escolas e famílias.

Apoiámos a participação do Miguel de Carvalho no evento Global Portugueses Mathematicians 2 que se realizou em Junho no Porto e o Workshop IWSM2019 que se realizou em Julho em Guimarães.

Ainda referente ao ano de 2019 quero realçar o Prémio Carreira SPE, que este ano homenageou o Professor Doutor Carlos Braumann e o Professor Doutor Kamil Feridun Turkman. Este Boletim inclui dois testemunhos sobre as suas carreiras.

A 28 de Novembro de 2020 a SPE faz 40 anos! Esta efeméride será celebrada com o XXV Congresso da SPE que terá lugar em Évora, entre 26 e 28 de Novembro. Este Congresso será o culminar de uma série de iniciativas a decorrer durante 2020 que se querem diversificadas e descentralizadas e que se destinam a comemorar o passado, refletir no presente e projetar o futuro da Estatística e dos Estatísticos em Portugal.

A Sociedade é dos sócios e para os sócios e é, essencialmente, o que os sócios fizerem dela. E por isso contamos com todos vós, os nossos sócios!

Até breve,

Porto, 10 de Novembro de 2019

Cordiais saudações

Notícias

• XXIV Congresso SPE

Foi num cenário inspirador com memórias do Românico que decorreu, de 6 a 9 de Novembro de 2019, o XXIV Congresso da Sociedade Portuguesa de Estatística, nas instalações do Hotel Casa da Calçada. Um palácio do século XVI que nos ofereceu uma experiência inesquecível, desde as salas onde decorreram as várias sessões, aos jardins envolventes e ao encanto das vinhas. Amarante recebeu-nos muito bem!

Estes 4 dias de intenso trabalho contaram com 208 participantes que vieram de norte a sul de Portugal e ilhas, mas também de Alemanha, Angola, Brasil, Escócia, Espanha, França, Inglaterra, entre outros. Decorreram 16 comunicações convidadas temáticas, 26 comunicações convidadas organizadas, 64 comunicações orais e foram apresentados 43 posters e 4 sessões plenárias.

Quarta-feira, 06 de Novembro de 2019

À semelhança dos anteriores, o XXIV Congresso da Sociedade Portuguesa de Estatística iniciou-se com o esperado minicurso. Este ano, sob o título "Análise estatística de dados financeiros" e proferido por Conceição Amado, Cláudia Nunes e Alberto Sardinha. O curso foi dividido em duas partes: na primeira parte foram apresentados vários conceitos fundamentais da área das finanças, enquanto que na segunda parte abordaram-se algumas das técnicas estatísticas usadas na análise de dados financeiros. Foram apresentadas algumas ferramentas de cálculo estocástico e exemplos de aplicação.

O minicurso terminou às 16h com uma pausa para café, que proporcionou o encontro e reencontro dos inúmeros participantes, num clima de grande afetuosidade.

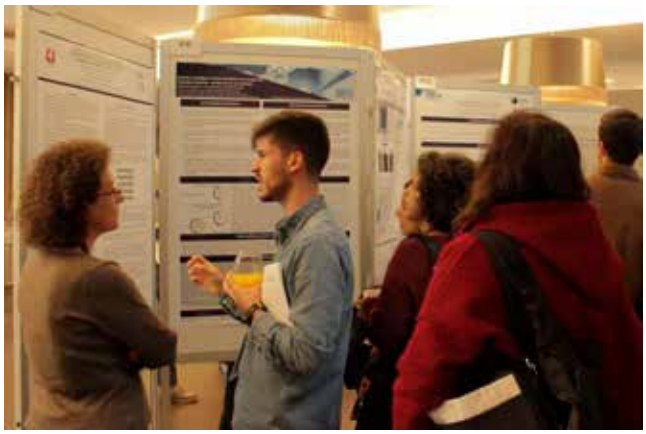
A cerimónia de abertura do congresso teve início às 16h30, com a presença e as palavras de boas vindas das Presidentes da comissão organizadora, Maria João Polidoro e Sandra Ramos, da Presidente da Escola Superior de Tecnologia e Gestão, Dorabela Gamboa, do Vice Presidente do Instituto Superior de Engenharia do Porto, Roque Brandão, do Vereador da Câmara Municipal de Amarante, André Magalhães, Diretor de Metodologias do Instituto Nacional de Estatística, Pedro Campos, e da Presidente da SPE, Maria Eduarda Silva. De seguida decorreu a primeira sessão plenária, proferida por Walter J. Radermacher, Presidente da FENStatS, Professor da Universidade de Roma, Itália, com o tema "The role of statistics in the digitised and globalised world".

Embora o dia já fosse longo, ainda decorreu o primeiro conjunto de comunicações orais em paralelo. Em 4 salas distintas falou-se sobre aplicações nas ciências sociais, análise de sobrevivência, estatística multivariada e métodos bayesianos.

Para culminar este dia, fomos acarinhados com uma receção de Boas Vindas, nos salões do Paço do Concelho, oferecida pelo Presidente da Câmara Municipal de Amarante, José Luís Gaspar, e por André Magalhães, InvestAmarante.
Fomos deliciados com vários produtos regionais, desde os fumeiros aos doces, passando pelo inconfundível vinho da região.

Quinta-feira, 07 de Novembro de 2019

Mais um dia intensivo de trabalho! O segundo dia do congresso iniciou com várias sessões temáticas a decorrer em 3 salas em paralelo. Uma das sessões foi organizada em parceria entre SPE e CLAD, e as restantes incidiram sobre equações diferenciais estocásticas e estatística industrial.
Logo a seguir, enquanto bebíamos um chá, um café ou um sumo de laranja, pudemos visitar a primeira sessão de posters e falar com os respetivos autores. Este foi mais um excelente momento de partilha de experiências.



As comunicações orais que se seguiram, em paralelo, centraram-se na bioestatística e epidemiologia, séries temporais e ciência de dados, dando destaque aos novos métodos ou adaptações necessárias a cada uma das áreas de aplicação. Em todos as sessões temáticas abordaram-se assuntos muito atuais e relevantes.

Depois da pausa para almoço, decorreu a segunda sessão plenária proferida por Maria do Rosário Oliveira, Professora e investigadora do Instituto Superior Técnico, Universidade de Lisboa, com o tema "Data Science, Data Science, Data Science,… Are you a data scientist?", que falou da ciência de dados e do papel dos estatísticos neste vasto campo multidisciplinar.

A meio da tarde pudemos desfrutar do passeio do congresso. Atravessámos o rio Tâmega e fomos visitar o Museu Amadeo de Souza-Cardoso e a Igreja Matriz de Amarante (o extinto Convento dominicano de São Gonçalo de Amarante). De regresso foi possível contemplar a magnífica vista da ponte que ao entardecer, tem outro encanto.

Como o passeio abriu o apetite, fomos gentilmente encaminhados para um momento de degustação na Dolmen Espaço Douro & Tâmega. Esta foi mais uma excelente oportunidade de convívio onde pudemos apreciar o vinho e outros produtos alimentares com origem no Douro Verde.

Sexta-feira, 08 de Novembro de 2019

Foi mais um dia de intensa atividade com o alinhamento da manhã igual ao do dia anterior: uma sessão organizada em parceira entre SPE e SEIO, e 2 sessões temáticas sobre estatística na saúde e estatística em ecologia e ambiente, tudo novamente em paralelo. A série de sessões foi seguida da segunda apresentação de posters e comunicações orais sobre probabilidade e processos estocásticos, aplicações em ambiente, clima, geociências e agricultura, bioestatística e epidemiologia e métodos não paramétricos.

Depois do almoço, decorreu a terceira sessão plenária proferida por Bruno Falissard, Professor da Universidade de Paris-Sud, com o tema "The crisis of statistical inference: time has come for an epistemological reflection". Nesta sessão foi feita uma reflexão sobre o papel da inferência estatística no paradigma atual. Bruno Falissard alertou para os perigos atuais e para a facilidade com que estes podem surgir. É preciso estarmos atentos e manter os fundamentos que são essenciais.

Ao longo do programa foi-se falando da evolução do papel da estatística na nossa sociedade e dos desafios daí decorrentes. Mas que dizer do ensino da estatística, o que mudou nos últimos 20 anos? Esta é uma reflexão fundamental que foi debatida em algumas sessões e que esperamos ver continuada nos próximos congressos. Nas sessões temáticas que decorreram durante a tarde, foram abordados os temas: o ensino da estatística e as novas tecnologias, investigação em saúde em África, estatísticas espácio-temporais; e decorreu mais uma sessão organizada, esta em parceria entre SPE e APDIO. Depois de uma pausa para café e de mais um conjunto de comunicações orais sobre aplicações em ambiente, clima, geociências e agricultura, análise de sobrevivência, extremos e ciência de dados, a decorrer em paralelo em 4 salas, decorreu ainda uma reunião informal da SPE. Os membros da SPE tiveram oportunidade para debater alguns assuntos e definir algumas estratégias. Todos os momentos são de debate e de reflexão!

Como é habitual, a sexta-feira do congresso culminou com mais um dos seus pontos altos, o Jantar, e a atribuição do Prémio Carreira. Este ano, foram distinguidas duas personalidades: Carlos Alberto dos Santos Braumann e Kamil Feridum Turkman. Russell Alpizar-Jara apresentou o seu colega e amigo Carlos Alberto dos Santos Braumann, fazendo referência à sua longa lista de contributos na área da estatística. Kamil Feridum Turkman, o segundo homenageado, foi apresentado por Patrícia de Zea Bermudez antiga aluna e agora sua colega, em mais uma homenagem a uma vida de trabalho dedicada à estatística. Foi momento de merecido rejúbilo, no qual tivemos uma grande honra em participar!

Ainda durante o jantar fomos presentados, a título gracioso, com o canto da viola amarantina renascida, pelas mãos de Professor Eduardo Costa, membro fundador da Associação Viola Amarantina, representante da Viola como Tocador e Professor da Escola de Viola Amarantina.

Sábado, dia 9 de Novembro de 2019

Logo depois de mais uma série de sessões paralelas, constituída por uma sessão organizada em parceria entre SPE e DstatG, outra em parceria entre SGAPEIO, SEB E SBioSPE e uma sessão temática sobre filas de espera, decorreu a 4ª e última sessão plenária proferida por Maria Manuela Neves, Professora e investigadora do Instituto Superior de Agronomia, Universidade de Lisboa, Portugal, com o tema "Procedimentos computacionais em Estatística".

A seguir à sessão plenária, decorreu mais uma série de sessões temáticas sobre estatística espacial, educação, estatística computacional e análise longitudinal.

Durante a tarde, já em jeito de despedida, foram atribuídos o Prémio SPE e os Prémios Estatístico Júnior. O Prémio SPE 2019 foi atribuído a Jessica Lomba, pelo seu trabalho intitulado "L-Moments for automatic threshold selection in extreme value ananlysis".

Os Prémios Estatístico Júnior foram atribuídos no decorrer das deliciosas brincadeiras do Circo Matemático.

O XXIV Congresso terminou com a respetiva sessão de encerramento, com as palavras dos membros da comissão organizadora, Luísa Hoffbauer, Maria João Polidoro, Sandra Ramos e Ana Borges, e da Presidente da SPE, Maria Eduarda Silva.

Com este texto, queremos agradecer a todos os que contribuíram para a realização deste excelente evento, com um agradecimento especial à comissão organizadora. Os nossos parabéns pela organização deste maravilhoso congresso!
Foi uma experiência fantástica em todas as vertentes, desde a vertente científica à vertente social, passando também pela vertente afetiva.
Houve um sentimento geral de um acolhimento muito bom!
Parabéns também pela escolha do local, pois há qualquer coisa em Amarante que nos toca e nos faz querer voltar.

Conceição Ribeiro e Cláudia Marisa Silvestre

• ECAS2019 on Statistical Analysis for Space-Time Data

O *ECAS2019 on Statistical Analysis for Space-Time Data*, organizado pela SPE (Sociedade Portuguesa de Estatística) e pelo SEIO (*Spanish Society of Statistics and Operational Research*) realizou-se na Faculdade de Ciências da Universidade de Lisboa, entre 15 e 17 de julho de 2019. Para além do patrocínio das duas sociedades organizadoras, foi apoiado pelo CEAUL (Centro de Estatística e Aplicações Universidade de Lisboa), CIM (Centro Internacional de Matemática), pela FCT (Fundação para a Ciência e Tecnologia) e pela Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, e pelo projeto PREFERENTIAL: PTDC/MAT-STA/28243/2017.

https://ecas2019.math.tecnico.ulisboa.pt/

No dia 15 de julho, vindo da Dinamarca, Ege Rukak ministrou um curso sobre padrões pontuais espaciais: metodologia e aplicações com o software R. O curso forneceu uma introdução detalhada ao pacote do R, *spatstat* e abordou tópicos que incluem análise exploratória de dados, estimação não-paramétrica da intensidade, modelos paramétricos (Poisson, Cox, Neyman-Scott, Gibbs), ajuste de modelos, simulação e testes de Monte Carlo. Durante a tarde, realizou-se uma sessão com 25 posters, permitindo que os participantes apresentassem o seu próprio trabalho.

No segundo dia do encontro Liliane Bel, vinda de França ministrou a primeira parte do curso intitulado novas tendências em geoestatística espaço-temporal. No curso foram investigados métodos para modelar e estimar a função de covariância em algumas estruturas espaciais, foram estudados alguns exemplos com dados reais e apresentados pacotes do R dedicados à estimação da covariância e à simulação e previsão espaço-temporal. Neste dia, Patrick Brown do Canadá também ministrou o curso sobre modelos estatísticos e inferência para dados espaço-temporais em áreas, dando especial foco aos aspetos práticos do uso desses modelos, bem como na interpretação e apresentação dos resultados, utilizando o R-INLA.

No último dia, Liliane Bel apresentou a segunda parte do curso sobre novas tendências em geoestatística espaço-temporal. Vindo da Arábia Saudita Hävard Rue, conjuntamente com Haakon Bakka, apresentaram modelos espaciais e espaço-temporais usando equações diferenciais parciais estocásticas (abordagem SPDE), dando ênfase à forma de realizar a análise Bayesiana desses modelos usando eficientemente o R-INLA.

(fotos de Paula Simões)

O Curso contou com 56 participantes, de 14 países.
Os participantes, na sua grande maioria alunos de doutoramento, puderam apresentar a sua investigação numa agradável sessão de posters onde, num ambiente descontraído de convívio, trocaram ideias e expuseram os seus trabalhos com os demais participantes e conferencistas.
Compareceram um total de 61 conferencistas, dos quais 26 eram alunos, superando as expetativas quanto à capacidade prevista do evento. Cerca de metade dos participantes eram de Portugal, 6 vieram da Alemanha, 5 de Espanha e os restantes vieram de locais tão variados como Suécia, Itália, Bélgica, França, Reino Unido, Irlanda, Moçambique, Canadá, Colômbia, México e Austrália.
Foi uma excelente oportunidade de aprendizagem de conhecimento de ponta nesta área da Estatística Espaço-Temporal e também de *networking*, proporcionados pelos diversos agradáveis momentos de pausa preparados pela organização. Esta organização ficou a cargo de Isabel Natário (presidente), Paulo Soares, Soraia Pereira, Tomás Goicoa, Anabel Forte e Giovani Silva.
A comissão científica do evento era constituída por Giovani Silva (presidente), Raquel Menezes, Maria Eduarda Silva, María Dolores Ugarte, Rubén Fernández Casal e Ricardo Cao.

No final foi unânime o reconhecimento do sucesso do encontro na sua componente científica e foi prestado um agradecimento público a todos quantos contribuíram para a organização do evento.

Andreia Monteiro
"Comissão Organizadora"

• Contributo do INE para o Boletim da SPE - outono 2019

Secção "A SPE e a Comunidade"

É com muito gosto que o INE passa a ter uma colaboração permanente no Boletim da Sociedade Portuguesa de Estatística.

Nesta secção será dado conhecimento de trabalhos diversos da responsabilidade do INE, a maioria dos quais no domínio de aplicações metodológicas e que foram apresentados em eventos científicos nacionais ou internacionais.

Esperamos que esta colaboração vá de encontro a todos aqueles que procuram uma componente mais prática na área da Estatística.

Nesta primeira participação no Boletim, incluem-se dois trabalhos que foram apresentados no recente *34th Internacional Workshop on Statistical Modelling*, que decorreu em Guimarães no passado mês de julho.

Pedro Campos
Carlos Marcelo
Serviço de Metodologia do Departamento de Metodologia e Sistemas de Informação do INE

• Statistics on Health Decision Making

No sentido de se partilhar o que melhor se faz usando a estatística médica e conjuntamente dinamizar a investigação desta área em Portugal surge a revista "Journal of Statistics on Health Decision" e o encontro "Statistics on Health Decision Making".
A Universidade de Aveiro, conjuntamente com o Centro Hospitalar do Baixo Vouga, promoveu a primeira edição do evento "Statistics on Health Decision Making", juntando investigadores do mundo clínico, académico e empresarial, em verdadeira discussão sobre as linhas estratégicas da estatística na decisão em saúde.
Este encontro teve a sua primeira edição a 10 de maio de 2019, na Universidade de Aveiro, com o subtema "STATE OF THE ART", juntou mais de 100 participantes, e contou com um vasto programa científico composto por quatro sessões convidadas, uma mesa-redonda e uma sessão de comunicações em pósteres.

As sessões convidadas, marcadamente desafiantes mas esclarecedoras, deram voz a António Vaz Carneiro – Universidade de Lisboa, Rumana Omar – University College London, Milton Severo – Universidade do Porto e José Aranda da Silva.

A sessão de pósteres juntou 43 trabalhos, envolvendo a estatística na tomada de decisão nas mais variadas áreas da saúde. O Prémio "Nascimento Leitão – *Bayer prize*", que visa premiar uma comunicação que se evidencie pela originalidade, impacto para a decisão em saúde, rigor científico e qualidade da apresentação, foi atribuído à comunicação intitulada "*The role of gender in survival after bilateral internal mammary artery in coronary artery bypass grafting: a propensity score analysis*", da autoria de Francisca Saraiva e co-autores (Faculdade de Medicina da Universidade do Porto e Centro Hospitalar e Universitário de São João). Os resumos estendidos de todos os trabalhos apresentados no evento foram publicados numa edição especial da revista *Journal of Statistics on Health Decision*.

Em 2020 teremos a segunda edição do encontro, a 8 de maio, com o subtema "CLINICAL TRIALS" (https://www.ua.pt/estatisticamedica).

Reserve a data e usufrua de uma oportunidade de partilhar conhecimento, promover novas colaborações, conversar com amigos e fazer novos!

Vera Afreixo e Ana Tavares

• Dia Europeu da Estatística 2019

A comunidade estatística europeia celebra anualmente o Dia Europeu da Estatística, a 20 de Outubro. O objetivo é sensibilizar os cidadãos para a importância das estatísticas oficiais na tomada de decisões fundamentadas, por parte de governos, empresas, investigadores, imprensa, bem como pela sociedade em geral.

O Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) não podia deixar passar este dia em vão e organizou um evento para a sua celebração, no dia 21 de outubro (segunda - feira), por iniciativa dos jovens investigadores da unidade. Este evento teve lugar na Faculdade de Ciências da Universidade de Lisboa e contou com mais de 30 participantes. O programa incluiu 3 comunicações sobre aplicações da Estatística em diferentes áreas de investigação e uma breve apresentação de colaborações desenvolvidas entre o CEAUL e outras unidades de I&D da FCUL. Houve ainda tempo para uma discussão aberta com café e bolinhos, promovendo novas interações.

A abertura do evento contou com um breve discurso de Lisete Sousa, coordenadora do CEAUL, acerca do contexto da celebração deste dia e da importância das estatísticas oficiais.

A qualidade e dinâmica das comunicações dos oradores convidados foi sem dúvida apreciada pela audiência e representativa da importância da Estatística nas mais diferentes áreas de investigação. Os três trabalhos apresentados são de Hygor Piaget (CFTC - Centro de Física Teórica e Computacional), Carlos da Câmara (IDL - Instituto Dom Luiz) e Victor Sousa (cE3c - Centro de Ecologia, Evolução e Alterações Ambientais). O primeiro orador falou-nos sobre o preço de um voto, num contexto de deseconomia em eleições proporcionais. Mostrou como a Estatística pode ganhar um papel relevante numa análise de campanhas eleitorais. O segundo orador falou sobre a Estatística ao serviço da prevenção e combate de incêndios rurais em Portugal, enquanto o terceiro orador fez uso de uma ponte entre a Estatística e a biologia evolutiva para inferir a história de determinadas espécies a partir dos seus genótipos.

Por fim, Tiago Marques falou-nos acerca da estratégia do CEAUL, áreas atuais de investigação e colaborações entre o CEAUL e outras unidades, dentro e fora da FCUL. Esta apresentação mostrou sem dúvida a dinâmica que esta unidade de investigação tem criado e a abertura para trabalhos multidisciplinares. As duas últimas frases desta sessão reforçam a ideia: "Visit us!" and "Invite us to visit you!".

Raquel Correia e Soraia Pereira

• Sobre a sessão de entrega dos Prémios "Estatístico Júnior 2019"

A Sociedade Portuguesa de Estatística promove anualmente o Prémio Estatístico Júnior.
Com esta iniciativa pretende-se incentivar o interesse pelas áreas de Probabilidades e Estatística dos estudantes dos Ensinos Básico e Secundário, e dos Cursos de Educação e Formação (CEF) e de Educação e Formação de Adultos (CEFA).
O Prémio Estatístico Júnior (PEJ) distingue anualmente sete trabalhos e é atribuído aos estudantes que os realizaram e a alguns dos professores orientadores, sendo a sua entrega formal realizada numa sessão que lhe é expressamente consagrada.
A entrega dos PEJ 2019 decorreu numa sessão especial, realizada em Amarante no passado dia 9 de novembro, integrada nas atividades do XXIV Congresso SPE e da qual, mais adiante neste Boletim, damos o devido destaque.
O Júri dos Prémios Estatístico Júnior 2019 integrou a Doutora Maria Eugénia Graça Martins e a Doutora Manuela Neves.

FR

• Prémio SPE 2019

O Prémio SPE, é promovido pela Sociedade Portuguesa de Estatística e pretende estimular a atividade de estudo e investigação científica em Probabilidades e Estatística entre os jovens.

Júri do Prémio SPE 2019:

- Prof. Manuel Scotto (Presidente), da Universidade de Lisboa.
- Prof. Ana Ferreira, da Universidade de Lisboa.
- Prof. Esmeralda Gonçalves, da Universidade de Coimbra.

O Prémio SPE 2019 foi atribuído a Jessica Silva Lomba, estudante de Doutoramento em Estatística na Faculdade de Ciências da Universidade de Lisboa.
No final desta edição do Boletim damos o devido destaque desta notícia.

FR

• Sessão Prémio Carreira SPE 2019

O Prémio Carreira – SPE foi instituído, em 2013, pela Sociedade Portuguesa de Estatística (SPE) e propõe-se reconhecer a atividade de estatísticos portugueses com papel de relevância no desenvolvimento científico, pedagógico e de divulgação da Estatística em Portugal.
Em 2019, o Prémio Carreira foi atribuído a dois Estatísticos com enorme relevo na SPE - o Professor Doutor Carlos Braumann da Universidade de Évora e o Professor Doutor Feridun Turkman da Universidade de Lisboa. A homenagem e entrega foram feitas numa sessão integrada no programa do XXIV Congresso da SPE, em Amarante; durante o Jantar do Congresso em 8 de novembro de 2019.
A Prof. Eduarda Silva, Presidente da SPE, fez a entrega dos diplomas e de lembranças comemorativas do Prémio Carreira – SPE aos Profs. Carlos Braumann e Feridun Turkman.
No final desta edição são apresentados dois testemunhos.

FR

• Comissão Especializada da Secção Biometria da SPE

As Comissões Especializadas e Representações na SPE dos actuais Órgãos Sociais da Sociedade Portuguesa de Estatística foram noticiadas no Boletim primavera de 2018.
A Comissão Coordenadora (CC) da Secção de Biometria da SPE, a partir de Julho de 2019, integra as colegas Inês Sousa (Universidade do Minho) e Laetitia Teixeira (ICBAS - Universidade do Porto) e Luzia Gonçalves (IHMT – Universidade Nova de Lisboa).

A Direcção da SPE

• A 28 de Novembro de 2020 a SPE faz 40 anos!

Em 28 de Novembro de 1980 foi fundada a, então denominada, Sociedade Portuguesa de Estatística e Investigação Operacional que daria, mais tarde, origem à SPE. Pelo que em 2020 a "nossa" SPE comemora 40 anos, o Conselho Geral propôs a organização de um Congresso para celebrar a data. Tendo a Direção abraçado a proposta com todo o entusiasmo, tenho o prazer de anunciar XXV Congresso da Sociedade Portuguesa de Estatística cuja Comissão Organizadora Local é constituída por: Russell Alpizar Jara (Presidente), Dulce Gomes, Patrícia Filipe, Lígia Henriques Rodrigues e peço a melhor atenção dos sócios para a mensagem da CO. Espero por todos em Évora!

Maria Eduarda Silva

• XXV Congresso SPE

O XXV Congresso da Sociedade Portuguesa de Estatística irá decorrer de 26 a 28 de novembro no Évora Hotel, em Évora, e será organizado pelo Departamento de Matemática da Escola de Ciências e Tecnologia e o Centro de Investigação em Matemática e Aplicações-IIFA da Universidade de Évora em parceria com a Sociedade Portuguesa de Estatística.

O objetivo principal deste Congresso é a celebração dos 40 anos da Sociedade que foi fundada no dia **28 de novembro de 1980**.

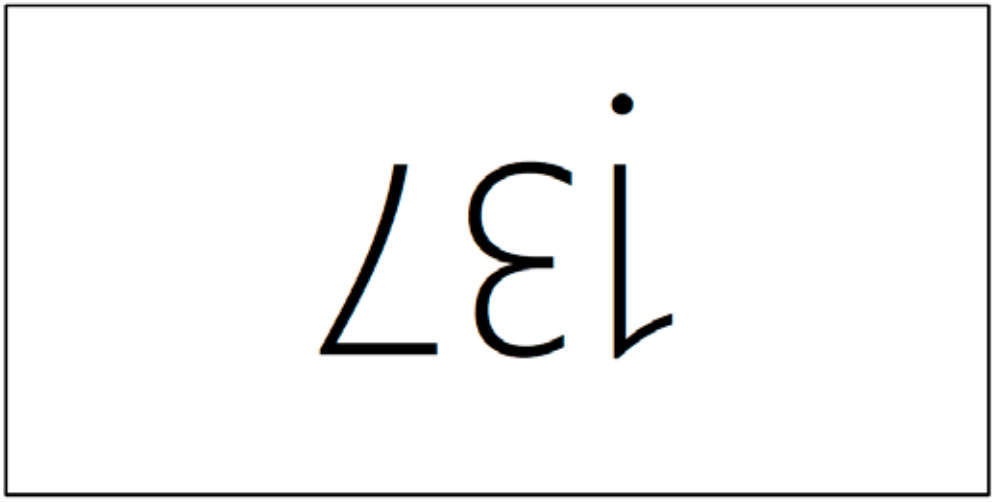
Assim, convidamos todos a participarem no XXV Congresso da SPE a decorrer na bela cidade de Évora, Património Mundial da UNESCO.

Mais informações poderão ser obtidas através do endereço *spe2020@uevora.pt* e oportunamente em *www.spe2020.uevora.pt*; bem como na página web da SPE e no próximo Boletim primavera de 2020.

Até novembro 2020, em Évora!

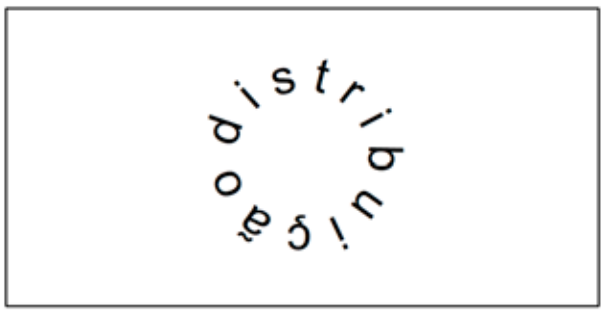
A Comissão Organizadora Local

Enigmística de mefqa



No Boletim SPE primavera de 2019 (p. 11):

big data



distribuição circular

SPE e a Comunidade

Nota Técnica sobre "diferenciação" e "diferençação"

Francisco Mercês de Mello, *facmm@live.com.pt*

Prof. Associado (aposentado) da Universidade de Évora

A análise das Sucessões Cronológicas é cada vez mais utilizada pelos alunos de Estatística. Econometria e das diferentes Engenharias. Parece pois óbvia a necessidade de clarificar a terminologia adequada, de modo a que designações que são aceitáveis na linguagem corrente, tratando-se de linguagem matemática deixam de o ser, e, daí, o porquê da presente Nota Técnica.

Escolhendo o livro de Box e Jenkins (1976) como texto, por excelência, das sucessões cronológicas, vamos dele extrair certas passagens, das páginas 8, 11 e 13, fazendo algumas considerações sobre as mesmas.

Na equação $\nabla Z_t = Z_t - Z_{t-1} = (1 - B)Z_t$ B refere-se ao operador atraso e ∇ é o operador diferença *(difference operator*). Mais adiante pode ler-se:

$$W_t = \nabla^d Z_t \qquad (1.2.6)$$

Homogeneous nonstationary behavior can threfore be represented by a model which calls for the d'th difference of the process to be stationary.

In practice d is usually 0,1, or at most 2.

Ora é aqui que reside o problema na tradução para português, pois é dito que se pode obter a estacionaridade, aplicando o operador diferença de ordem d, o que é diferente de diferenciar d vezes o processo.

Na página 13 da mesma obra (1.2.2 *Transfer function models*) refere-se o facto importante de nos modelos dinâmicos e contínuos se recorrer às equações diferenciais (*differential equations*), onde surge o operador diferencial D, e de nos modelos dinâmicos e discretos se recorrer às equações às diferenças (*difference equations*), onde aparece, como seria de esperar, o operador diferença ∇.

Convém lembrar que existem na língua inglesa os verbos *to difference* e *to differentiate,* que se podem traduzir por fazer a diferença e por fazer o diferencial. Assim, pelo rigor que a linguagem matemática exige, e em coerência com o que se ensina aos alunos no Cálculo Diferencial, não se pode dizer nas sucessões cronológicas, ao traduzir o símbolo $\nabla^d Z_t$ atrás citado, que se diferenciou d vezes z_t, mas sim que se diferençou d vezes z_t, ou, em alternativa, que se calcularam as diferenças de ordem d. Os termos "diferençar" e "diferençação" existem nos dicionários da língua portuguesa. E certamente por isso, já César de Freitas (1991) escrevia no seu livro Análise Numérica: "A somação é a operação inversa da diferençação, tal como a integração o é para a derivação" (página 31).

Neste contexto propõe-se que não se utilize a expressão incorrecta "diferenciando d vezes…" mas sim que se diga "fizeram-se as diferenças de ordem d…" ou "que se diferençou d vezes…"

Bibliografia

BOX, G; JENKINS,G. (1976) *Time Series forecasting and control* (Revised edition). Holden-Day.

SHIEL, Francis. (1991) *Análise Numérica* (2ª ed.). Tradução e Adaptação de A. César de Freitas. Shaum McGraw-Hill

Dados sintéticos como Ficheiros de Uso Público: uma aplicação ao Inquérito às Despesas das Famílias

Inês Rodrigues, *ines.rodrigues@ine.pt*

Instituto Nacional de Estatística

Dados relativos a unidades estatísticas individuais no formato de Ficheiros de Uso Público – em inglês, *Public Use Files* (PUF) – correspondem a registos de acesso universal e gratuito, preparados de modo a evitar a identificação direta ou indireta da respetiva unidade estatística. Com este trabalho, pretende-se desenvolver uma metodologia de controlo da divulgação estatística para efeitos de produção de PUF do Inquérito às Despesas das Famílias (IDEF) através da geração de dados sintéticos.
Foram utilizados métodos paramétricos (regressões logísticas multinomiais e log-lineares) e não paramétricos (algoritmos de árvores de decisão) para modelar a sequência de distribuições condicionais necessária à geração de um importante conjunto de variáveis, considerando as relações entre as mesmas. As duas abordagens foram comparadas, em particular no que se refere ao risco de divulgação de informação confidencial a partir do PUF resultante. A quantificação deste risco foi efetuada com base em duas medidas: risco esperado de ligação (entre as unidades originais e as unidades sintéticas), que reflete a possibilidade de um utilizador conseguir, de forma aleatória e para cada um dos registos do ficheiro de dados original, encontrar no ficheiro sintético uma unidade cujos valores relativos a um conjunto de variáveis identificadoras indiretas e variáveis confidenciais igualam os do registo original; e o risco real de ligação, que expressa a possibilidade de um utilizador identificar única e corretamente cada um dos registos do ficheiro original de dados, a partir do ficheiro sintético (casos em que existe um único registo no ficheiro sintético que coincide com o registo original, em termos das variáveis identificadoras indiretas e das variáveis confidenciais).
A geração de dados sintéticos permite produzir ficheiros com um nível de risco de divulgação muito reduzido, mantendo-se um elevado grau de utilidade. Os resultados obtidos pelas duas abordagens são idênticos, sendo que apenas se tende a verificar um ligeiro aumento do risco quando se recorre aos métodos não paramétricos; tal é justificado pela maior facilidade em modelar relações não lineares e interações entre as variáveis com base nestes métodos, comparativamente à abordagem paramétrica.
Para informação mais detalhada, poderá consultar a apresentação efetuada e o respetivo artigo em *https://www.ine.pt/xurl/doc/384908263*.

A Estimação em Pequenos Domínios para Uso e Ocupação do Solo

Pedro Campos, Suelma Pina, A. Manuela Gonçalves, *pedro.campos@ine.pt*

Instituto Nacional de Estatística

A Estimação em Pequenos Domínios (ou *Small Area Estimation - SAE*) é uma abordagem estatística que combina amostragem e inferência em populações finitas com modelação estatística. O principal objetivo deste trabalho é analisar e testar a implementação de diferentes tipos de estimadores de pequenos domínios, de modo a melhorar a qualidade das estimativas produzidas no âmbito do Inquérito à Estrutura das Explorações Agrícolas (FSS) ao nível das NUTS III. A aplicação é efetuada no âmbito do projecto LUCAS (*Land Use and Land Cover Survey*), que fornece informação estatística harmonizada e comparável sobre o uso e ocupação do solo em todo o território da UE.
Os resultados deste trabalho mostram que os estimadores modificados e indiretos Reg, SEBLUP e EBLUP, apresentam maiores ganhos de precisão quando o tamanho da amostra é maior e quando a correlação entre a variável dependente (área agrícola) e as variáveis independentes é maior. Ao analisar as estimativas do coeficiente de variação (CV) dos diferentes estimadores estudados por NUTS III para as variáveis mais importantes, a SAU (Superfície Agrícola Utilizada) das regiões do Baixo Alentejo (184) e do Alentejo Central (187) são as que apresentam os maiores valores de CV quando comparados com os das restantes regiões do nível III da NUTS.
Para informação mais detalhada, poderá consultar a apresentação efetuada e o respetivo artigo em *https://www.ine.pt/xurl/doc/384908714*.

Limitações dos estudos baseados em amostras de doentes/utentes

Vera Afreixo[1,2] *vera@ua.pt*, Ana Helena Tavares[1,3] *ahtavares@ua.pt*,
Tiago Gregório[4] *tiagogreg@gmail.com*

[1] *Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), Universidade de Aveiro*
[2] *Departamento de Matemática, Universidade de Aveiro*
[3] *Instituto de Biomedicina (iBiMED), Universidade de Aveiro*
[4] *Centro Hospitalar de Vila Nova de Gaia/Espinho, EPE*

O termo "evidência do mundo real" (Real-World evidence) é amplamente usado em ciências da saúde, referindo-se a informações sobre cuidados de saúde que advêm não apenas de registos clínicos, mas também de várias fontes fora da pesquisa clínica típica [1]. As duas principais fontes de dados do mundo real na pesquisa médica são os registos clínicos eletrónicos e os *claim data*. Os registos clínicos contêm dados sobre diagnósticos, exames e tratamentos. Os *claim data* representam a parte mais "administrativa" do histórico de saúde de um doente, consistindo em códigos de registos hospitalares. Estes dados têm o benefício de seguir um formato relativamente consistente e de usar um conjunto padrão de códigos pré-estabelecidos que descrevem diagnósticos, procedimentos e medicamentos específicos, criando uma fonte abundante e padronizada de informações do utente.

Em ciências da saúde muitos estudos reportam resultados baseados em ambos os tipos de registos. Além das vantagens óbvias que existem nestes estudos no que diz respeito ao custo (monetário e de tempo) e à capacidade de recrutamento de um elevado número de doentes, aquilo que os clínicos mais procuram neles é a avaliação da "efetividade". O mundo idílico dos ensaios clínicos está ainda muito distante do mundo real dos doentes, pelo que se ganha confiança quando os resultados dos ensaios clínicos são confirmados neste tipo de estudos.

Este artigo de opinião incide sobre a temática dos estudos baseados em doentes e, em particular sobre algumas questões que achamos merecerem reflexão. Apresentamos uma reflexão sobre as limitações dos estudos baseados em doentes e, apesar de não haver soluções cabais para as superar, discutimos alguns cuidados a ter em conta no desenvolvimento deste tipo de estudos. Apesar do tema em discussão não ser extensamente discutido na literatura científica, existem outros temas relacionados que o são. Por exemplo, as limitações da investigação médica (e.g. [2-3]) é um tema amplamente abordado, demonstrando a existência de preocupação relativamente à confiança nos resultados dos estudos da saúde.

Cada indivíduo e, em particular cada doente, é um sistema extremamente complexo; um efeito observado num doente pode estar associado à ocorrência de um número elevado de condições (variáveis) e às muitas possíveis interações que ocorrem entre estas. Tal cenário leva-nos a assumir cada doente como um sistema quase caótico! Essas diferenças entre indivíduos, mesmo que em pequena escala, podem ser suficientes para alterar o efeito do tratamento do indivíduo, eventualmente mudando a direção do efeito. A ocorrência de múltiplos efeitos, em cada indivíduo, é uma possibilidade consensual. Alguns efeitos que são, geralmente, levados em conta nos estudos da saúde, por exemplo, os efeitos de idade, raça ou sexo. No contexto do estudo da ação de um tratamento, é bem relatado o efeito placebo, e alguns estudos (por exemplo, ensaios clínicos duplamente-cegos) são cuidadosamente preparados para salvaguardar este efeito. No entanto, outros efeitos, como exposições ambientais ou estados emocionais individuais, são efeitos difíceis de avaliar e/ou controlar. Em

estudos observacionais, como estudos retrospetivos, parece ser quase impossível controlar este tipo de variáveis. Esta tarefa pode ser difícil de realizar mesmo em estudos prospetivos observacionais. Pelo que prevalecem algumas questões: Quantas variáveis devem ser controladas? A recolha de todas essas variáveis compromete a viabilidade do estudo?

Os dados registados nos processos clínicos dos doentes são frequentemente omissos, por vezes devido a letra ilegível ou abreviaturas incompreensíveis e, não raras vezes, são omitidos dados. Os dados não foram registados para efeitos de investigação e por isso muitas vezes falta-lhes qualidade. No caso dos *claim data*, os dados podem estar limitados a informações que suportam reembolso, o que vai influenciar a captura de eventos. Porém, existem exemplos de registos prospetivos de doentes que tendem a ter dados completos e primam pela qualidade, como por exemplo o RIETE (base de dados internacional sobre doentes com tromboembolismo venoso) ou o GLORIA-AF (base de dados internacional sobre doentes com fibrilação auricular não valvular em risco de AVC).

A recolha da amostra é um importante passo do estudo, que exige um bom planeamento. A atribuição do grupo a cada indivíduo segue um procedimento aleatório? Efetivamente, alguns estudos baseados em doentes utilizam dados não aleatorizados (por exemplo, o indivíduo dirige-se ao hospital para receber um tratamento específico) o que leva à violação do pressuposto basal da inferência estatística. No entanto, é prática corrente realizar análise estatística sobre esse tipo de dados, incluindo inferência estatística. Evidencie-se que a falta de amostragem aleatória compromete fortemente a generalização dos resultados.

Os estudos observacionais não são randomizados: a verdade é que, apesar de os ensaios clínicos aleatorizados capturarem melhor eficácia que efetividade, estes são mais robustos no que diz respeito à validade interna. Nos últimos anos temos visto estudos de mundo real a tentar mimetizar os ensaios aleatorizados usando por exemplo a técnica de *propensity score matching* mas a verdade é que esta técnica estatística: i) só ajusta para variáveis conhecidas e medidas (ao contrario dos ensaios clínicos aleatorizados onde os grupos são semelhantes quer em termos de factores de confusão conhecidos, quer em termos de outros fatores); e ii) está dependente da qualidade dos dados. No entanto, estes métodos não substituem a randomização, como exemplo digno de reflexão, considere-se a discussão do artigo de Sabina A. Murphy [4] sobre dois trabalhos que, apesar de incidirem sobre o mesmo estudo e os mesmos doentes, alcançam conclusões contraditórias!

Se todos os indivíduos em estudo provierem de uma mesma instituição (por exemplo, um centro hospitalar) é pouco provável que estes sejam representativos da população em estudo. Por exemplo, cada região tem características ambientais específicas, cada instituição tem formas específicas de acompanhar os doentes, os tratamentos,… Os estudos multicentro permitem colmatar este problema, mas a realidade é que a grande maioria dos estudos publicados refere-se a estudos unicentro. No entanto, como é do conhecimento geral, os estudos locais desempenham um papel muito importante para a instituição onde é desenvolvido e, nesses casos, a representatividade global pode ser um problema menor!

Por outro lado, e não menos importante, algumas características dos indivíduos alteram-se durante o período de estudo, havendo poucos estudos que explorem essa fonte de variabilidade/viés. O controlo de todas as variáveis que afetam a resposta do indivíduo a um tratamento específico parece ser um grande desafio ou até mesmo impossível. Esse tipo de limitação deve ser claramente assumido/declarado e resolvido quando possível, tornando o estudo o mais transparente possível.

Uma questão geral relacionada com estudos de efeito do tratamento prende-se com o facto dos critérios de inclusão/exclusão não refletirem exatamente a população em risco, isto é, a população que efetivamente irá usufruir do tratamento (caso haja evidência do seu efeito). Trata-se de um problema? Os ensaios clínicos são desenhados para mostrar eficácia e segurança, de forma a levar à introdução das tecnologias/tratamentos na prestação de cuidados. Para este fim, os medicamentos são usados em populações muito selecionadas. Ora quando se parte para o mundo real, vemos que os doentes/indivíduos são bem mais heterogéneos entre si do que em grande parte dos ensaios clínicos. Neste contexto, poderíamos ser levados a considerar que o resultado de tais estudos são pouco informativos e, até mesmo, que poderiam contribuir para uma tomada de decisão errada, por parte do clínico. Será? A representatividade da população alvo é uma questão que merece extrema ponderação no delineamento de qualquer estudo da saúde. Na prática, sucede que à medida que os clínicos vão ganhando confiança no produto testado num ensaio clínico, começam a prescrevê-lo em doentes cada vez "piores", gerando mais informação. Os estudos de mundo real permitem obter informação sobre a

utilização de tecnologias/tratamentos em doentes "reais", “piores” que os doentes dos ensaios clínicos aleatorizados.

O senso comum indica-nos que um estudo é considerado apropriado se tiver capacidade para distinguir as alterações devidas à ação (o tratamento) das alterações devidas ao acaso (ruído), em particular, o mesmo se aplica aos estudos baseados nos doentes. Se o efeito produzido por um tratamento for forte, a sua deteção pode ser simples, mesmo na presença de ruído. Porém, se o efeito do tratamento for mais moderado ou fraco, a sua deteção pode tornar-se difícil ou impossível de ser efetuada com confiança. Um grande número de tratamentos de saúde produz efeitos pequenos, e estes podem ser ofuscados ou confundidos pelo grande número de variáveis que muitas vezes não estão sob análise. Como ultrapassar esta limitação em estudos baseados em doentes?

Por fim, falemos da reprodutibilidade dos estudos. Será realmente possível reproduzir estudos baseados em doentes? Talvez esta seja uma questão filosófica, mas é importante que seja discutida ou pelo menos refletida [5], uma vez que a reprodutibilidade é um princípio basal da ciência. Para tornar um estudo reprodutível, os dados e os métodos utilizados devem estar completamente disponíveis. A reprodutibilidade é um assunto de enorme interesse, mas também o é a proteção de dados e a confidencialidade das informações dos doentes. Portanto, resta reconhecer que esta dualidade resulta num problema sem uma solução à vista. Assim, a credibilidade dos estudos baseados em doentes é essencialmente baseada na confiança dos que a executam e dos que a validam!

Os estudos baseados em doentes não são perfeitos, no entanto, muitos dados de saúde estão disponíveis para alguns investigadores/profissionais de saúde e a maioria dos dados está informatizada. Estes dados contêm um potencial de informação útil à espera de ser desbravada, seria uma má opção não os aproveitar. Assim, os estudos baseados nos dados dos processos dos doentes, apesar de potencializadores de conclusões abusivas, se utilizados com a devida cautela, cumprem e continuarão a cumprir um papel importante na investigação em saúde.

Referências

[1] Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., ... & Shuren, J. (2016). Real-world evidence—what is it and what can it tell us. N Engl J Med, 375(23), 2293-2297.

[2] Morris, A. H., & Ioannidis, J. P. (2013). Limitations of medical research and evidence at the patient-clinician encounter scale. Chest, 143(4), 1127-1135.

[3] Theofanidis, D., & Fountouki, A. (2018). Limitations and delimitations in the research process. Perioperative nursing, 7(3) 155-163.

[4] Murphy, S. A. (2013). When ‘digoxin use’ is not the same as ‘digoxin use’: lessons from the AFFIRM trial, European Heart Journal, 34(20) 1465-1467.

[5] Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. Nature human behaviour, 1(1), 0021.

Análise de Dados Multiestado: aplicação a uma base de dados de cancro da mama

Luís F. Meira Machado, *lmachado@math.uminho.pt*

Departamento de Matemática, Universidade do Minho

Gustavo Soutinho, *gustavo.soutinho@ispup.up.pt*

EPIUnit, ICBADS, Universidade do Porto

1. Introdução

Os modelos multiestado são uma forma útil de descrever um processo no qual um indivíduo se move através de um número finito de estados em tempo contínuo. Em aplicações biomédicas, os estados podem representar condições de saúde (por exemplo, saudável, doente e morte), estados de uma doença (tais como, estados de um cancro ou de infeção por HIV) ou uma complicação não fatal no curso de uma doença (por exemplo, recidiva de um cancro, transplante de um órgão, etc.). A análise desses estudos, em que os indivíduos podem observar vários eventos, é em geral efetuada recorrendo ao modelo de regressão de Cox com covariáveis dependentes no tempo. Com este trabalho pretendemos ilustrar de que modo os modelos multiestado podem ser utilizados como alternativa a este modelo. Para demonstrar o potencial da metodologia descrita, utilizou-se uma base de dados de cancro da mama. Foram realizadas várias análises para avaliar os efeitos das diferentes variáveis preditoras (covariáveis) para as transições entre os diferentes estados. *Software* na forma de uma biblioteca para o R foi desenvolvido pelos autores.

Em muitos estudos clínicos, os doentes podem observar vários eventos ao longo de um período de acompanhamento. A análise destes estudos é frequentemente realizada recorrendo a modelos multiestado (Andersen et al., 1993; Hougaard, 2000; Meira-Machado et al., 2009, Meira-Machado e Sestelo, 2019). Estes modelos podem ser usados com sucesso para investigar o progresso de doentes ao longo de um determinado número de estados. Em geral, os estados representam a ocorrência de um evento que pode estar relacionado ao prognóstico de sobrevida, como complicações após uma cirurgia, recidivas ou episódios não fatais.

Graficamente, os modelos multiestado podem ser ilustrados recorrendo a diagramas com caixas representando os estados e com setas entre os estados representando as possíveis transições. A complexidade do modelo multiestado depende muito do número de estados e também das possíveis transições. O modelo de doença-morte (Figura 1) desempenha um papel central na teoria e na prática destes modelos, descrevendo a dinâmica de indivíduos 'saudáveis' que podem passar para um estado intermédio 'doente' antes de entrar num estado absorvente que em muitas situações é representado pela morte do indivíduo.

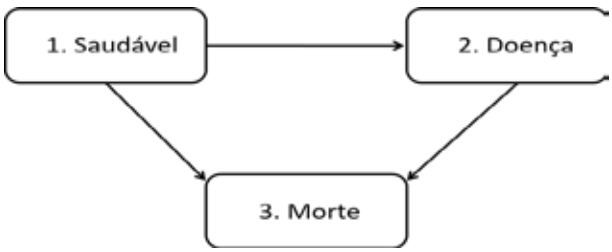


Figura 1: Modelo de doença-morte progressivo

Um objetivo importante na modelação multiestado é avaliar o possível efeito de um conjunto de fatores de prognósticos no curso de uma doença. Para relacionar as características individuais com as intensidades de transição, diversos modelos têm sido utilizados na literatura. Uma estratégia comum e que permite uma simplificação da análise, consiste em desagregar todo o processo em vários modelos de sobrevivência, ajustando modelos de regressão de Cox (Cox, 1972) para cada uma das transições, considerando alguns ajustes apropriados aos conjuntos de risco. De um modo geral o modelo pode ser escrito da seguinte forma:

$$h_{ij}(t;X) = h_{ij,0}(t)exp\left(\beta_{ij}^{T}X\right)$$

Em que $h_{ij,0}(t)$ denota a função de risco de suporte entre os estados *i* e *j*, β_{ij} é um vetor com os parâmetros de regressão e *X* é um vetor de covariáveis.

Para a função intensidade de mortalidade sem a doença, $\alpha_{13}(t;X)$, os tempos de sobrevida dos indivíduos que observaram a doença são considerados como censurados no tempo da doença. Indivíduos que permanecem vivos e livres de doença ('saudáveis') também contribuem com tempos de sobrevivência censurados. Para a intensidade da doença, $\alpha_{12}(t;X)$, o ponto final é o tempo de início da doença. Os tempos de sobrevida dos indivíduos que não adoeceram são considerados censurados, estejam eles vivos ou tenham falecido sem terem sido afetados pela doença. Por fim, para modelar $\alpha_{23}(t;X)$, a intensidade de mortalidade após a ocorrência da doença, apenas são considerados os tempos de sobrevivência (censurados ou não) truncados no tempo de doença dos indivíduos que observaram a doença. Note-se que os indivíduos estão em risco apenas depois de entrar no estado intermédio 2.

De realçar que, em alguns casos, podemos impor algumas condições sobre as funções de risco suporte. Por exemplo, para o modelo de doença-morte, uma abordagem que é frequentemente considerada, consiste em assumir que as funções de risco suporte para a transição do estado 1 para o estado 3 ($1 \rightarrow 3$) e para a transição entre o estado 2 e o estado 3 ($2 \rightarrow 3$) sejam proporcionais. Nestes casos, o modelo para estas transições é dado por:

$$\alpha_{13}(t;X) = \alpha_{13,0}(t)exp(\beta_{13}^{T}X) \text{ e } \alpha_{23}(t;X) = \alpha_{13,0}(t)exp(\beta_{23}^{T}X + \delta)$$

Na implementação de um modelo de regressão de Cox tal como apresentado nas equações em cima, é assumido que o efeito de cada uma das covariáveis tem uma forma funcional linear (ou log-linear). A presença de um efeito não-linear pode levar a consequências sérias com uma incorreta especificação do modelo, resultando em enviesamentos e numa diminuição do poder dos testes de significância estatística (Struthers e Kalbfleisch 1986; Anderson e Fleming 1995). Uma forma funcional incorreta pode levar também a um diagnóstico de riscos não proporcionais.
A ausência de flexibilidade dos modelos de sobrevivência (semi-)paramétricos levou, nas últimas décadas, ao desenvolvimento de uma variedade de métodos de regressão não paramétricos baseados em vários modelos estatísticos dos quais salientamos: a abordagem pelo modelo de riscos aditivos de Aalen (Martinussen e Scheike, 2006) e os modelos de regressão de Cox com preditores aditivos (Hastie e Tibshirani, 1990). Para introduzir flexibilidade no modelo de regressão de Cox, vários métodos de suavização podem ser aplicados, mas as splines penalizadas (P-splines), introduzidas por Eilers e Marx (1996) são as mais consideradas neste contexto. O modelo pode escrever-se da seguinte forma:

$$h_{ij}(t;X) = h_{ij,0}(t)exp\left(\sum_{k=1}^{q} f_{k,ij}\left(X_k\right)\right)$$

em que $f_{k,ij}(\cdot), k = 1, \ldots, q$ são funções suaves associadas a covariáveis quantitativas.

No contexto multiestado, dois diferentes modelos podem ser considerados tendo por base pressupostos feitos sobre a dependência das intensidades de transição e da história do processo. As intensidades de transição podem ser modeladas usando modelos de Cox separadamente, assumindo que o processo seja

Markoviano (que afirma que o passado e o futuro são independentes, dado o estado atual do processo). Nas situações em que o processo não verifica o pressuposto de Markov, é usual recorrer-se a um modelo semi-Markov em que se assume que o futuro do processo não depende do tempo atual, mas sim da duração no estado atual. Os modelos semi-Markov também são chamados de modelos de '*clock-reset*', pois cada vez que o indivíduo entra num novo estado, o tempo é redefinido para 0.

O pressuposto de Markov pode ser verificado pela inclusão de covariáveis dependendo da história. No caso do modelo doença-morte, o pressuposto de Markov apenas é relevante para a transição da mortalidade após a recidiva. Podemos por isso testar este pressuposto averiguando se o tempo de permanência no estado inicial 'Saudável' (ou seja, o passado) é importante na transição do estado de recidiva para a morte (ou seja, o futuro). Para testar este pressuposto na prática, denotemos por X o tempo de permanência no estado inicial. Ajustando um modelo de regressão de Cox $\alpha_{23}(t;X) = \alpha_{23,0}(t)exp(\beta X)$, precisamos testar a hipótese nula, H_0: $\beta = 0$, contra a alternativa mais geral, H_1: $\beta \neq 0$. Isto permite avaliar se a intensidade de transição do estado de doença para a morte não é afetada pelo tempo de permanência no estado anterior (i.e., se o pressuposto de Markov é válido). Métodos alternativos para testar o pressuposto de Markov em modelos multiestado foram discutidos no trabalho de Soutinho, Meira-Machado e Oliveira (2019).
Na seção seguinte são apresentados os resultados da aplicação dos métodos aqui referidos a um caso real de cancro da mama realizado no âmbito do *German Breast Cancer Study Group*.

2. Aplicação a um caso real

Vários estudos foram desenvolvidos nas últimas décadas sobre o cancro da mama. Entre 1983 e 1989, quatro ensaios clínicos foram realizados pelo '*German Breast Cancer Study Group (GBSG)*', incluindo 2746 pacientes com cancro da mama primário positivo. Detalhes sobre esses estudos podem ser encontrados no artigo de Schumacher et al. (1994). Neste trabalho, usamos dados de um destes ensaios, no qual um total de 720 mulheres com cancro da mama foi recrutado no período entre julho de 1984 e dezembro de 1989. Os dados, com informação completa para 686 mulheres, estão disponíveis como parte das bibliotecas do *software* R (*www.r-project.org/)* **mfp**, **TH.data** e **survidm**. Neste estudo, os doentes foram acompanhados desde a data do diagnóstico do cancro da mama até a censura ou a morte por cancro da mama. Do total de 686 mulheres, 299 desenvolveram uma recidiva e 171 morreram (21 das quais sem que se tenha observado uma recidiva). Além dos dois momentos dos dois eventos (recidiva e morte) e as correspondentes funções indicadoras de censura, um vetor de covariáveis incluindo a idade, tamanho do tumor, número de nodos positivos, recetor de progesterona e recetor de estrogênio, terapia hormonal e grau do tumor também estão disponíveis. Uma descrição das variáveis disponíveis na base de dados é apresentada na Tabela 1.

Tabela 1: Descrição das variáveis presentes no estudo sobre o cancro da mama

Variável	Descrição
rectime	Tempo até à recidiva
censrec	Ocorrência de recidiva (0: dado censurado)
survtime	Tempo de sobrevivência
censdead	Ocorrência de censura (0: dado censurado)
age	Idade aquando do diagnóstico
size	Tamanho do tumor (mm)
nodes	Número de nódulos linfáticos envolvidos (1-51)
prog_recp	Número de recetores de progesterona (1 – 2380)
estrg_recp	Número de recetores de estrogénio (1-1144)
menopause	Estado relativamente à menopausa (1: pre, 2: post)
hormone	Terapia hormonal (1: sim, 2: não)
grade	Grau do tumor (I, II e III)

A informação de uma possível recidiva, e correspondentes tempos, levam a que se possa considerar esta covariável como dependente do tempo. Esta covariável pode ser considerada como um estado transiente (intermédio) e modelada usando um modelo de doença-morte com estados 'Vivo e livre da doença', 'Vivo com Recidiva' e 'Morte'. Assim, começou por se analisar os fatores de prognóstico para a mortalidade pelo cancro da mama cujos resultados são apresentados na Tabela 2. Para tal foram utilizados modelos de regressão simples e múltipla recorrendo ao modelo de riscos proporcionais de Cox, considerando a ocorrência de recidiva como covariável dependente do tempo.

Pela análise dos resultados obtidos pode concluir-se que existe um efeito forte da ocorrência de recidiva na sobrevivência dos indivíduos ($P < 0.001$) em que a razão entre as funções de risco (HR - *hazard ratio*), no modelo de regressão múltipla ajustado, é de 33.5 superior nos caso das doentes em que se verificou recidiva. A idade, o tamanho do tumor e o número de recetores de progesterona são outros dos fatores que melhor explicam a mortalidade de acordo com o modelo de regressão múltipla ajustado.

Tabela 2: Modelos de regressão de Cox com *recidiva* como covariável dependente do tempo

		Simples			Múltipla	
Variável	n	HR	95%CI	valor-p	HR	95%CI	valor-p
recurrence	686	42.299	25.93-69.01	<0.001	33.565	20.434-55.134	<0.001
age	686	1.002	0.987-1.016	0.836	1.016	1.002-1.030	0.028
size	686	1.021	1.012-1.029	<0.001	1.013	1.004-1.022	0.007
nodes	686	1.071	1.053-1.088	<0.001	1.014	0.991-1.038	0.228
prog_recp	686	0.993	0.991-0.996	<0.001	0.996	0.994-0.998	0.001
estrg_recp	686	0.998	0.997-0.999	0.028			
Menopause							
Pre	290	1	-				
Post	396	1.116	0.821-1.517	0.484			
Hormone							
não	440	1	-		1	-	
sim	246	0.771	0.559-1.061	0.111	0.918	0.656-1.285	0.619
Grade							
I	81	1	-		1	-	
II	444	3.465	1.522-7.885	<0.001	1.149	0.493-2.678	0.747
III	161	6.438	2.776-14.930	<0.001	1.551	0.639-3.762	0.332

A utilização de interações entre as covariáveis fixas e a covariável dependente no tempo pode ser considerada como uma forma flexível (mas menos ambiciosa) de modelação multiestado. Neste caso, em que a covariável dependente no tempo é binária, esta modelação corresponde a uma situação em que se assume a proporcionalidade dos riscos para as transições $1 \rightarrow 3$ e $2 \rightarrow 3$ enquanto a transição $1 \rightarrow 2$ não é modelada. A interação entre a covariável dependente do tempo e as covariáveis fixas, permite modelar as situações em que a covariável tem diferentes efeitos antes e depois de ocorrer a covariável dependente no tempo (evento intermédio - recidiva). Os resultados da aplicação deste modelo multiestado 'parcial' aos dados do cancro da mama são apresentados na Tabela 3.

Os resultados sugerem que a idade tem um efeito sobre o tempo de sobrevivência pré-recidiva ($P=0.038$) com um *HR* de 1.049 (IC 95%: 1.003-1.099), mas também indicam que a mesma covariável terá uma menor importância para explicar a sobrevivência dos indivíduos que recidivaram ($P=0.107$). Há igualmente evidência estatística para assumir que o tamanho do tumor ($P=0.017$) e os números de recetores de progesterona ($P=0.003$) são fatores de prognóstico sobre o tempo de sobrevivência pós-recidiva, com *HR* ajustados de 1.012 e 0.997, respetivamente. Pode-se observar também que as restantes covariáveis não foram identificadas como fatores de risco para a ocorrência de morte (com ou sem recidiva).

Tabela 3: Modelo de regressão de Cox múltiplo com interações com *recidiva*

Variável	HR	95%CI	valor-p
recurrence	244.26	8.048-7413.9	0.002
recurrence0:age	1.049	1.003-1.099	0.038
recurrence1:age	1.012	0.997-1.027	0.107
recurrence0:size	1.016	0.989-1.044	0.246
recurrence1:size	1.012	1.002-1.022	0.017
recurrence0:nodes	1.041	0.986-1.098	0.144
recurrence1:nodes	1.010	0.986-1.036	0.416
recurrence0:prog_recp	0.995	0.989-1.001	0.106
recurrence1:prog_recp	0.997	0.994-0.999	0.003
recurrence0:hormone(s)	0.862	0.348-2.131	0.747
recurrence1:hormone(s)	0.938	0.653-1.349	0.731
recurrence0:gradeII	0.884	0.191-4.097	0.875
recurrence0:gradeIII	1.397	0.268-7.299	0.692
recurrence1:gradeII	1.241	0.446-3.455	0.700
recurrence1:gradeIII	1.663	0.574-4.822	0.349

De seguida, pretendeu-se estudar os fatores de prognóstico não apenas relativamente à mortalidade (com e sem recidiva), mas também sobre a ocorrência de recidiva. Para tal foram utilizados modelos de Cox para cada uma das transições do modelo multiestado associado aos dados do cancro da mama. Para o caso dos modelos ajustados para as transições do estado inicial 1 para os estados 2 e 3 (1→2 e 1→3), foram considerados os 686 indivíduos que iniciaram o estudo. Para a transição 2→3 foram apenas consideradas as 261 mulheres que recidivaram. Antes de proceder à realização desta modelação multiestado (via regressão de Cox), é necessário verificar o pressuposto de Markov que assume que o passado e a evolução da doença dependem apenas do estado atual em que se encontra o doente. Para tal foi analisada a influência do tempo em que o indivíduo permanece saudável (vivo e sem evidências de doença – estado 1) na transição do estado intermédio para o estado absorvente (i.e., intensidade de mortalidade em indivíduos que sofreram de recidiva). Pelos resultados obtidos há evidência de que o tempo de permanência não tem influência sobre os tempos de sobrevivência pós-recidiva e consequentemente pode assumir-se que um modelo de Markov é satisfatoriamente para o estudo em causa (P=0.121). A influência das covariáveis para cada uma das três transições, assumindo o modelo de Markov, é apresentada na Tabela 4 (modelos de regressão simples) e Tabela 5 (modelo de regressão múltipla). Os resultados da análise dos modelos de regressão de Cox simples foram considerados para a escolha do modelo de regressão de Cox múltipla.

Tabela 4: Modelos de Markov via regressão de Cox simples para todas as transições

	Recidiva			Mortalidade sem recidiva			Mortalidade com recidiva		
Variável	HR	95%CI	valor-p	HR	95%CI	valor-p	HR	95%CI	valor-p
age	0.992	0.980-1.004	0.183	1.046	0.999-1.094	0.051	1.011	0.997-1.025	0.129
size	1.014	1.007-1.022	<0.001	1.023	1.001-1.047	0.043	1.010	1.001-1.019	0.039
nodes	1.060	1.045-1.074	<0.001	1.073	1.023-1.126	0.004	1.025	1.001-1.089	0.040
prog_recp	0.997	0.996-0.999	<0.001	0.994	0.988-0.999	0.049	0.996	0.994-0.999	0.001
estrg_recp	0.999	0.998-1.000	0.056	0.999	0.995-1.002	0.443	0.999	0.998-1.001	0.232
menopause			0.846			0.204			0.296
Pre	1	-		1	-		1	-	
Post	1.024	0.806-1.301		1.847	0.716-4.765		1.192	0.858-1.655	
hormone			0.003			0.774			0.383
não	1	-		1	-		1	-	
sim	0.682	0.528-0.880		0.878	0.363-2.127		1.167	0.825-1.650	
grade									
I	1	-		1	-		1	-	
II	2.527	1.517-4.210	<0.001	1.299	0.291-5.809	0.732	1.760	0.645-4.807	0.270
III	3.234	1.880-5.563	<0.001	2.703	0.559-13.07	0.216	2.627	0.940-7.344	0.066

De salientar o facto de a covariável idade revelar uma maior importância quando ajustada pelo modelo de regressão múltipla. Este facto já acontecia quando se ajustou o modelo de regressão com

covariáveis dependentes no tempo. Em sentido inverso, nas transições 1→3 e 2→3, as covariáveis grade (grau do tumor) e nodes (número de nódulos linfáticos com o tumor) viram a sua importância diminuir no modelo de regressão múltipla, em parte explicadas pela sua correlação com a covariável size (tamanho do tumor).

Tabela 5: Modelos de Markov via regressão de Cox múltipla para todas as transições

	Recidiva			Mortalidade sem recidiva			Mortalidade com recidiva		
Variável	HR	95%CI	valor-p	HR	95%CI	valor-p	HR	95%CI	valor-p
age	0.997	0.985-1.010	0.665	1.053	1.006-1.102	0.028	1.013	0.998-1.027	0.095
size	1.007	0.999-1.015	0.111	1.019	0.992-1.046	0.166	1.011	1.001-1.021	0.030
nodes	1.050	1.034-1.066	<0.001	1.052	0.995-1.112	0.076	1.009	0.984-1.034	0.493
prog_recp	0.998	0.997-0.999	<0.001	0.995	0.989-1.001	0.071	0.997	0.995-0.999	0.004
hormone									
não	1	-		1	-		1	-	
sim	0.717	0.551-0.932	0.013	0.826	0.334-2.038	0.677	0.957	0.665-1.378	0.814
grade									
I	1	-		1	-		1	-	
II	2.034	1.214-3.407	<0.001	0.907	0.197-4.179	0.900	1.215	0.436-3.387	0.710
III	2.269	1.301-3.955	<0.001	1.602	0.302-8.068	0.576	1.593	0.558-4.630	0.392

A abordagem pelo modelo de Markov e pelo modelo de regressão de Cox, com recidiva como covariável dependente no tempo, proporcionam resultados similares, revelando impacto da idade na mortalidade em doentes sem recidiva e do tamanho do tumor e recetores de progesterona na mortalidade em doentes com recidiva. A vantagem do modelo de Markov é que ele permite considerar efeitos dos fatores de prognóstico sobre a recidiva, revelando o impacto do número de nódulos com tumor (nodes), do estatuto sobre a terapia hormonal, e do grau do tumor na recidiva.

Os resultados pela abordagem do modelo Markov indicam que excetuando a idade (P=0.665) e o tamanho do tumor (P=0.111) as restantes covariáveis têm influência no desenvolvimento de recidiva de cancro da mama. Relativamente à sobrevivência, para indivíduos em que não ocorreu recidiva, apenas a idade, com um *HR* de 1.053, tem uma influência direta (P=0.028). Já no caso da sobrevivência pós-recidiva, como fatores de prognóstico há a considerar o tamanho do tumor (P=0.030) e o número de recetores de progesterona (P=0.004) com HR, respetivamente, de 1.011 e 0.997.

O efeito de covariáveis contínuas sobre o logaritmo da função de risco (*log-hazards*) assume-se usualmente como tendo uma forma linear para cada uma das intensidades de transição de um modelo multiestado. Acontece que nem sempre este comportamento é verificado. Existem inúmeras abordagens para lidar com este problema, sendo os métodos de suavização de splines penalizadas (P-splines) propostos por Eilers e Marx (1996) frequentemente utilizados para este contexto. Estes métodos permitem introduzir alguma flexibilidade ao modelo de Cox, nomeadamente nos efeitos das covariáveis quantitativas. Os resultados da aplicação desta abordagem para estimar os efeitos dos preditores contínuos, nomeadamente da idade, sobre a intensidade para a ocorrência de recidiva são apresentados na Tabela 6. Efetivamente, a partir dos dados amostrais, é possível observar a presença de um efeito não-linear que provavelmente não teria sido detetado através de uma análise paramétrica, devido à ausência de informação prévia sobre a forma da curva HR correspondente. De realçar que o modelo ajustado na Tabela 5 (para a recidiva), e que considerava um efeito linear para as covariáveis quantitativas, não identificou um efeito da idade na recidiva (HR=0.997; IC 95%: 0.985-1.010)

Tabela 6: Modelo de regressão de Cox múltiplo para a intensidade de ocorrência de recidiva com efeitos não-lineares

Variável	HR	95%CI	valor-p
age			
age (linear)	0.993		2.1e-01
ps(age, df=4.9) (nonlinear)			1.6e-05
nodes	1.047	1.032-1.063	6.9e-10
size	1.009	1.001-1.017	3.1e-02
prog_recp	0.998	0.997-0.999	4.4e-04
Hormone			3.7e-03
não	1	-	
sim	0.675	0.518-0.880	
Grade			
I	1	-	
II	1.960	1.169-3.287	1.0e-02
III	2.078	1.188-3.634	3.7e-03

A variável contínua idade foi ajustada com uma componente linear e uma componente não linear. A componente linear tem uma função de razão de risco de 0.9929 (P=0.21). A componente não linear, com um valor de prova de 1.6e-05 indica que o efeito da idade na recidiva é não linear. Para obter resultados interpretáveis de maneira simples e resumida, construímos curvas flexíveis para a *HR* com intervalos de confiança a 95% para descrever a relação entre a idade e a (logaritmo) razão de riscos para a recidiva, considerando um valor específico como referência. A Figura 2 apresenta a correspondente curva para um valor de referência de 50 anos, valor selecionado como um possível valor para o início da menopausa. O correspondente gráfico confirma a existência de um efeito não linear entre a idade da mulher e o risco de recidiva, revelando uma relação decrescente com a idade, com um ligeiro crescimento após os 47 anos de idade e permanecendo aproximadamente constante posteriormente. Os dados são bastante dispersos em idades mais avançadas, conforme refletido pelo amplo intervalo de confiança nestas idades. Esta representação gráfica revela que o risco de recidiva é maior para mulheres mais jovens, por exemplo, a função razão de riscos toma o valor de exp(1.1032) = 3.0138 (com IC 95% 1.8083-5.0230) quando uma doente com 30 anos é comparada com uma doente de 50 anos (valor de referência).

3. *Software*

Vários investigadores desenvolveram nos últimos anos *software* para a análise de dados de sobrevivência multiestado. Uma lista abrangente das bibliotecas disponíveis na rede de distribuição do *software* R (CRAN) pode ser obtida no 'CRAN task view' sob o tópico 'Survival Analysis' (Allignol e Latouche, 2019).

Para fornecer aos investigadores biomédicos uma ferramenta de utilização fácil no contexto dos modelos multiestado, foi desenvolvida uma biblioteca para o *software* estatístico R chamada **survidm**. Esta biblioteca pode ser utilizada para efetuar regressão multiestado recorrendo a modelos de (semi-)Markov de Cox. A biblioteca vai muito para além da regressão multiestado, proporcionando aos investigadores biomédicos a possibilidade de obterem outros resultados interpretáveis de uma forma simples e resumida. Isso inclui estimativas de várias probabilidades com carácter preditivo, tais como, probabilidades de transição, probabilidades de ocupação, função de incidência cumulativa e a distribuição do tempo de permanência em cada estado. Uma limitação da biblioteca **survidm** é que esta só pode ser utilizada para o modelo progressivo de doença-morte. No entanto, isso acaba por se revelar uma vantagem para os usuários que desejam analisar dados de um modelo com esta estrutura. Nestes casos, a biblioteca **survidm** é ideal, pois é fácil de usar e tem uma forte semelhança com a biblioteca **survival**, bem conhecida e amplamente utilizada na comunidade de usuários do *software* R. Para modelos com uma estrutura diferente do modelo doença-morte recomendamos a utilização da biblioteca **mstate**.

As bibliotecas **survival** e **mgcv** de Terry Therneau e Simon Wood, respectivamente, podem ser utilizadas para introduzir flexibilidade no modelo de regressão de Cox. A biblioteca **smoothHR**,

desenvolvida por Araújo e Meira-Machado, permite o cálculo de estimativas pontuais para a razão da função de risco - e seus correspondentes limites de confiança - de preditores contínuos considerando um efeito não linear, introduzidos recorrendo a splines penalizadas.

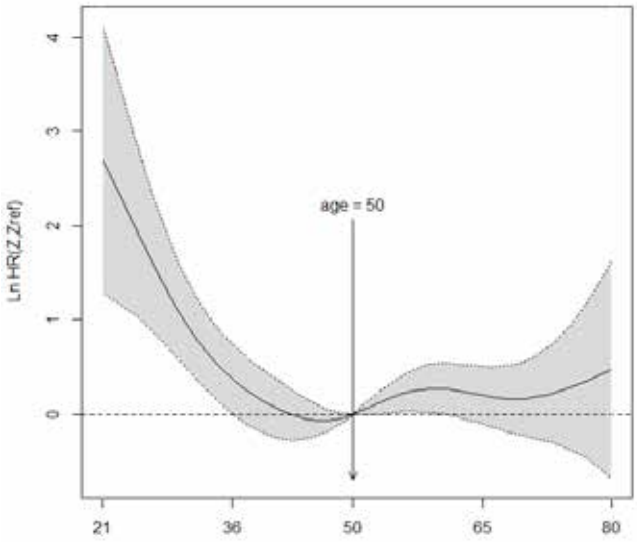


Figura 2: Estimativas não paramétricas da dependência da taxa de risco de recidiva (com limites de confiança de 95%) em pacientes com cancro da mama. Valor de referência de 50 anos de idade.

Agradecimentos

Este trabalho recebeu o apoio financeiro por parte da Fundação para a Ciência e a Tecnologia (FCT) no âmbito do projeto PTDC/MAT-STA/28248/2017 e da bolsa de doutoramento PD/BD/142887/2018.

Bibliografia

Allignol, A. e Latouche, A. (2019). CRAN Task View: Survival Analysis. Version 2019-09-01, URL http://CRAN.Rproject.org/view=Survival

Andersen, P. K., Borgan, O., Gill, R. D. e Keiding, N. (1993) *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

Anderson, G. L. e Fleming, T. R. (1995). Model misspecification in proportional hazards regression, *Biometria*, 82, 527-541.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34, 187-220.

Hastie, T. e Tibshirani, R. (1990). *Generalized additive models*, Chapman and Hall.

Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer, New York.

Martinussen, T. e Scheike, T. H. (2006). *Dynamic regression models for survival data*. Springer, New York.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C. e Andersen, P. K. (2009). Multi-state models for the analysis of time to event data. *Statistical Methods in Medical Research*, 18:195-222.

Meira-Machado, L. e Sestelo M. (2019). Estimation in the progressive illness-death model: a nonexhaustive review. *Biometrical Journal*, 61(2), 245-263.

Meira-Machado, L., Sestelo, M. e Soutinho, G. (2019). *survidm: survidm: Inference and Prediction in an Illness-Death Model*. R package version 1.2, URL https://CRAN.R-project.org/package=survidm

Soutinho, G., Meira-Machado, L. e Oliveira, P. (2019). Methods for checking the Markov condition in multi-state survival data. Proceeding of the International Workshop on Statistical Modelling (IWSM2019), 29-34.

Struthers, C. A. e Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2), 363-369.

Eilers P. H. C. e Marx B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11: 89-121.

Schumacher, M., Bastert, G., Bojar, H., Hiibner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R. L. A. e Rauschecker, H. F. for the German Breast Cancer Study Group (GBSG). (1994). A randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12, 2086-2093.

Gripe Pandémica 2009-2010: evolução espaço-temporal dos primeiros casos em Portugal

Isabel Natário*, *icn@fct.unl.pt*
M. Lucília Carvalho**, *mlcarvalho@gmail.com*

** Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa; CMA*
*** Faculdade de Ciências da Universidade de Lisboa; CEAUL*

1. Introdução

Durante o período da gripe pandémica de 2009-2010, o Instituto Nacional de Saúde Dr. Ricardo Jorge criou uma base de dados com todos os casos suspeitos de infeção por Gripe A (H1N1) participados, no contexto da rede de laboratórios de diagnóstico de infeção que coordenava. Esta base de dados inclui informação sobre as características dos pacientes e dos seus antecedentes epidemiológicos, sobre as estadias fora de casa dos doentes nos sete dias antes da ocorrência, sobre os correspondentes produtos biológicos que foram analisados, sobre a severidade da doença e sintomas associados. A análise destes dados pode ajudar a esclarecer os padrões de dispersão da doença, tanto no tempo como no espaço, evidenciando a importância da sinalização precoce de um surto deste tipo. Este estudo está assim confinado à chamada fase de contenção da epidemia (que é seguida da fase de mitigação, aqui não considerada) . Este período correspondeu aos dias entre 10 de Abril e 20 de Agosto de 2009, onde se registaram 2394 casos confirmados. A Organização Mundial de Saúde declarou que a pandemia de gripe A(H1N1) se encontrava no nível mais alto possível da fase de alerta no dia 29 de abril.

Detalham-se neste texto algumas ferramentas disponíveis e adequadas a um primeiro estudo de exploração e deteção de aglomerados espaço-temporais, não considerando a modelação. Apresentam-se desde medidas resumo simples a métodos de deteção de aglomerados, como a metodologia SaTScan^{TM} de Kulldorff [1] e variações [2], e funções-K espaço-temporais [3]. Utilizam-se estes recursos para tentar caracterizar e compreender os padrões de dispersão da doença nos primórdios da epidemia, conduzindo a análise por subgrupos populacionais definidos de acordo com fatores inerentes, tais como características demográficas, características clínicas dos pacientes, etc. [4].

2. Descrição dos dados

Analisam-se os dados que incluem todos os primeiros casos de gripe A do subtipo H1N1 confirmados laboratorialmente, ocorridos em Portugal, testados no período da fase de contenção da doença, de 10 de abril a 20 de agosto de 2009. Durante esta fase testavam-se todos os pacientes suspeitos de estarem doentes.

Esta base de dados foi coligida pelo Departamento de Doenças Infecciosas, Rede de Laboratórios para o Diagnóstico da Gripe para os Dados Portugueses da Gripe Pandémica de 2009, no Instituto Nacional de Saúde Dr. Ricardo Jorge. Dos 5991 pacientes na base de dados com sintomas de gripe, 2394 (40%) resultaram num teste positivo para a gripe A(H1N1).

Juntamente com a informação sobre a doença há outros dados, nomeadamente o género e a idade do paciente, a data do aparecimento dos sintomas, a escala de severidade dos sintomas do paciente, informação sobre estadias anteriores ou residência em áreas afetadas e datas de chegada a Portugal, informação sobre possíveis contactos com casos confirmados, localização do centro de saúde ou

hospital onde foi atendido, o resultado do teste para a gripe A(H1N1), o distrito e a freguesia de residência.

3. Aglomeração espaço-temporal

Na propagação no tempo e no espaço de uma doença contagiosa como a gripe, é frequente acontecer que alguns casos que ocorrem perto no tempo, porque foram infetados em simultâneo, não aconteçam necessariamente perto no espaço, devido à grande mobilidade dos infetados. Contudo, casos que ocorrem perto no espaço podem ter de facto uma tendência para ocorrer perto no tempo. Adicionalmente, apesar de frequentemente nos estádios iniciais de uma epidemia os efeitos de propagação se conseguirem ver muito claramente, isto pode já não ser verdade para os estádios mais avançados, porque quando a epidemia já está bem espalhada através de uma região os padrões de ocorrência de novos casos pode ser muito disperso.

Quando ocorrências observadas mais próximas no espaço também estão mais próximas no tempo, mais do que seria de esperar devido apenas ao acaso, fala-se de aglomeração espaço-temporal.

Neste estudo para se investigar a existência de aglomeração espaço-temporal de casos de gripe A(H1N1) confirmados em Portugal, os dados vão ser tratados como um padrão espacial pontual no tempo. Na análise consideram-se tanto medidas de associação espaço-temporal globais, como o teste de Mantel [5] e a função de Ripley [6], como locais como a dada pela estatística scan espaço-temporal de Kulldorff [7].

Medidas globais de associação

O teste de Mantel é uma versão melhorada do teste de Knox [8], baseado na medida de covariância de Mentel [5],

$$M = \sum_i \sum_j s_{ij} t_{ij} ,$$

onde s_{ij} representa a distância espacial entre dois casos, t_{ij} é a correspondente distância temporal e a soma é feita sobre todos os pares de casos.

Testa a independência entre as distâncias espaciais e temporais através de um teste de permutações que se faz usando o método Monte Carlo, onde as observações temporais são repetidamente baralhadas entre os locais espaciais e a estatística M calculada, gerando uma distribuição de referência sob a hipótese nula e permitindo o cálculo do valor-p do teste, posicionando o valor de M observado entre os valores gerados aleatoriamente.

A função-K espaço-temporal de Ripley [6] define-se como o número esperado de acontecimentos por unidade de área e de tempo,

$$K(s,t) = \frac{AT}{n^2} \sum_i \sum_j \frac{I_{s,t}(i,j)}{W_{ij}},$$

onde A é a área total em estudo, T é o período de tempo total, n é o número de observações, $I_{s,t}(i,j)$ é uma função indicatriz da distância espacial e temporal entre as observações i e j ser menor do que s e t, respetivamente, e W_{ij} é uma correção de efeito fronteira.

No caso de não ser esperada nenhuma relação entre espaço e tempo é razoável supor que $K(s,t) = K(s)K(t)$, onde $K(s)$ é o número esperado de acontecimentos por unidade de área e $K(t)$ é o número esperado de acontecimentos por unidade de tempo, dadas respetivamente por:

$$K(s) = \frac{A}{n} \sum_i \sum_j \frac{I_s(i,j)}{W_{ij}} \quad \text{e} \quad K(t) = \frac{T}{n} \sum_i \sum_j \frac{I_t(i,j)}{W_{ij}},$$

onde $I_s(i,j)$ é uma função indicatriz da distância espacial entre as observações i e j ser menor do que s e $I_t(i,j)$ é uma função indicatriz da distância temporal entre as observações i e j ser menor do que t. A estatística de teste resultante é então:

$$\widehat{D}(s,t) = K(s,t) - K(s)K(t).$$

Novamente se faz um teste de Monte Carlo, gerando um certo número de valores da estatística de teste K de Ripley sob a hipótese nula de não interação espaço-temporal, contra os quais o valor observado da mesma é classificado e o valor-p do teste calculado.

Medidas locais de associação

A estatística scan de permutação espaço-temporal [7] é uma generalização da estatística scan espacial de Kulldorf [1]. Este método usa uma janela cilíndrica em três dimensões onde a base do cilindro representa o espaço e a altura o tempo. Tal como com a estatística scan espacial, a estatística scan espaço-temporal considera uma base circular centrada, por exemplo, nos centroides de divisões administrativas da região, ou qualquer outra grelha de pontos cobrindo a área, para cada um dos intervalos de tempo disponíveis. Para cada centroide, faz-se variar o raio do círculo de zero até um valor máximo pré-definido, específico de cada aplicação.

Assim, escolhendo um número de pontos espaço-temporais, vários cilindros de tamanhos crescentes, tanto no espaço como no tempo, são considerados; então a estatística scan mede se o número de casos observados incluídos nestes cilindros é plausível para aquele volume, comparando com toda a área/período de tempo em estudo. Represente

$$C = \sum_c \sum_d c_{cd}$$

o número total de casos e c_{cd} o número de casos que aconteceram no centroide c e no período (e.g, dia) d. Condicional nos marginais observados, o número esperado de casos em (c, d) pode ser dado como a proporção de todos os casos que ocorreram no centroide c vezes o número total de casos que ocorreram no dia d,

$$\mu_{cd} = \left(\frac{\sum_c c_{cd}}{C}\right) \times \sum_d c_{cd}.$$

O número esperado de casos num dado cilindro A é dado então pela soma dos valores esperados atrás descritos sobre os centroides e sobre os dias que formam o cilindro,

$$\mu_A = \sum_{(c,d)\in A} \mu_{cd}.$$

Sob a hipótese nula de inexistência de interação espaço-temporal, a probabilidade de um caso estar no centroide c dado que foi observado no dia d é o mesmo todos os dias. Representando c_A o número de casos observados no cilindro A, condicional nas margens, sob a hipótese de não existência de interação espaço-temporal e assumindo que ambos $\sum_{c\in A} c_{cd}$ e $\sum_{d\in A} c_{cd}$ são pequenos quando comparados com C, c_A segue aproximadamente uma distribuição Poisson com parâmetro μ_A. Baseado nesta aproximação, o método considera a razão de verosimilhança generalizada de Poisson para avaliar se o cilindro A contém um número pouco normal de casos:

$$\left(\frac{c_A}{\mu_A}\right)^{c_A} \left(\frac{C - c_A}{C - \mu_A}\right)^{C - c_A}.$$

Dos cilindros considerados, aquele com a maior razão de verosimilhança generalizada de Poisson é aquele com menor chance de ser um aglomerado apenas devido ao acaso e é tomado como o principal candidato a um surto real.

Porque a população em risco não está disponível, tem de se fazer um teste de simulação Monte Carlo, onde as datas e os instantes são baralhados e reatribuídos ao conjunto original das localizações, garantindo que não há alterações nas margens espaciais e temporais. Então o aglomerado mais provável é determinado da mesma forma que para os dados reais. A significância estatística pode ser avaliada pela posição do aglomerado observado entre a distribuição estimada dos aglomerados simulados.

4. Análise dos dados referentes aos primeiros casos de gripe A(H1N1) em Portugal

Na Figura 1 representa-se a cinzento o número total diário de novos casos, a preto o número de casos diários correspondentes a pessoas que estiveram no estrangeiro numa área afetada antes da infeção (infeção primária), e a amarelo mostarda o número de casos diários que contactaram com um caso confirmadamente infetado. Para estas duas últimas quantidades há apenas informação para 1193 casos de 2394, cerca de 50% do número total de casos. Os primeiros casos confirmados vieram essencialmente de áreas afetadas no estrangeiro, mas em Agosto este efeito torna-se comparativamente menos importante na dispersão da epidemia, dado o grande aumento do número de casos. É interessante ver que o número de casos que se sabe ter estado em contacto com uma pessoa infetada nunca é muito elevado, sugerindo que essa informação é frequentemente desconhecida.

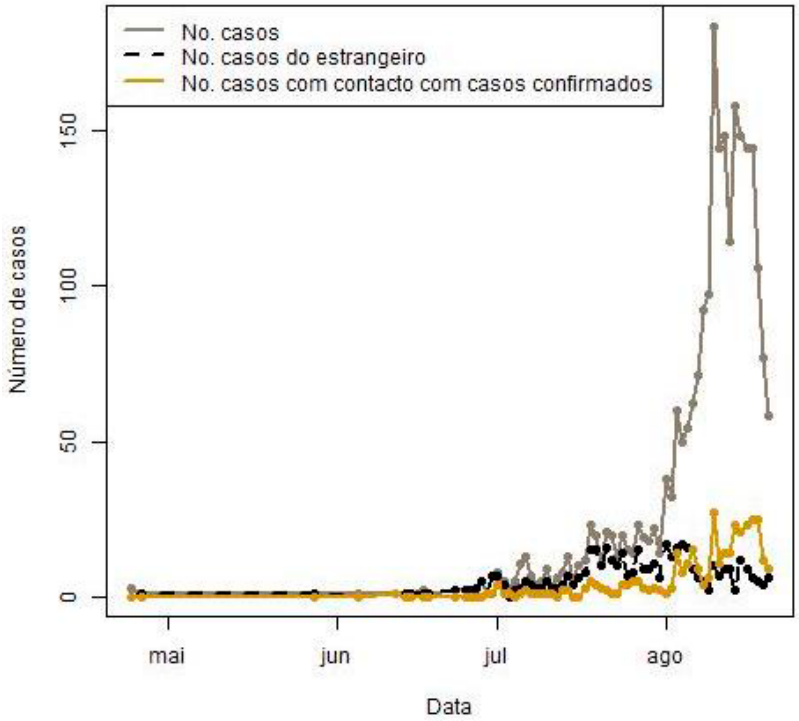


Figura 1: Número diário de novos casos (cinzento), número de casos que se sabe terem estado na área infetada antes da infeção (preto) e número de casos que se sabe terem contactado com casos infetados (mostarda).

Calcularam-se medidas resumo para 5991 pacientes da base de dados (**TODOS**) e apenas para o grupo dos 2394 pacientes que se confirmaram como sendo casos de gripe A(H1N1) (**A(H1N1)**), apresentadas na Tabela 1. Uma coluna adicional resume os dados para os pacientes cujo teste deu positivo em Portugal Continental (**A(H1N1), PTC**) – uma vez que não estando ligados por terra nem próximos os Arquipélagos dos Açores e da Madeira de Portugal Continental, as viagens para esses Arquipélagos são essencialmente feitas por avião e possivelmente não tão frequentemente como dentro de Portugal Continental.

Os aspetos que mais se destacam dos dados analisados são que os mais afetados são homens jovens, com idades entre 20 e 30 anos, das áreas mais densamente populadas. O fenómeno das férias é muito claramente percebido em julho, com uma grande percentagem de casos sendo residentes do Reino Unido, conhecidos por terem em Portugal um destino de férias preferido, e em agosto quando a maioria dos casos atendidos nos centros de saúde e hospital são de um dos destinos de praia favoritos dos Portugueses (Algarve), apesar de não serem residentes na área. Em relação a ter viajado para o estrangeiro nos dias que antecederam o aparecimento de sintomas, não se consegue estabelecer uma evidência clara da sua importância, apesar de para os primeiros casos registados tal parecer bastante relevante.

A média de idades dos casos confirmados é de 23.6 anos. Comparando por género, a maioria dos casos confirmados correspondem a homens.

Variável	Medidas Resumo	TODOS	A(H1N1)	A(H1N1), PTC
Sexo	Percentagem homens Percentagem mulheres Percentagem NAs	53.6 45.8 0.6	57.4 42.2 0.4	57.1 42.5 0.4
Idade	Média Desvio padrão Percentagem de NAs	23.6 16.4 1.7	22.3 12.5 1.6	22.2 12.5 1.6
Data primeiros sintomas	Mínimo Máximo Mediana	10/04/2009 20/08/2019 11/08/2009	24/04/2009 20/08/2009 11/08/2009	24/04/2009 20/08/2009 11/08/2009
Severidade dos sintomas	% sem sintomas % sintomas fracos % sintomas moderados % sintomas severos % mortes % NAs	2.9 59.9 8.7 0.7 0.02 27.8	2.6 64.8 8.1 0.5 0.0 24.0	2.7 64.1 8.3 0.4 0.0 24.5
Deslocação a zona afetada	% sim % não % NAs	19.8 17.0 63.2	18.3 18.6 63.1	23.8 27.1 49.2
Data de chegada, se sim na questão anterior	Min Max Mediana % Nas	20/04/2009 30/08/2009 24/07/2009 12.7	26/04/2009 21/08/2009 26/07/2009 22.2	26/04/2009 21/08/2009 26/07/2009 21.0
Contactou caso suspeito ou confirmado nos 7 dias anteriores?	% sim % não % NAs	9.8 14.0 76.1	14.3 10.3 75.4	19.6 12.6 67.8
Local de contacto	% Algarve % Lisboa % Campismo Ericeira % Acampamento de Verão	0.7 0.4 0.3 0.1	1.0 0.7 0.5 0.5	1.1 0.5 0.4 0.6
Distrito de residência ou país se outro que não Portugal	% Lisboa % Porto % Açores % Braga % Coimbra % França % Setúbal % Faro % Reino Unido % NAs	10.6 10.3 3.3 2.6 2.3 2.2 2.5 1.8 1.2 49.4	11.9 9.7 3.0 2.8 1.5 0.6 2.8 2.7 1.4 51.6	29.5 24.1 - 7.1 3.7 - 7.0 6.6 - 0
Localização de atendimento dos cuidados de saúde	% Algarve % Porto % Lisboa	33.4 20.4 11.1	36.2 18.7 14.0	5.2 36.7 20.9

Tabela 1: Medidas resumo para todos os pacientes que foram testados para a gripe (**TODOS**), para o grupo dos pacientes que se confirmaram como sendo casos (**A(H1N1)**) e para os pacientes que se confirmaram como sendo casos em Portugal Continental (**A(H1N1), PTC**).

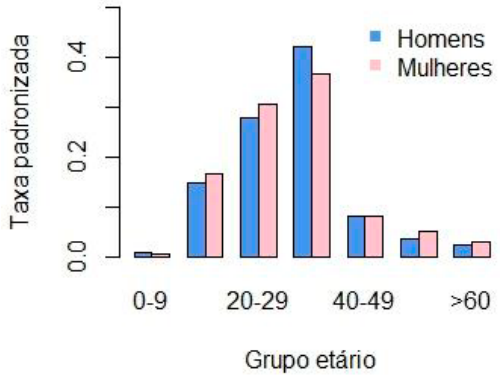


Figura 2: Taxa padronizada de casos confirmados por idade e por sexo.

A Figura 2 apresenta um gráfico da taxa padronizada de casos confirmados por idade e por género, onde se consegue visualizar que as taxas mais elevadas correspondem a jovens adultos e as taxas mais baixas a crianças pequenas. Estas taxas foram calculadas como o número de casos por faixa etária e por sexo dividido pela correspondente população em risco, e ainda dividindo cada uma destas taxas pela soma de todas elas. Padronizando apenas por género resulta numa taxa de 0.60 para homens e 0.40 para mulheres.

Para 83.8% dos casos para os quais há informação sobre sintomas, estes foram essencialmente fracos. Relativamente aos casos, há uma grande proporção de falta de informação sobre a estadia dos pacientes no estrangeiro em áreas infetadas mas, para os que há essa informação, cerca de metade estiveram nesse tipo de áreas. Os mesmos comentários se aplicam à informação sobre se os casos contactaram com outros casos confirmados, com um pouco mais de casos tendo contactado, essencialmente no Algarve e em Lisboa. Quanto ao local de residência, apesar de cerca de metade dos casos não ter esta informação associada, a maioria dos outros casos residiam em Lisboa e no Porto. Em relação ao centro de saúde/hospital em que foram atendidos, a maioria localizava-se no Algarve, Porto e Lisboa.

A Figura 3 apresenta mapas do número de casos confirmados por freguesia de residência em três pontos distintos do tempo: 30 de julho de 2009, 10 de agosto de 2009 e o último dia da fase de contenção, 20 de agosto de 2009, de onde se consegue visualizar bem a dispersão do número de casos das regiões em torno das principais cidades de Portugal Continental, Lisboa, Porto e Algarve, para o resto do país.

No painel da esquerda da Figura 4 representa-se o número de novos casos diários por residência dos pacientes e no painel direito o número de novos casos diários por localização do centro de saúde/hospital onde o paciente foi seguido. Pode-se constatar que o número de casos aumentou mais expressivamente no final de julho, começo de agosto, e que houve mais casos nas áreas mais populadas e no Algarve (período de férias).

Figura 3: Mapas do número de casos confirmados de gripe A(H1N1), por freguesia de residência, em 30 de julho de 2009 (esquerda), 10 de agosto de 2009 (meio) e o último dia da fase de contenção, 20 de agosto de 2009 (direita).

Conduziram-se então os testes de aglomeração globais de Mantel e função-K de Ripley, apenas para o conjunto de casos confirmados e residentes em Portugal Continental.

A correlação observada de Mantel é de $M_{obs} = 0.012$ com um intervalo de confiança bootstrap a 95% de (0.004;0.019). O teste de permutação aceita a hipótese de que a correlação é nula (valor-p=0.83), indicando que não há associação global entre o espaço e o tempo para estes dados. Todos os cálculos foram feitos usando o pacote do R `ecodist` [9].

A aplicação da função-K de Ripley espaço-temporal resultou igualmente na aceitação da inexistência de relação espaço-temporal global (valor-p=0.954), confirmando o resultado anterior. Os cálculos foram feitos usando os pacotes do R `spatstat`[10] e `splancs`[11].

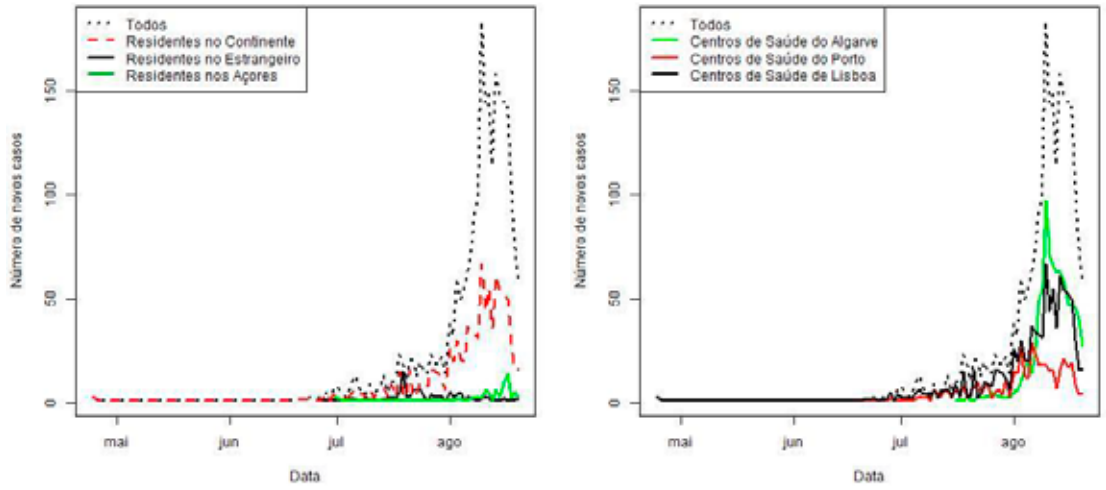


Figura 4: Número de novos casos diários por local de residência (esquerda) e número de novos casos diários por local de atendimento no centro de saúde/hospital (direita).

Seguidamente aplicou-se o teste da estatística scan espaço-temporal de Kulldorf à base de dados dos casos confirmados residentes em Portugal Continental. Tal resultou na detação de cinco aglomerados, três dos quais se encontram representados na Figura 5. O quarto aglomerado corresponde essencialmente a residentes de três freguesias no distrito de Leiria a 15 de agosto e o quinto aglomerado corresponde a residentes de apenas uma freguesia do distrito do Porto a 28 de julho. Esta análise foi feita usando o software SaTScan™, http://www.ststscan.org/.

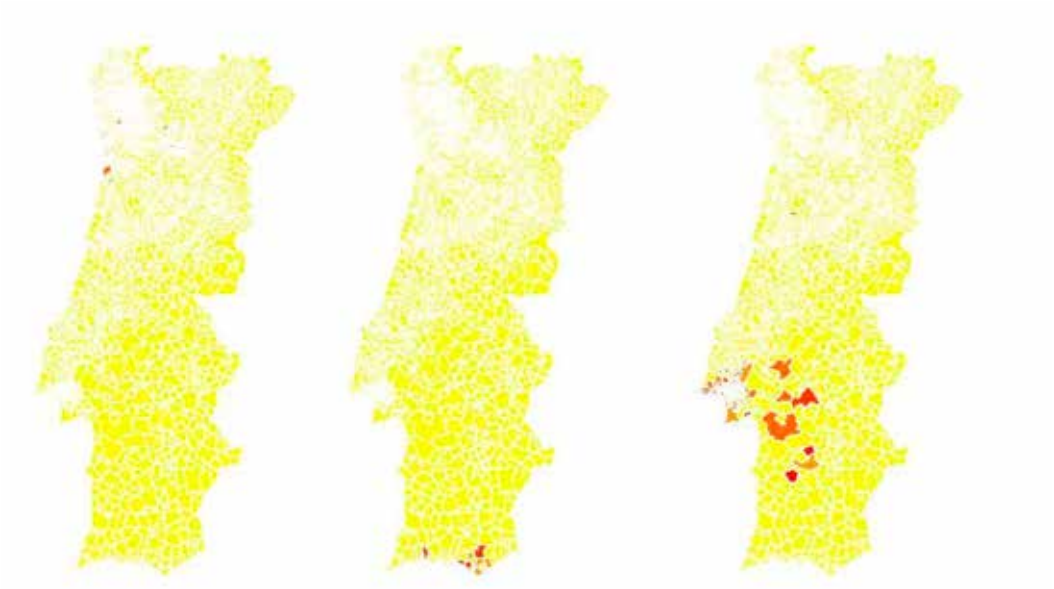


Figura 5: Mapas dos três aglomerados detetados: 29/07/2009 a 07/08/2009, Norte (esquerda), 25/07/2009 a 30/07/2009, Sul - Algarve (meio) e 11/08/2009 a 20/08/2009, Lisboa e Alentejo (direita).

5. Conclusão

Assim se conclui esta análise preliminar dos primeiros casos da gripe pandémica A(H1N1) de 2009-2010 em Portugal, relativamente à sua fase de contenção. Os principais resultados apontam para que os mais afetados tenham sido homens jovens, com idades entre 20 e 30 anos, das áreas mais densamente populadas, o que está em linha com o que aconteceu noutros lados [4]. O fenómeno das férias é bem percebido no Algarve em julho com residentes do Reino Unido e em agosto com os Portugueses, que em comum elegem esta região do país para as suas férias balneares. Estas conclusões constituem uma importante base para a modelação da heterogeneidade através de modelos que incluam, por exemplo, covariáveis e expliquem que fatores são os relevantes para essa heterogeneidade – diferente distribuição populacional, diferentes níveis de exposição à contaminação pela mobilidade, número de

contactos, etc.. Outros métodos não tradicionais de modelação, como modelos de redes para relacionar padrões de viagens poderão ser adequados para este tipo de dados [12].

Agradecimentos
Os autores agradecem ao Instituto Nacional de Saúde Dr. Ricardo Jorge, Departamento de Doenças Infecciosas, Rede de Laboratórios para o Diagnóstico da Gripe, base de dados Mercúrio, pelos dados Portugueses da Gripe Pandémica, nas pessoas do Professor José Calheiros e a Doutora Raquel Guiomar, e ao Doutor Baltazar Nunes ter estabelecido os contactos necessários.

Este trabalho foi parcialmente suportado por fundos nacionais através da Fundação para a Ciência e Tecnologia, Portugal, projeto UID/MAT/00297/2019.

Referências

[1] Kulldorff M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**, 1481-1496.
[2] Takahashi, K.; Kulldorff, M.; Tango, T.; Yih, K. (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. International Journal of Health Geographics, **7**, 1-14.
[3] Xiao H.; Lin X.; Chowell G.; Huang H.; Gao L.; Chen B.; Wang Z.; Zhou L.; He X.; Liu H.; Zhang X.; Yang H. (2014). Urban structure and the risk of influenza A (H1N1) outbreaks in municipal districts. *Chinese Science Bulletin*, **59**, 554-562.
[4] Roll U.; Yaari, R.; Katriel G.; Barnea O.; Stone L.; Mendelson E.; Mandelboim M.; Huppert A. (2011). Onset of a pandemic: characterizing the initial phase of the swine flu (H1N1) epidemic in Israel. *BMC Infectious Diseases*, **11**, 92.
[5] Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.
[6] Ripley, B.D. (1981). *Spatial Statistics*. Wiley, New-York.
[7] Kulldorff M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. Journal of Royal Statistical Society A, **164**, 61-72.
[8] Knox, G. (1964). The detection of space-time interactions. *Applied Statistics*, **13**, 25-29.
[9] Goslee S.C.; Urban D.L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1-19.
[10] Baddeley A.; Turner R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, **12**, 1-42.
[11] Bivand, R.; Rowlingstone, B.; Diggle P.; Petris G.; Eglen S. (2017). Package splancs, url: https://cran.r-project.org/web/packages/splancs/splancs.pdf.
[12] Newman, M.E.J. (2010). *Networks: An Introduction*, Oxford University Press, Oxford.

Ambiente, Saúde e Estatística

Ana Luísa Papoila**,** *ana.papoila@nms.unl.pt*

NOVA Medical School/Faculdade de Ciências Médicas,
Universidade Nova de Lisboa e CEAUL

1. Introdução

O papel da Estatística no Mundo actual tem vindo a destacar-se em várias áreas do saber das quais saliento a área da Saúde por ser esta a minha área de eleição no que diz respeito à investigação desde há muitos anos atrás. De facto, o aumento da utilização de resultados obtidos por métodos quantitativos na estimação do risco inerente a determinadas exposições e na tomada de decisão clínica/política, favorece a aplicação das metodologias Estatísticas de uma forma mais intensa na área da Saúde. A Epidemiologia Ambiental não é uma excepção, salientando-se o desenvolvimento de vários métodos estatísticos, e de *software* apropriado para os implementar, que permitem estimar o risco em saúde associado às condições atmosféricas e à exposição a determinados agentes ambientais como, por exemplo, os pólens e os poluentes, habitualmente quantificados em estudos sobre a qualidade do ar.

No que diz respeito às condições atmosféricas, é exemplo de grande utilidade para a Saúde Pública toda a investigação em torno do impacto das temperaturas extremas do ambiente sobre a saúde humana, nomeadamente o desenvolvimento de novas metodologias estatísticas neste contexto. Vários estudos têm vindo a ser publicados sobre o impacto das ondas de calor e das alterações climatéricas na saúde humana (Gosling *et al*., 2007; Gosling *et al*., 2009; Li *et al*., 2015; Wang et al., 2015; Isaksen *et al*., 2016) e sobre as epidemias de gripe durante o Inverno (Brooks *et al*., 2015; da Costa *et al*., 2018; Hosseini *et al*., 2018; Zhang *et al*., 2019). As próprias Organização Meteorológica Mundial e Organização Mundial de Saúde propuseram algumas orientações sobre Sistemas de Alerta relacionados com as ondas de calor (WMO e WHO, 2015) e, em Portugal, tem sido o Instituto Nacional de Saúde Dr. Ricardo Jorge a liderar este assunto. Efectivamente, no que diz respeito aos excessos de temperatura, é implementado sazonalmente o Sistema de Vigilância ICARO (**I**mportância do **CA**lor: **R**epercussões sobre os **Ó**bitos), em funcionamento desde 1999, com o objectivo de vigiar e monitorizar as ondas de calor. Também respeitante à época invernal, várias contribuições metodológicas na área da Estatística foram propostas por esta Instituição, nomeadamente no que diz respeito à estimação do excesso de mortalidade associado às epidemias de gripe (Nunes *et al*., 2011) e à previsão em tempo real da evolução destas epidemias (Nunes *et al*., 2013).

Neste âmbito, as abordagens estatísticas utilizadas, bayesianas e frequencistas, envolvem, maioritariamente, séries temporais e cadeias de Markov, modelos aditivos generalizados, habitualmente com desfasamentos (*lags*), e considerando a distribuição de Poisson para a variável resposta, e ainda modelos espaço-temporais.

Por sua vez, a poluição ambiental é um tema actual e representa uma séria preocupação a nível mundial devido às consequências nefastas em várias vertentes. Sérias alterações climatéricas, problemas de saúde e até mortalidade são alguns exemplos graves do impacto da má qualidade do ar sobre o planeta e sobre os seus habitantes em que nós nos incluímos. A preocupação é a nível mundial e várias iniciativas têm vindo a ser tomadas. Assim, já em 1980, nos Estados Unidos da América, foi fundado um Instituto para estudar os efeitos da poluição na saúde (Health Effects Institute). Este Instituto é financiado pelo Governo e pela Indústria e tem vários projectos de investigação na América do Norte, América Latina, Europa e Ásia. Estes projectos incidem sobre vários tipos de poluentes como, por exemplo, o dióxido de carbono, dióxido de nitrogénio, dióxido de enxofre, monóxido de

carbono, ozono e matéria particulada. Mais recentemente, em 2018, a Organização Mundial de Saúde (OMS) realizou a primeira conferência sobre Poluição e Saúde em Geneva (Global Conference on Air Polution and Health). Na minha opinião peca pela hora tardia mas, como é habitual dizermos, "antes tarde que nunca"!!!. De facto, várias análises foram efectuadas no contexto desta conferência e conclui-se que os níveis de poluição do ar ambiente são demasiado altos numa grande parte do mundo e que nove de dez pessoas respiram ar altamente poluído. A OMS estima que sete milhões de pessoas morram todos os anos devido à exposição a partículas muito finas que se encontram em supensão no ar poluído, provocando doenças como o cancro de pulmão, doenças pulmonares obstrutivas crónicas e infecções respiratórias incluindo a pneumonia, doenças cardio-vasculares e acidentes vasculares cerebrais, entre outras. Sem dúvida um panorama assustador e naturalmente desencadeador de atitudes que permitem não só minimizar a poluição e seus efeitos mas também compreender todas estas associações. A Estatística tem seguramente um papel importante neste contexto e, muito recentemente, já começam a surgir estudos de investigação em que o principal objectivo é compreender o futuro impacto sobre a saúde das alterações climatéricas (Vicedo-Cabrera *et al*., 2019).

Pelo já referido, quando se fala em ambiente, saúde e estatística, é patente a heterogeneidade de temas dificilmente abordáveis no contexto do presente texto. Assim sendo, considerando apenas o par poluição/saúde, seguir-se-ão alguns conceitos básicos no que diz respeito ao desenho dos estudos utilizados nesta área e serão ainda apresentadas as principais metodologias estatísticas envolvidas neste âmbito, nomeadamente para dados de séries temporais. A ideia é munir o leitor da informação necessária ao planeamento e análise dos dados resultantes de um projecto de investigação cujo objectivo seja estimar a influência da qualidade do ar na saúde humana.

Como base de apoio, foi utilizado um livro exclusivamente sobre métodos estatísticos aplicados à Epidemiologia Ambiental cujos autores constituem uma referência nesta área de investigação (Dominici & Peng, 2008).

2. Desenhos de estudo

O estudo pode envolver séries temporais em que se pretende associar a exposição à poluição com o número de eventos de interesse ao longo do tempo (Health Effects Institute, 2003, Bell *et al*., 2004). Estes estudos podem ser classificados de ecológicos, na medida em que analisam resultados de saúde diários médios populacionais e a sua associação com os níveis de exposição. Os modelos estatísticos mais comumente utilizados são os modelos lineares generalizados com *splines* paramétricos (como, por exemplo, os *splines* cúbicos naturais) e os modelos aditivos generalizados com suavizadores não paramétricos (como, por exemplo, *splines* suavizadores ou os suavizadores *lowess*). Quando faça sentido, é ainda habitual considerar nesta análise as flutuações na mortalidade que poderão ser consideradas factores de confundimento na estimação do efeito da poluição.

Outro tipo de desenho é o *case-crossover*. Este tipo de estudo inclui apenas indivíduos afectados pelo efeito em estudo e compara a probabilidade da sua ocorrência quando sujeitos à exposição em investigação durante um período considerado como período de risco (*Time Case Period* ou *Index Time*) com a probabilidade de ocorrência fora do período de risco (*Time Control Period* ou *Referent Times*). Assim, os períodos de tempo em que o efeito não se manifestou constituem períodos controlo permitindo diminuir não só o viés de selecção mas também o número de variáveis de confundimento dado que os casos são controlos de si mesmos. A ausência de associação entre a ocorrência do efeito e a exposição, significa que a proporção de casos ocorridos dentro e fora do período de risco é semelhante. Pela sua própria essência, os estudos *case-crossover* são os mais apropriados para investigar associações entre exposições intermitentes e consequente ocorrência imediata de efeitos transitórios, comparando o número de casos ocorridos dentro e fora do período de risco. Assim sendo, no presente contexto, este desenho foi originalmente proposto por Maclure (1991, 2000) dado ser esta uma situação que surge frequentemente nos estudos sobre o efeito da poluição na saúde (Jaakkola, 2003). Dado o emparelhamento implícito dos dados resultantes deste desenho, a modelação deverá contemplar esse facto através da aplicação de, por exemplo, modelos simples de regressão logística condicional (Fisher *et al*., 2019) e modelos lineares/aditivos generalizados de efeitos mistos, mais abrangentes e também eles apropriados para esta situação (Figueiras *et al.*, 2005).

Finalmente, temos os estudos longitudinais de painel e de coorte. Embora surjam na literatura como sendo o mesmo tipo de desenho com alguma frequência, existem pequenas diferenças que os distinguem. Assim, num estudo de coorte, é habitual considerar um grupo de pessoas que partilham

uma determinada característica (e.g. expostas a determinado fármaco ou poluente, ou submetidas a determinado tratamento médico) e que são seguidas ao longo do tempo, habitualmente períodos longos para avaliar a mortalidade ou outros eventos relacionados com a saúde. No presente contexto, uma medida dos níveis acumulados da poluição ambiente ao longo do tempo é considerada a exposição pelo que, estudos de coorte associam exposições a longo prazo com eventos de saúde (Pope, 2007). No que diz respeito a estudos de painel, estes são também estudos longitudinais que seguem os mesmos indivíduos de determinada amostra ao longo do tempo sem, no entanto, terem que partilhar uma mesma característica. No âmbito de estudos ambiente/saúde, este tipo de desenho é eficaz para estudar os efeitos a curto prazo dos poluentes do ar na saúde.
No que diz respeito à escolha do desenho de estudo, este dependerá principalmente do tipo de relação que se pensa existir entre a exposição e o evento de interesse. Obviamente, serão necessárias abordagens próprias para situações distintas no que diz respeito, por exemplo, ao tempo acumulado de exposição (e.g. exposição a curto ou a longo prazo) ou mesmo ao tipo de efeito (e.g. efeitos agudos ou crónicos). Assim, efeitos agudos podem ser estimados a partir de estudos de painel, séries temporais e *case-crossover*, e efeitos crónicos situam-se mais no âmbito de estudos de coorte. No entanto, dado que este tipo de estudos se prolongam habitualmente por longos períodos de tempo, também é possível utilizá-los em situações em que se pretende quantificar as consequências sobre a saúde de picos de exposição atingidos em curtos espaços de tempo.
Para maior detalhe sobre este tópico, consultar Dominici & Peng, 2008.

3. Modelação dos dados

Na maioria dos casos, quando utilizamos modelos de regressão na análise de dados, temos em mente dois possíveis cenários, um deles tem a ver com predição, o outro tem a ver com a estimação da força de associação entre um determinado número de variáveis e uma variável resposta. Ao analisarmos dados sobre poluição e saúde, encontramo-nos maioritariamente na segunda destas situações. De facto, neste âmbito, pretende-se estimar e compreender a associação entre uma determinada exposição ambiental e as consequências sobre a saúde resultantes desta exposição, após ajustar por outras variáveis relevantes onde se incluem potenciais variáveis de confundimento, habitualmente presentes neste tipo de estudos (e.g. temperatura, humidade e estação do ano).
Como já referido, dado ser este um tema muito abrangente, vamos abordar apenas metodologias estatísticas apropriadas para analisar dados correspondentes a uma exposição a poluentes ambientais X_t e correspondente efeito na saúde, avaliado, por exemplo, pelo número de eventos de interesse ocorridos em determinado instante/período de tempo Y_t (e.g. número diário de internamentos por problemas respiratórios). Assim sendo, considerando $Y_t \sim Poisson(\mu_t), (t = 1, \dots, n)$ e o desfasamento l, um dos modelos clássicos vem dado pela seguinte expressão,

$$\log \mu_t = \alpha + \beta X_{t-l} + \eta \boldsymbol{Z}_t + s(t; \lambda),$$

em que, α, β e η representam os coeficientes de regressão, correspondendo β ao logaritmo do risco relativo associado à exposição X no instante $t - l$, $\boldsymbol{Z}_t$ um vector de covariáveis no instante t e o termo $s(t; \lambda)$ representa uma função suavizadora do tempo t controlada pelo parâmetro λ (parâmetro suavizador). Este termo é muito importante dado que permitirá ajustar as estimativas, de uma forma indirecta, a potenciais variáveis de confundimento que variam com o tempo e que, por algum motivo, não foram medidas. Este modelo não é mais do que um modelo aditivo generalizado (Hastie e Tibshirani, 1990) em que a variável resposta tem uma distribuição de Poisson, podendo ainda ter uma maior flexibilidade caso permitamos que também Z_t tenha uma associação não linear com a resposta

$$\log \mu_t = \alpha + \beta X_{t-l} + s_1(Z_t;\ \lambda_1) + s_2(t;\ \lambda_2). \qquad (1)$$

Neste primeiro modelo assume-se que o efeito da exposição em estudo ocorre apenas num único dia, determinado pelo desfasamento l. Assim sendo, e se considerarmos como exemplo de variável resposta o número de óbitos devidos à exposição a determinado poluente, o ajustamento do modelo conduzirá a um resultado que nos permitirá retirar conclusões àcerca da associação entre os níveis do poluente no dia t e a mortalidade l dias mais tarde. Este exemplo será um caso claramente apropriado para utilizar o modelo (1) dado se tratar de um caso com dados de mortalidade. No entanto, se

considerarmos estudos em que se pretende estudar a associação entre um poluente e o número de idas à urgência hospitalar por problemas respiratórios, é natural que esta associação se estenda por mais tempo do que apenas um dia.
Assim, novos modelos foram propostos em que este facto pode ser considerado através da inclusão de múltiplos desfasamentos simultaneamente. Estes modelos designam-se por modelos de desfasamentos distribuídos (*distributed lag models*) pelo facto de permitirem que o efeito de uma determinada exposição seja distribuído por um período específico, usando vários parâmetros para explicar as contribuições nos diferentes desfasamentos. Foram originalmente propostos para dados de séries temporais em econometria por Almon (1965) e posteriormente em epidemiologia por Schwartz (2000).
De uma forma genérica, um modelo de desfasamentos distribuídos de ordem *K* é definido pela seguinte expressão

$$\log \mu_t = \alpha + \sum_{l=0}^{K} \beta_l X_{t-l} + \eta \boldsymbol{Z}_t,$$

em que *K* representa o desfasamento máximo. Estes modelos admitem que o impacto global num instante *t* provocado pelo aumento de uma unidade da exposição distribui-se, mais tarde, ao longo de *K* dias. Assim sendo, o valor acumulado deste efeito vem dado por $\sum_{l=0}^{K} \beta_l$. Devido a problemas óbvios de colinearidade, foi necessário introduzir restrições sobre os coeficientes de regressão e alguns desenvolvimentos surgiram em torno deste assunto (para mais detalhes, consultar Zanobetti *et al*., 2000).
De novo, para a obtenção de modelos mais flexíveis, a evolução para a introdução dos modelos aditivos generalizados neste contexto foi inevitável, tendo-se obtido o seguinte modelo

$$g\{E(Y_t)\} = \alpha + \boldsymbol{\gamma^T Z}_t + \sum_{j=1}^{m} f_j(S_{jt}) + \sum_{l=0}^{K} \beta_l X_{t-l},$$

em que *g* representa uma função de ligação, $\boldsymbol{Z}_t$é um vector de variáveis modeladas de forma linear e S_{jt} representa a *j*-ésima variável modelada através de um suavizador f_j (Zanobetti *et al*., 2000). Outros autores se seguiram e, ainda que noutros contextos, apresentaram propostas numa vertente frequencista (Rushworth *et al*., 2013; Obermeier *et al*., 2015) tendo Welty *et al*., já em 2009, proposto uma abordagem bayesiana para este tipo de modelos.
Embora já com uma muito boa flexibilidade, estes modelos não permitem uma associação "dose-resposta" não linear. Assim, surgiram novas propostas e Armstrong (2006) e Gasparrini *et al*. (2010) propõem modelos de desfasamentos distribuídos não lineares, enquanto que, mais tarde, Gasparrini (2014) generaliza estes modelos, já não lineares, a outros desenhos e estruturas de dados que não séries temporais. De facto, além de muitas aplicações práticas, este autor tem vindo a desenvolver ao longo dos anos novas metodologias estatísticas com vista ao estudo da associação entre exposições ambientais e saúde (Gasparrini *et al*., 2017) e é uma referência nesta área. É ainda o autor do pacote do R que permite a implementação destes modelos (Gasparrini, 2011).
Mais recentemente, novos modelos que permitem a modelação do efeito de vários poluentes em simultâneo foram propostos por Chen *et al*. (2019) e é patente a grande evolução da estatística nesta área com um tão grande número de novos desenvolvimentos num espaço de dez anos. Ciente de que muito ficou por abordar, espero que a informação contida neste texto seja útil a quem deseje iniciar a sua investigação nesta área tão interessante e útil do ponto de vista da Saúde Pública.

Agradecimentos

A investigação que deu origem ao presente documento foi suportada por fundos nacionais através da Fundação Nacional para a Ciência e Tecnologia, Portugal-FCT no âmbito do projecto UID/MAT/UI0006/2019.

Referências Bibliográficas

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, **33**, 178–196.

Armstrong, B. (2006). Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, **17**, 624–631.

Bell, M. L., Samet, J. M. & Dominici, F. (2004). Time-series studies of particulate matter. *Annual Review of Public Health*, **25**, 247–280.

Brooks, L.C., Farrow, D.C., Hyun, S., Tibshirani, R.J. & Rosenfeld, R. (2015). Flexible modeling of epidemics with an empirical Bayes framework. *PLoS computational biology*, **28**;11(8): e1004382.

Chen, Yin-Hsiu, Mukherjee, B. and Berrocal, V.J. (2019). Distributed lag interaction models with two pollutants. *Journal of the Royal Statistical Society, Applied Statistics, Series C*, 68, Part 1, 79–97.

da Costa, A., Codeço, C. T., Krainski, E. T., Gomes, M., & Nobre, A. A. (2018). Spatiotemporal diffusion of influenza A (H1N1): Starting point and risk factors. *PloS one*, **13**(9), e0202832.

Dominici, F. & Peng, R.D. (2008). *Statistical Methods for Environmental Epidemiology with R*. Springer, New York, NY, USA.

Figueiras, A., Carracedo-Martínez, E., Saez, M. & Taracido, M. (2005). Analysis of Case-Crossover Designs Using Longitudinal Approaches: A Simulation Study. *Epidemiology*, **16** (2) 239–246.

Fisher, J.A., Puett, R.C., Laden, F., Wellenius, G.A., Sapkota, A., Liao, D., Yanosky, J.D., Carter-Pokras, O., He, X. & Hart, J.E. (2019). Case-crossover analysis of short-term particulate matter exposures and stroke in the health professionals follow-up study. *Environment International*, **124**, 153–160.

Gasparrini, A., Armstrong, B., & Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, **29**, 2224–2234.

Gasparrini, A. (2011). Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*, **43**, 1–20.

Gasparrini, A. & Leone, M. (2014). Attributable risk from distributed lag models. *BMC Medical Research Methodology*, **14**, 55.

Gasparrini, A., Scheipl, F., Armstrong, B. & Kenward, M.G. (2017). A penalized framework for distributed lag non-linear models. *Biometrics*, **73**, 938–948.

Gosling, S.N., McGregor, G.R. & Paldy, A. (2007). Climate change and heat-related mortality in six cities part 1: model construction and validation. *International Journal of Biometeorology*, **51** (6), 525–540.

Gosling, S.N., McGregor, G.R. & Lowe, J.A. (2009). Climate change and heat-related mortality in six cities Part 2: climate model evaluation and projected impacts from changes in the mean and variability of temperature with climate change. *International Journal of Biometeorology*, **53** (1), 31–51.

Hastie, T.J. & Tibshirani, R. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Health Effects Institute (2003). *Revised Analyses of Time-Series Studies of Air Pollution and Health. Special Report*. Health Effects Institute, Boston MA.

Hosseini, S., Karami, M., Farhadian, M. & Mohammadi, Y. (2018). Seasonal Activity of Influenza in Iran: Application of Influenza-like Illness Data from Sentinel Sites of Healthcare Centers during 2010 to 2015. *Journal of Epidemiology and Global Health*, **8** (1-2), 29–33.

Isaksen, T.B., Fenske, R.A., Hom, E.K., Ren, Y., Lyons, H. & Yost, M.G. (2016). Increased mortality associated with extreme-heat exposure in King County, Washington, 1980–2010. *International Journal of Biometeorology*, **60** (1), 85–98.

Jaakkola, J. J. K. (2003). Case-crossover design in air pollution epidemiology. *European Respiratory Journal Suppl*, **40**, 81s-85s.

Li, M., Gu, S., Bi, P., Yang, J. & Liu, Q. (2015). Heat Waves and Morbidity: Current Knowledge and Further Direction-A Comprehensive Literature Review. *International Journal of Environmental Research and Public Health*, **12**, 5256–5283.

Maclure, M. (1991). The case-crossover design: A method for studying transient effects on the risk of acute events. *American Journal of Epidemiology*, **133**, 144–153.

Maclure, M. & Mittleman, M. A. (2000). Should we use a case-crossover design? *Annual Review of Public Health*, **21**, 193–221.

Nunes, B., Natário, I., & Carvalho, M. L. (2011). Time series methods for obtaining excess mortality attributable to influenza epidemics. *Statistical Methods in Medical Research*, **20**(4), 331–345.

Nunes, B., Natário, I., & Lucília Carvalho, M. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Statistics in Medicine*, **32**(15), 2643–2660.

Obermeier, V., Scheipl, F., Heumann, C., Wassermann, J., & Kuhchenhoff, H. (2015). Flexible distributed lags for modelling earthquake data. *Journal of the Royal Statistical Society: Series C*, **64**, 395–412.

Pope, C.A. (2007). Mortality effects of longer term exposures to fine particulate air pollution: review of recent epidemiological evidence. *Inhalation Toxicology*, **19** Suppl 1:33–38.

Rushworth, A.M., Bowman, A.W., Brewer, M.J., & Langan, S. J. (2013). Distributed lag models for hydrological data. *Biometrics*, **69**, 537–544.

Schwartz, J. (2000). The distributed lag between air pollution and daily deaths. *Epidemiology*, **11**, 320–326.

Vicedo-Cabrera, A.M., Sera, F. & Gasparrini, A. (2019). Hands-on tutorial on a modeling framework for projections of climate change impacts on health. *Epidemiology*, **30** (3), 321–329.

Wang, X.Y., Guo, Y., FitzGerald, G., Aitken, P., Tippett, V., Chen, D., Wang, X. & Tong, S. (2015). The impacts of heatwaves on mortality differ with different study periods: a multi-city time series investigation. *PLoS One* **10** (7), e0134233.

Welty, L.J., Peng, R.D., Zeger, S.L. & Dominici, F. (2009). Bayesian distributed lag models: estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, **65**(1), 282–91.

WMO e WHO (2015). *Heatwaves and Health: Guidance on Warning-System Development*. G.R. McGregor, lead editor P. Bessemoulin, K. Ebi and B. Menne, editors.

Zanobetti, A.,Wand, M., Schwartz, J. & Ryan, L. (2000). Generalized additive distributed lag models: quantifying mortality displacement. *Biostatistics*, **1**, 279–292.

Zhang, Y., Yakob, L., Bonsall, M.B. & Hu, W. (2019). Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data. *Scientific Reports*, **9** (1), Article number: 3262.

A velha questão do cálculo do tamanho da amostra numa era em que os dados longitudinais ganham terreno

Luzia Gonçalves, *luziag@ihmt.unl.pt*

Unidade de Saúde Pública Internacional e Bioestatística, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa e CEAUL

Introdução

Na investigação nas áreas da saúde surge, cada vez com mais frequência, o problema do cálculo do tamanho da amostra. Neste aspeto há uma notória evolução nos últimos anos, mas nem sempre a estatística tem uma resposta atempada. Principalmente nos estudos longitudinais ainda há muitas situações práticas para as quais o problema do cálculo do tamanho amostral não tem uma solução imediata. Mesmo com os desenvolvimentos em termos da estrutura da matriz de correlação, na prática aparecem algumas barreiras difíceis de ultrapassar. Sem fornecer detalhes matemáticos, neste trabalho evidenciam-se algumas potencialidades em termos de investigação estatística com elevada aplicabilidade prática.

Nas instituições ligadas à saúde, a não ser que o estatístico trabalhe com dados oficiais, de uma forma geral tem o trabalho divido pela análise de dados recolhidos no âmbito de investigações e pela preparação de protocolos de investigação com vista, por exemplo, à submissão para aprovação de comissões de ética ou para a obtenção de financiamento para a investigação proposta. Em caso de sucesso destas submissões, os estatísticos dedicam também tempo para a implementação da recolha de dados no terreno. Em caso de insucesso, estes protocolos de investigações não geram dados para análise estatística futura, mas a estatística tem um papel crucial na prevenção destes desfechos desfavoráveis. A "estatística sem dados" tem, portanto, um espaço importante para desenvolvimentos teóricos com futura aplicação prática de grande relevo.

A questão do cálculo ou da justificação do tamanho amostral (n) está atualmente enraizada na comunidade científica, no âmbito da preparação dos protocolos de investigação, e o papel do estatístico, nesta etapa, está em fase de expansão. No entanto, as fórmulas de cálculo do tamanho amostral que dominam ainda são as clássicas destinadas à estimação de uma proporção, de um valor médio, ou das diferenças de duas proporções ou dois valores médios, num contexto de estudos transversais. Em estudos de caso-controlo, as dimensões das amostras de casos e de controlos também aparecem ainda descritas de forma redutora, pois fixam-se frequentemente na frequência de uma única exposição nos controlos, quando por vezes se querem estudar várias exposições. Os programas disponíveis, além do nível ou do coeficiente de confiança e da potência requerida, exigem uma estimativa para a razão das chances (*odds ratio*) e a estimativa da frequência de exposição nos controlos. Repare-se que apesar do amplo uso da regressão logística múltipla em saúde, em termos do cálculo do tamanho da amostra a literatura atual ainda está limitada a situações muito restritas (van Smeden et al, 2019, Bujang et al, 2018, van Smeden et al, 2016). Nos estudos de coorte, de forma semelhante, partindo de uma estimativa para a incidência da doença nos não expostos e de uma estimativa do risco relativo, a dimensão da amostra pode ser determinada.

Num estudo longitudinal, em que observações repetidas de uma variável resposta são obtidas ao longo do período de seguimento, num ou mais grupos em estudo, o cálculo do tamanho amostral necessita de desenvolvimentos para contemplar as diferentes situações que aparecem na prática.

Cabral & Gonçalves (2011) dão realce a dois tipos de modelos: o marginal e o modelo de efeitos aleatórios, para a análise de dados longitudinais, havendo ainda a classe dos modelos de transição tal como sugerido por Diggle et al (2002). Será que a ampla gama de opções para o tratamento de dados longitudinais tem sido acompanhada com desenvolvimentos sólidos para o cálculo do tamanho amostral?

Tal como refere Baghfalaki (2019), apesar do problema da determinação do tamanho amostral não ser um assunto novo, referindo trabalhos datados de 1973, é um assunto que, embora pouco estudado, pode ser desafiante. A partir de 2000, surgiram diversos trabalhos para determinar o *n* associado aos modelos lineares mistos (LMM) e uso de equações de estimativas generalizadas (GEE) na análise dos dados longitudinais (e.g., Liu & Colditz, 2017, Ahn et al, 2014, Guo et al, 2013, Dang et al, 2008, Muller et al, 2007, Muller and Stewart, 2006, Jung & Ahn, 2003, Ahn & Jung 2004, 2003, Diggle et al, 2002, Verbeke & Molenberghs, 2000). Anteriormente, entre outros, também Liu & Liang (1997) e Liang & Zeger (1986) já tinham explorado este assunto. Trindade et al (2011) apresentam as fórmulas e comparam as propostas de Twist (2003), Kirby et al (1994) e Overall & Doyle (1994). Baghfalaki (2019) and Wang and Gelfand (2002) exploram este problema numa perspetiva Bayesiana.

Mesmo sem fazer uma revisão da literatura exaustiva, nota-se uma expansão teórica neste campo, embora ainda muito centrada numa perspetiva da comparação de dois grupos acompanhados ao longo do tempo. Por exemplo, no âmbito dos ensaios clínicos ou noutros estudos experimentais, com alguma frequência, surgem situações com mais de dois grupos ou braços acompanhados ao longo do tempo (e.g., Gasparinho, 2019). Por outro lado, acrescendo a presença de um conjunto de covariáveis ou as dificuldades típicas dos estudos longitudinais, com as habituais perdas de seguimento e presença de valores omissos, frequentemente, há problemas práticos sem um enquadramento teórico imediato. Assim, quer do ponto de vista de desenvolvimentos teóricos, quer do desenvolvimento de *software* ainda existe muito por fazer. Entre diversos aspetos a considerar no processo da determinação da dimensão da amostra, neste trabalho, focamos a problemática da escolha da matriz da correlação.

A quase "obsessão" pelo cálculo do tamanho da amostra

O cálculo do tamanho amostral além de dependente do objetivo do estudo, parte de suposições que podem variar muito facilmente (Trindade et al, 2011). Por esta razão, os estatísticos, juntamente com os investigadores de outras áreas do conhecimento, devem procurar explorar várias tentativas para o cálculo amostral. No entanto, numa situação destinada a financiamento para a investigação, é necessário elaborar um orçamento que pode ter um teto muito limitado, havendo necessidade de fixar um valor final para o tamanho amostral também defendido por outros aspetos como o tempo para a investigação, recursos humanos e mesmo disponibilidade das unidades amostrais (Liu & Colditz, 2017, Dattalo, 2008). Por exemplo, no caso de querer efetuar um estudo numa maternidade de uma cidade de uma província de Moçambique, que visa acompanhar recém-nascidos no mês de janeiro de 2020, durante 24 meses, o tamanho amostral "real" está limitado ao número médio de nascimentos naquele mês que pode ser inferior a 300 nascimentos. Como, por vezes, surgem dificuldades para obter o consentimento informado dos pais para participar no estudo, facilmente o *n* poderá ficar bastante abaixo dos 300. Neste caso, pode ser indicado optar pelo cálculo da potência estatística em função do tamanho *n* que se poderá obter na prática, de modo a perceber se vale a pena realizar a investigação.

Quer se trate de "justificar" ou "calcular" o tamanho amostral (ou a potência estatística para um dado *n*), os investigadores da área da saúde sabem que numa futura publicação, este aspeto é de apresentação obrigatória na secção de Material e Métodos. Com a crescente adoção nas revistas científicas das linhas de orientação para reportar estudos observacionais, como é o caso da iniciativa denominada *Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE), com diferentes adaptações, por exemplo, para dados longitudinais (Zavada et al., 2014) e traduções para diversas línguas (e.g., em português, Malta et al, 2010), ou para ensaios clínicos como é o caso da CONSORT (*Consolidated Standards of Reporting Trials*), dificilmente à posterior, um trabalho será

publicado sem explicar o cálculo do tamanho amostral, conjugando aspetos estatísticos com outros de natureza distinta. Estas linhas orientadoras, vieram cimentar o papel da estatística, mas por vezes conduzem a um certo exagero. Algumas publicações apresentam até as fórmulas matemáticas, mas para um parâmetro que não obedece ao objetivo principal do estudo. Por exemplo, num estudo transversal para estudar os conhecimentos, atitudes e perceções face a uma doença, por parte de mulheres, aparece a fórmula para estimar a prevalência da doença em crianças em idade escolar, assumindo que esta é semelhante à dos adultos, não estando o objetivo principal e a escolha da fórmula em harmonia (e.g., Mutsaka-Makuvaza et al, 2019). Essa recolha de dados fez-se através de um questionário, assim o tamanho amostral obtido pelo método anterior poderá não estar muito longo do alvo, mas também podemos ter discrepâncias consideráveis, com custos desnecessários.
Nos estudos transversais para estimar a prevalência de um dado acontecimento, por norma há mais informação disponível. Por exemplo, pode haver estudos pilotos, ou outras investigações em contextos similares com resultados publicados, ou não, que poderão dar uma ideia acerca das estimativas iniciais para os parâmetros de interesse. Por vezes, as estatísticas oficiais também podem ajudar neste processo. Também, numa abordagem quase Bayesiana, poderá haver informação de especialistas que pode ser quantificada para este fim. Mesmo num estudo transversal que tenha como objetivo estimar uma proporção binominal (p), pequenas variações de p conduzem a diferenças consideráveis no tamanho amostral, por exemplo, para o método de Wald (ainda o mais popular nas aplicações). Entre métodos alternativos ao método de Wald, para alguns intervalos de variação de p, também existem discrepâncias consideráveis para n (Gonçalves et al, 2012).

No processo editorial e de revisão das publicações, esta "fixação" no n, com uma maior experiência em estudos transversais, também penaliza as situações em que o tamanho amostral foi calculado para um estudo longitudinal. Como é conhecido, geralmente, nos estudos longitudinais o n necessário, para atingir uma dada potência, é menor que nos estudos transversais. Porém, os investigadores queixam-se de comentários na linha da "a amostra é reduzida", notando-se um pensamento enviesado pela maior experiência no outro tipo de estudos.

Tipos de matrizes de correlação em estudos longitudinais com medidas de natureza contínua

Nos estudos longitudinais, o *software* disponível é mais escasso do que para estudos transversais. Nestes, sendo mais simples, as estimativas de alguns parâmetros são mais fáceis de obter através da literatura ou de estudos piloto. Porém, nos longitudinais ao trabalhar com medições repetidas, a questão da estrutura da correlação, torna ainda o problema mais complicado. Entre diferentes tipos de matrizes de correlação, Guo et al (2013) discutem as vantagens e desvantagens das escolhas de matrizes com (a) correlações entre pares nulas; (b) correlações entre pares iguais; (c) correlações com padrões estabelecidos; (d) correlações sem uma estrutura definida. No caso de correlações iguais, habitualmente, os programas solicitam o coeficiente intraclasses (Guo et al, 2013). Um dos programas com ampla utilização prática, o *G*Power* (versão 3.1), calcula o n baseado na ANOVA clássica com medições repetidas, permitindo especificar o valor da correlação.

Caso exista um padrão definido, teoricamente, no contexto de dados longitudinais consideram-se as matrizes de correlação com uma estrutura auto-regressiva de primeira ordem – AR1 – que é um caso particular da família LEAR (*Linear exponent first-order autoregressive*) que assume que as correlações, entre pares de medições, decrescem exponencialmente com o tempo ou distância (Guo et al, 2013). O programa GLIMMPSE permite calcular o n com esse tipo de matrizes. No caso de uma matriz não-estruturada, cada par de medições apresenta um valor único, sem um padrão definido. Guo et al (2013) consideram que a matriz LEAR constitui um bom compromisso entre a simplicidade da matriz com correlações iguais e complexidade da matriz não-estruturada que exige um número elevado de parâmetros a estimar. Para variáveis resposta contínuas, usando GEE para testar a diferença entre dois declives, no Programa PASS *Sample Size Software* estão disponíveis essas matrizes e adiciona outras variantes, nomeadamente a *Banded(1)* e *Banded(2)* que contemplam correlações não-nulas para pares que distam um tempo ou um ou dois tempos, respetivamente, e nulas

para os outros casos. Por exemplo, a *Banded(2)* , sendo a correlação (ρ) do momento *baseline*, tem o seguinte aspeto:

$$\begin{bmatrix} 1 & \rho & \rho & 0 & \dots & 0 \\ \rho & 1 & \rho & \rho & \dots & 0 \\ \rho & \rho & 1 & \rho & \dots & 0 \\ 0 & \rho & \rho & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Entre outras variantes, a matriz auto-regressiva também se desdobra noutra variante proposta por Ahan et al (2015), com diagonal igual a um e restantes elementos iguais a $\rho^{|t_j - t_k|}$.

O Programa PASS, tendo a desvantagem de não ser de utilização livre, tem várias opções de cálculo com base em modelos LMM e GEE, notando-se avanços entre as últimas versões. Recentemente, no Programa R (R Core Team, 2018), Donohue (2019) propôs o package "longpower" (versão 1.0-19) que oferece diferentes possibilidades de cálculo (*n* ou potência), seguindo os desenvolvimentos de Liu and Liang (1997), Diggle et al (2002) e Lu et al (2008). Pelo menos em algumas opções, estes programas oferecem a possibilidade de incorporar as perdas de seguimento e valores omissos. Guo et al (2013), na Tabela 1 descrevem o que é necessário especificar no programa GLIMMPSE, constando além dos requisitos habituais, as estimativas da variância para cada medida repetida e da matriz de correlação para os diferentes pares de medidas repetidas. Guo et al (2013) realçam que a variância requerida deve ser a residual. Ora, a não ser que haja algum estudo piloto prévio, em princípio, será difícil ter estimativas fiáveis, quer para essa variância, quer para as matrizes de correlação.

Exemplo: O estado nutricional de crianças entre os 2 e os 5 anos, após seguimento de dois anos no Bengo, Angola, usando um estudo experimental com 4 braços (Gasparinho, 2019)

Na literatura médica, as matrizes de correlação de estudos similares não são geralmente publicadas e, neste caso, seriam extremamente valiosas. Gasparinho (2019) implementou um estudo experimental com 4 braços, com seguimento durante dois anos no Bengo em Angola. Este estudo teve como objetivo de investigar se uma dose anual de desparasitação com albendazol (ALB) ou se uma abordagem que inclua o diagnóstico e o tratamento dos parasitas intestinais de quatro em quatro meses, a nível individual ou do agregado familiar, têm impacto sobre o estado nutricional de crianças entre os 2 e os 5 anos. O crescimento das crianças foi avaliado pelos indicadores: estatura, peso, estatura para idade (EIZ), peso para estatura (PEZ), peso para idade (PIZ) e perímetro braquial em Z-scores no início do estudo e aos 4, 8, 12, 16, 20 e 24 meses de seguimento. Posteriormente, a análise por intenção-de-tratar foi realizada após a análise e tratamento de valores omissos e usando uma abordagem não paramétrica *nparLD*, modelos de efeitos mistos e equações de estimativas generalizadas para dados longitudinais. Para calcular o tamanho amostral previamente, deu-se destaque à estatura para idade (EIZ) por ser um indicador que reflete o crescimento linear da criança, sendo importante para o diagnóstico da desnutrição crónica. No entanto, noutras situações, também poderia ser importante dar destaque às outras medições. Neste caso, para estudos futuros, na Tabela 1 ilustram-se algumas matrizes de correlações amostrais obtidas para a amostra de 121 crianças analisadas nos 6 momentos do seguimento.

Tabela 1. Matrizes de correlação obtidas para EIZ, PEZ e PIZ

		Momentos de seguimento (meses)				
EIZ	4	8	12	16	20	24
4	1	0,968	0,966	0,956	0,938	0,927
8		1	0,971	0,963	0,952	0,934
12			1	0,983	0,972	0,960

16				1	0,985	0,970
20					1	0,970
24						1
PEZ	4	8	12	16	20	24
4	1	0,813	0,788	0,768	0,763	0,725
8		1	0,864	0,809	0,867	0,806
12			1	0,844	0,868	0,813
16				1	0,837	0,766
20					1	0,822
24						1
PIZ	4	8	12	16	20	24
4	1	0,938	0,909	0,888	0,890	0,864
8		1	0,947	0,913	0,941	0,916
12			1	0,921	0,943	0,913
16				1	0,926	0,881
20					1	0,934
24						1

A Tabela 1 mostra correlações amostrais elevadas para os diferentes pares, com suaves decréscimos à medida que aumenta o espaçamento entre medições para o EIZ. Para os indicadores dependentes do peso da criança (PIZ e PEZ) as correlações são também elevadas, mas existem mais oscilações, embora também se note alguma tendência para um decréscimo com o distanciamento dos momentos de observação da criança. Se quiséssemos efetuar um novo estudo similar, na área em estudo, estas matrizes de correlação poderiam servir como estimativas iniciais, mas mesmo assim ao conjugar com as matrizes teóricas podem surgir dúvidas.

Por exemplo, numa matriz AR(1), partindo de uma correlação de 0,97 teríamos uma primeira linha para a matriz associada ao EIZ da ordem de:

0,97	0,94	0,91	0,89	0,86

Ou seja, com um decréscimo aparentemente mais acentuado que o registado em termos amostrais. Porém, outras matrizes do tipo *Banded(1)* and *Banded(2)* parecem estar afastadas quer para EIZ quer para PIZ e PEZ. Para este caso, inicialmente foi usado o Programa GLIMMPSE pois tratava-se de um estudo experimental com 4 braços, partindo de uma correlação de 0,95 com decréscimos de 1%, tendo-se chegado a 38 crianças por braço, para uma potência de 0,80 e outros requisitos associados que não são apresentados aqui. No entanto, num contexto de pobreza, mesmo tendo um bom suporte logístico, as 152 crianças previstas inicialmente pelos cálculos, reduziram-se a 121, pois estas crianças deviam estar infetadas pelo menos com um parasita intestinal. Neste caso, é muito vantajoso o cálculo da potência atingida. Alguns estudos de simulação, em paralelo com os fundamentos teóricos, são uma mais valia na prática.

Notas finais

Fora do enquadramento clássico dos estudos epidemiológicos, o problema do cálculo do tamanho da amostra pode não ter uma resposta imediata. Este aspeto é importante, numa situação de prazos para cumprir, como é o caso da submissão de projetos a financiamento. Os desenvolvimentos teóricos e de *software* devem ser concomitantes, de forma a que a disseminação de conhecimento, para as áreas de aplicação, seja mais veloz. Com a expansão dos estudos longitudinais, quer em termos teóricos, quer em termos de implementação em diversas áreas da saúde, nota-se que ainda existem lacunas a preencher no campo da determinação dos tamanhos amostrais em diferentes situações. Os estudos de

simulação e os estudos comparativos com diferentes métodos propostos na literatura são escassos e merecem um maior investimento.

Agradecimentos

Um agradecimento especial à Júlia Teles e à Maria Helena Gonçalves pela leitura e sugestões. Trabalho financiado pela Fundação para a Ciência e Tecnologia - UID/MAT/00006/2019 e UID/Multi/04413/2019.

Referências Bibliográficas

Ahn C, Heo M, Zhang S. *Sample Size Calculations for Clustered and Longitudinal Outcomes in Clinical Research*. Boca Raton: CRC Press, 2014.

Ahn C, Jung SH. Efficiency of general estimating equations estimators of slopes in repeated measurements: Adding subjects or adding measurements? *Drug Information Journal* 2003; 37(3):309–316.

Ahn C, Jung SH. Effect of dropouts on sample size estimates for test on trends across repeated measurements. *Journal of Biopharmaceutical Statistics* 2004;15(1):33–41.

Baghfalaki T. Bayesian sample size determination for longitudinal studies with continuous response based on different scientific questions of interest. *Journal of Biopharmaceutical Statistics* 2019, 29:2, 244-270.

Bloch DA. Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Statistics in Medicine* 1986; 5(6):663–667.

Bujang MA, Sa'at N, Tg Abu Bakar Sidik TMI, Lim CJ. Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *Malays J Med Sci.* 2018; 25(4):122–130. https://doi.org/10.21315/ mjms2018.25.4.12

Cabral MS, Gonçalves MH. *Análise de dados longitudinais*. Sociedade Portuguesa de Estatística. 2011.

Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Computer Methods and Programs in Biomedicine* 2008; 91(2):122–127.

Dawson JD. Sample size calculations based on slopes and other summary statistics. *Biometrics* 1998; 323–330.

Dattalo P. *Determining Sample Size - Balancing Power, Precision, and Practicality*. Oxford University Press, 2008.

Diggle P, Heagerty P, Liang K, Zeger S. *Analysis of Longitudinal Data.* Oxford: Oxford University Press, 2002.

Gasparinho C. *Malnutrition and enteric infections in children in Bengo province, Angola a four-arm experimental study*. Tese de Doutoramento em Saúde Internacional, Universidade Nova de Lisboa, Instituto de Higiene e Medicina Tropical, Lisboa, Portugal, 2019.

Gonçalves L, de Oliveira MR, Pascoal C & Pires A. Sample size for estimating a binomial proportion: comparison of different methods. *Journal of Applied Statistics* 2012; 39:11, 2453-2473

Guo Y, Logan HL, Glueck DH, Muller KE. Selecting a sample size for studies with repeated measures. *BMC Medical Research Methodology* 2013,13:100.

Jung SH, Ahn C. Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statistics in Medicine* 2003; 22(8):1305–1315.

Lefante JJ. The power to detect differences in average rates of change in longitudinal studies. *Statistics in Medicine* 1990; 9(4):437-446.

Liang KY & Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73, 13-22.

Liu G, Liang KY. Sample size calculations for studies with correlated observations. *Biometrics* 1997; 937–947.

Liu J, Colditz GA. Optimal design of longitudinal data analysis using generalized estimating equation models. *Biom J*. 2017; 59(2): 315–330.

Malta M, Cardoso LO (tradutores). In: Malta M, Cardoso LO, Bastos FI, Magnanini MMF, Silva CMFP. Iniciativa STROBE: subsídios para a comunicação de estudos observacionais. *Rev Saude Publica* 2010; 44(3):559-65.

Donohue MC. longpower: Power and sample size calculations for longitudinal data. R package version 1.0-19, 2019.

Kirby AJ, Galai N & Muñoz A. Sample size estimation using repeated measurements on biomarkers as outcomes. *Controlled Clinical Trials* 1994; 15: 165-172.

Mutsaka-Makuvaza MJ, Matsena-Zingoni Z, Tshuma C, Katsidzira A, Webster B, Zhou X-N, Midzi N. Knowledge, perceptions and practices regarding schistosomiasis among women living in a highly endemic rural district in Zimbabwe: implications on infections among preschool-aged children. *Parasites & Vectors* 2019; 458, 1.

Overall JE, Doyle SR. Estimating sample size for repeated measurement designs. *Controlled Clinical Trials* 1994; 15: 100-123.

Trindade DB, Esquivel RM, Amorim LDAF. Tamanho amostral para análise de medidas repetidas em estudos longitudinais. In: Simpósio Nacional de Probabilidade e Estatística, 2011; São Pedro, São Paulo. São Paulo: *Associação Brasileira de Estatística*; 2011.

Twisk, JWR. *Applied longitudinal data analysis for Epidemiology*. Cambridge, 2003.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2008.

von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Lancet* 2007; 370(9596):1453-1457.

van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC and Reitsma JB. Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research* 2019; 28(8) 2455–2474.

van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016; 16: 163.

Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*. Springer Science and Business Media, Newyork, NY, 2000.

Zavada J, Dixon WG, Askling J. Launch of a checklist for reporting longitudinal observational drug studies in rheumatology: a EULAR extension of STROBE guidelines based on experience from biologics registries. *Ann Rheum Dis*. 2014; 73(3):628.

Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science* 2002; 17(2):193–208.

Ciência Estatística

• Artigos em Revistas

Nuno M. Brites, Carlos A. Braumann (2019). Fisheries management in randomly varying environments: Comparison of constant, variable and penalized efforts policies for the Gompertz model. *Fisheries Research* **216**: 196-203.
https://doi.org/10.1016/j.fishres.2019.03.016
(http://www.sciencedirect.com/science/article/pii/S0165783619300803)

Nuno M. Brites, Carlos A. Braumann (2019). Harvesting in a random varying environment: optimal, stepwise and sustainable policies for the Gompertz model. *Statistics, Optimization and Information Computing* **7 (3)**: 533–544. DOI: 10.19139/soic-2310-5070-830 http://www.iapress.org/index.php/soic/article/view/soic.20190902

Papança, F. (2018). As Lições de Balística na Escola Central de Sargentos na década de 1940. *Revista Proelium da Academia Militar*. VIII (1) 245-254.

• Livros

Título: *Introduction to Stochastic Differential Equations with Applications to Biology and Finance*
Autor: Carlos A. Braumann
Ano: 2019. Editora: Wiley, Hoboken NJ.
ISBN: 978-1-119-16606-1 (Hardcover), 978-1-119-16608-5 (E-Book), 978-1-119-16609-2 (O-book).

Título: *XXIV Congresso SPE – Programa e Livro de Resumos.*
Autoras: Sandra Ramos, Maria João Polidoro, Ana Borges e Luísa Hoffbauer.
Ano: 2019. Editora: Edições SPE. ISBN: 978-972-8890-44-5. Depósito Legal: 462497/19.

Título: *Análise Estatística de Dados Financeiros*
Autores: Conceição Amado, Cláudia Nunes e Alberto Sardinha
Ano: 2019. Editora: Edições SPE. ISBN: 978-972-8890-43-8. Depósito Legal: 462496/19.

• Teses de Mestrado

Título: Dynamic prediction of long-term survival in patients with early-stage breast cancer
Autora: Ana Sofia da Silva Azevedo, *asazevedo@fc.ul.pt*
Orientadoras: Lisete Sousa e Susana Esteves

Título: Variantes da Solução de Scheffé para o Problema de Behrens-Fisher Baseadas em Reamostragem
Autor: João Luís Nogueira de Oliveira, *jlnoliveira@outlook.com*
Orientador: Sílvio Velosa

• Teses de Doutoramento

Título: Contributos Computacionais e Metodológicos na Estimação de Valores Extremos
Autora: Helena Penalva, *helena.penalva@esce.ips.pt*
Orientadoras: M. Ivette Gomes, M. Manuela Neves e Sandra Nunes
Esta informação completa a página 62 do Boletim SPE primavera de 2019.

Título: Métodos paramétricos de screening em classificação supervisionada, na presença de populações assimétricas
Autora: Ana Sofia Monteiro Araújo Soares, *assoares@fc.ul.pt.*
Orientadoras: Marília Antunes e Lisete Sousa

Na minha tese apresentam-se as contribuições resultantes de um trabalho de investigação sobre métodos bayesianos de *screening* em classificação supervisionada num cenário bivariado, ou seja, métodos que permitem atribuir a um novo indivíduo uma categoria de entre um conjunto de categorias mutuamente exclusivas, com base na observação de vectores de características bidimensionais nesse indivíduo. Considerando a formulação do problema de *screening* do ponto de vista preditivo bayesiano mostra-se como se pode construir uma região de classificação óptima quando se admite um modelo gaussiano assimétrico bivariado para o vector de características, condicional ao grupo. A utilização deste modelo revela um desempenho bastante bom, dado que possui a maior parte das propriedades do modelo gaussiano bivariado e lida com a assimetria dos dados, o que é uma mais valia, pois em dados reais é uma situação muito frequente.
É, então, proposta uma regra de classificação baseada num modelo, havendo, portanto, uma construção probabilística, que resulta da combinação da classificação e de quantidades preditivas *a posteriori* resultantes da aplicação do método a cada par de valores simulados, dado não se dispor de dados reais.
Para ultrapassar os problemas de cálculo associados com a obtenção da região de classificação e das probabilidades preditivas desenvolveu-se um conjunto de algoritmos assentes em métodos de simulação estocástica, nomeadamente o de Monte Carlo ordinário e o de População de Monte Carlo. Todos os programas computacionais foram implementados em ambiente R e permitem obter a região de especificação de forma praticamente automática.
Esta metodologia foi aplicada inicialmente com dois grupos de classificação seguida da extensão a três grupos de classificação. Neste caso colocou-se em prática a mesma metodologia que para dois grupos, que se revelou um mau desempenho, até mesmo não praticável. Optou-se, então, por utilizar a metodologia que consiste em seleccionar o grupo para o qual a probabilidade preditiva condicional é maior. Por último resolveu-se o problema para três grupos, através da aplicação da metodologia de dois grupos de forma sequencial: dois grupos, um contra os outros dois como se fossem um grupo, determinando o classificador óptimo (primeiro cenário binário); de seguida, resolve-se outro problema binário, com um grupo contra outro, com os dois grupos que foram considerados como um só no primeiro cenário binário.
Compararam-se as várias abordagens bayesianas bem como com a abordagem clássica de uma análise discriminante quadrática, tendo esta um pior desempenho.
Os classificadores bayesianos obtidos permitem a obtenção de fronteiras com expressão analítica explícita, sem necessidade de fixar previamente a sua forma e possibilita o cálculo de um conjunto de quantidades preditivas de interesse.

Ana Sofia Soares

Título: Contributions to Spatial and Temporal Modelling
Autora: Andreia Alves Forte de Oliveira Monteiro, *andreiaforte50@gmail.com*
Orientadoras: Raquel Menezes da Mota Leite e Maria Eduarda Silva

Na minha tese, foi abordado o problema de modelar séries temporais com tempos de observação informativos. Tradicionalmente, a modelação espacial e temporal assume que as localizações amostradas (no tempo ou no espaço) são fixas ou estocasticamente independentes do fenómeno espacial e temporal em estudo. No entanto, é bem conhecido que, por exemplo, em estudos de poluição do ar, normalmente as estações de monitorização são colocadas perto das fontes de poluição mais prováveis e em áreas de alta densidade populacional. Em estudos médicos, um paciente é geralmente observado com maior frequência quando apresenta pior condição clínica. Nestes exemplos, nem as observações são obtidas de forma regular no tempo/espaço, nem as localizações das observações (no tempo ou no espaço) são estocasticamente independentes do processo em estudo. Ignorar essa dependência pode levar a estimativas tendenciosas e inferências
enganosas. Neste contexto, foi introduzido o conceito de Amostragem Preferencial na dimensão temporal e discutidas diferentes abordagens baseadas em modelos para fazer inferência e previsão debaixo deste esquema de amostragem.
Numa primeira abordagem, foi apresentado um modelo para lidar com séries temporais irregularmente espaçadas em que o desenho amostral depende do valor contemporâneo do processo subjacente, sob a hipótese de uma variável de resposta Gaussiana. Para este modelo, foram apresentados dois métodos de estimação, um baseado em simulações de Monte Carlo e outro baseado numa aproximação de Laplace e na adoção de uma técnica baseada em equações diferenciais parciais estocásticas para aproximar o processo subjacente.
Na segunda abordagem, foi proposto um modelo para séries temporais nas quais o desenho amostral depende de toda a história passada dos processos observados, tempos das observações, bem como os valores dessas observações. Tendo em conta a ordem natural do tempo subjacente aos dados disponíveis representados por uma série temporal, uma abordagem de modelação baseada em processos evolucionários foi uma escolha natural para dados com estas caraterísticas. Os modelos propostos foram validados através de estudos numéricos com conjuntos de dados reais e simulados.
O principal objetivo desta tese foi desta forma centrado na apresentação de contribuições para a modelação espacial e temporal nomeadamente no contexto da análise de dados irregularmente espaçados e em que o processo dos tempos/localizações de observação é estocástico e fornece informações adicionais sobre os fenómenos em estudo.

Andreia Monteiro

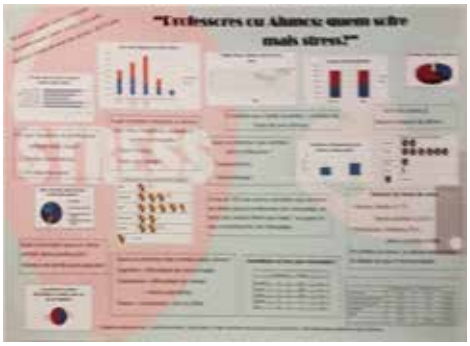
Sessão de Entrega dos
Prémios Estatístico Júnior 2019

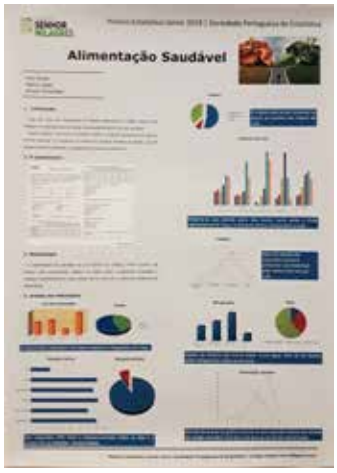
O Prémio Estatístico Júnior 2019, promovido pela Sociedade Portuguesa de Estatística e patrocinado pelo Centro de Matemática da Universidade de Coimbra, pretende incentivar o interesse pelas áreas de Probabilidades e Estatística dos estudantes dos Ensinos Básico e Secundário, e dos Cursos de Educação e Formação (CEF) e de Educação e Formação de Adultos (CEFA).
Nesta sessão, que teve lugar a 9 de Novembro de 2019, pelas 14h30m, no Hotel Casa da Calçada em Amarante, durante o XXIV Congresso da Sociedade Portuguesa de Estatística, foram entregues os Prémios atribuídos aos trabalhos distinguidos este ano.

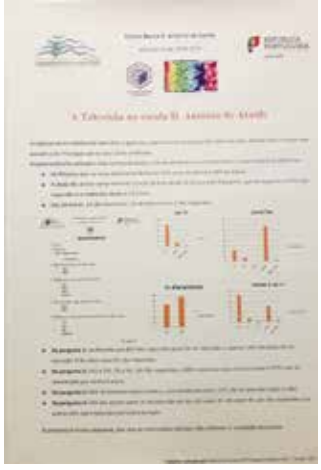
A sessão iniciou-se com uma actividade assegurada pelo núcleo de Aveiro do Circo Matemático, sediado no Departamento de Matemática da Universidade de Aveiro. Este núcleo contribui para dar corpo, na região norte do país, ao projecto da associação LUDUS, que visa espalhar, de forma itinerante, o deslumbramento e o fascínio da Matemática através de experiências lúdicas.

Surgiram assim vários momentos recreativos com truques envolvendo um baralho de cartas, cordas, nós e cálculo matemático, com ampla participação da assistência, e com a intervenção inopinada, mas sempre assertiva e alegre, do palhaço Totó. A intervenção do Circo Matemático incluiu também a explicação matemática subjacente a alguns dos truques exibidos.

Seguiu-se a entrega dos Prémios, com a presença da Doutora Manuela Neves, em representação do júri, e do Doutor Paulo Eduardo de Oliveira, em representação do CMUC, e um breve lanche para os laureados.







Prémios Estatístico Júnior 2019

Trabalho classificado em 2º lugar (3º ciclo do Ensino Básico)

Título: ***Hábitos Alimentares***
Autores: Ana Filipa Bernardes da Graça, Maria Almeida Marques Lopes e Bruna Alexandra Santos Fernandes.
Professora orientadora: Ana Rita Orfão Ramos.
Estabelecimento de Ensino: Colégio Senhor dos Milagres, Leiria

Trabalho classificado em 1º lugar (Ensino Secundário)

Título: ***Professores ou alunos: quem sofre mais stress?***

Autor: Guilherme Oliveira Pedro, João Miguel Isidro Antas e Tiago R. de Sousa Lopes Quinteiro.
Professor orientador: Maria Alice da Silva Martins
Estabelecimento de Ensino: Escola Básica e Secundária Artur Gonçalves, Torres Novas

Trabalho classificado em 3º lugar *ex-aequo* (Ensino Secundário)

Título: ***Alimentação Saudável***

Autor: Daniela Leitão dos Santos, Ana Paula Cunha Saraiva e Tânia Coelho dos Santos
Professor orientador: Mary Cristina Ferreira da Rocha
Estabelecimento de Ensino: Ensiguarda - Escola Profissional da Guarda

Trabalho classificado em 3º lugar *ex-aequo* (Ensino Secundário)

Título: ***Estrangeiros Matriculados no Agrupamento Adelaide Cabette de Odivelas***

Autores: Abdul Kadir Asaraf Satar, Beatriz Mendes de Figueiredo e Mariana Faria Pinto
Professor orientador: Vanda Sofia Branco Dias Cerejeira
Estabelecimento de Ensino: Escola Secundária de Odivelas

Trabalho premiado (Cursos EFA/CEF)

Título: ***A Televisão na escola D. António de Ataíde***

Autores: Raquel Alexandra Fernandes Godinho e Marlene Cristina Correia Pereira
Professora orientadora: Sónia Maria Mendes Gonçalves
Estabelecimento de Ensino: Agrupamento de Escolas D. António de Ataíde, Avintes

Nota: Não foram atribuídos o 1º e 3º lugar (Ensino Básico) e 2º lugar (Ensino Secundário)


SOCIEDADE
PORTUGUESA
DE ESTATÍSTICA '19


DE ESTATÍSTICA '19


SOCIEDADE
PORTUGUESA
DE ESTATÍSTICA '19

Prémio SPE 2019

L-moments for automatic threshold selection in extreme value analysis of wave heights from the Gulf of Mexico

Jessica Silva Lomba, *jslomba@fc.ul.pt*
CEAUL e DEIO, Faculdade de Ciências da Universidade de Lisboa

Na análise de valores extremos, sabe-se que a sensibilidade da inferência associada à definição do que se considera um evento extremo é uma questão primordial, que se traduz diretamente na necessidade de escolher um nível apropriado para a utilização de metodologias baseadas em excessos de *threhsold*. Esta escolha deve ser estabelecida anteriormente a qualquer inferência formal. Na abordagem *peaks-over-threshold* (POT), o ajustamento da distribuição Generalizada Pareto (GP) aos excessos de um *threshold* demasiado baixo gera estimação sujeita a um grande viés (a aproximação pode não ser válida). Por outro lado, selecionar apenas observações muito elevadas resulta numa redução do tamanho amostral, aumentando a variabilidade. Assim, é necessário escolher um nível adequado que procure um equilíbrio destas vertentes. Esta é uma questão não raramente alvo de críticas, no que concerne à objetividade (ou à falta dela), sendo usualmente este tópico classificado como um problema em aberto em muitas publicações na área de especialidade.

Encontram-se na literatura metodologias que enfrentam problemas recorrentes: subjetividade inerente a ferramentas como os gráficos de diagnóstico visual, ou alta intensidade computacional de processos mais objetivos, tornando por sua vez impraticável a sua aplicação à análise simultânea de conjuntos de dados em massa (*Big Data*) – exemplos de tais metodologias podem ser encontrados em Northrop and Coleman (2014), Lee, Fan, and Sisson (2015) e Bader, Yan, and Zhang (2018).

A nossa sugestão é um método verdadeiramente automático para a seleção do *threhsold*, visando eficiência computacional e eliminação da análise subjetiva. Baseada na bem estabelecida teoria dos L-momentos, apresentada por Hosking (1986), esta técnica versátil é útil no processamento de grandes coleções de conjuntos de dados extremais, apresentando ainda boa performance quando aplicada a pequenas amostras.

Os L-momentos são mais robustos do que as suas contrapartes clássicas na presença de valores muito extremais na amostra, e não apresentam limitações algébricas relativas ao tamanho amostral, mostrando-se assim apropriados para o nosso contexto.

Empregamos uma abordagem heurística na automatização de uma conhecida ferramenta de diagnóstico, o Diagrama de Rácios de L-momentos (LMRD do inglês *L-moment Ratio Diagram*), habitualmente utilizada em Hidrologia e na Análise de Frequência Regional para discernir entre várias distribuições candidatas para dados regionais, tirando vantagem da forma da GP caracterizada pelo par *(L-simetria, L-curtose)*.

A performance da metodologia é avaliada num grande estudo de simulação e a sua aplicação ilustrada num conjunto de alturas significativas de ondas da literatura. Concluímos que se compara favoravelmente às restantes metodologias do estado-da-arte no que toca à escolha do *threhsold*, estimação dos parâmetros associados e ao objetivo final de estimação de níveis de retorno.

Outros trabalhos foram desenvolvidos sob esta abordagem com base no comportamento assintótico dos L-momentos. Uma versão mais completa do trabalho apresentado, Silva Lomba and Fraga Alves (2019), encontra-se na fase de revisão para publicação numa revista científica internacional.

Referências

Bader, B., Yan, J., & Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, *12(1),* 310–329.

Hosking, J. R. M. (1986). The theory of probability weighted moments. *Research Report RC12210–IBM Research*.

Lee, J., Fan, Y., & Sisson, S. A. (2015). Bayesian threshold selection for extremal models using measures of surprise. *Computational Statistics and Data Analysis,85,* 84–99.

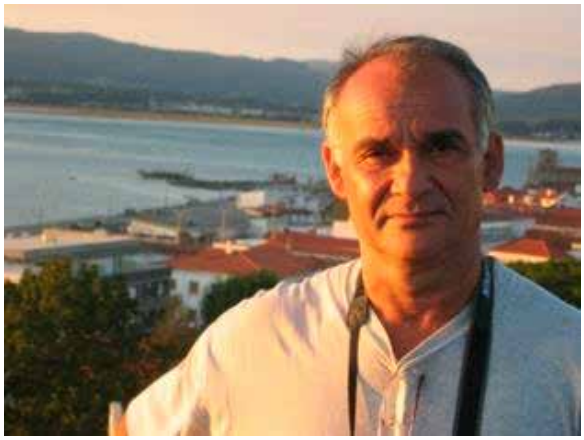
Northrop, P. J. & Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes,17,* 289–303.

Silva Lomba, J. & Fraga Alves, M. I. (2019). L-moments for automatic threshold selection in extreme value analysis. arXiv:1905.08726.

Jessica Silva Lomba, **galardoada com o Prémio SPE 2019**, é Licenciada em Matemática Aplicada e Mestre em Estatística e Investigação Operacional com Especialização em Estatística pela Faculdade de Ciências da Universidade de Lisboa. Atualmente prepara um Doutoramento em Estatística e Investigação Operacional na mesma instituição, sob a supervisão da Professora Doutora Maria Isabel Fraga Alves e com o apoio da bolsa de Doutoramento SFRH/BD/130764/2017 concedida pela Fundação para a Ciência e a Tecnologia, I.P. Executa também funções como Assistente Convidada na Nova School of Business and Economics na área de Métodos Quantitativos.

Prémio Carreira SPE 2019 - Feridun Turkman

É uma grande honra para mim ter sido convidada pela Direcção da Sociedade Portuguesa de Estatística (SPE) para fazer a apresentação do Professor Kamil Feridun Turkman, aquando do Jantar do XXIV Congresso da SPE que se realizou em Amarante, de 6 a 9 de Novembro de 2019. Esta apresentação antecedeu a entrega do prémio Carreira que a SPE outorga aos Estatísticos cujo percurso seja relevante, tanto em termos científicos, como pedagógicos e/ou ainda de divulgação da Estatística em Portugal. Na qualidade de sua ex-aluna de mestrado e de doutoramento é um grande privilégio para mim poder falar sobre a vida e a obra do grande Estatístico que é o Professor Kamil Feridun Turkman.



O Professor Kamil Feridun Turkman nasceu na Turquia em 1953 e foi em Ankara, na Middle East Technical University, que fez o seu bacharelato em Matemática. Em 1976 foi para a Universidade de Sheffield, no Reino Unido, para prosseguir os seus estudos na área de Estatística. Uma das suas colegas de Mestrado era a Professora Antónia Turkman. Pode dizer-se que este encontro marcou o destino do Professor Feridun Turkman.

Com efeito, após terminar o doutoramento na Universidade de Sheffield em 1980, o Professor Feridun Turkman veio para Portugal com a Professora Antónia Turkman, já sua mulher. O Professor Tiago de Oliveira, sabendo que o Professor Feridun Turkman tinha feito a dissertação de mestrado e a tese de doutoramento na área de séries temporais com o Professor Morris Walker, desde logo o convidou para leccionar a disciplina de Séries Temporais na Licenciatura em Matemática Aplicada da Faculdade de Ciências da Universidade de Lisboa (FCUL). Foi assim que em 1981 o Professor Feridun integrou um corpo de docentes recém-doutorados (Professores Dinis Pestana, Ivette Gomes, Antónia Turkman, Amílcar Sernadas, Cristina Sernadas e José Coelho) que, juntamente com os Professores Tiago de Oliveira e Fátima Fontes de Sousa, criaram em Setembro de 1981 o actual Departamento de Estatística e Investigação Operacional (DEIO). Este grupo fundou as primeiras Licenciaturas em Portugal em *Probabilidades e Estatística* e em *Estatística e Investigação Operacional*, assim como os primeiros mestrados e doutoramentos na área, que tantos frutos deram.

O Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) fora já criado em 1975 pelo Professor Tiago de Oliveira e foi óbvia a integração imediata do Professor Feridun Turkman no CEAUL, assim como a sua filiação na "recém-nascida" SPE. Foi assim que o Professor Feridun Turkman iniciou o seu percurso académico-científico, sempre ligado e dedicado ao DEIO, CEAUL e SPE. Em 1983 fez parte da Comissão Organizadora de uma conferência em Extremos (actualmente designada por EVA0), que decorreu no Vimeiro, a qual reuniu uma excelente colecção de jovens investigadores que hoje são internacionalmente famosos na área da Teoria de Valores Extremos.

Grande impulsionador da internacionalização e sempre preocupado em difundir a Estatística em Portugal, o Professor Feridun Turkman trouxe ao DEIO e CEAUL inúmeros investigadores estrangeiros de renome. Destas ligações surgiram muitas actividades importantes das quais vou apenas mencionar duas, em virtude da sua enorme relevância. Em 1989 o convite feito ao Professor Bob Loynes resultou no primeiro programa Erasmus entre o DEIO e a Universidade de Sheffield. A Dra. Zilda Mendes foi a primeira estudante do DEIO a usufruir deste programa. Foram muitos os alunos que puderam aproveitar desta iniciativa do Professor Feridun Turkman, facto que é com certeza desconhecido de muitos alunos que posteriormente participaram no programa Erasmus. Em 1991 um convite feito ao Professor Vic Barnett resultou na criação do projecto SPRUCE (*Statistics in Public Resources and Utilities and Care for the Environment*), o qual teve início num memorável encontro na Fundação Calouste Gulbenkian, em Lisboa, em Abril de 1992. Vários outros SPRUCE *workshops* dedicados à Estatística Ambiental se seguiram em diversas partes do mundo, tendo sido o último

destes eventos realizado em Portugal em 2004. Esta iniciativa deu origem à criação do *SPRUCE TRUSTEE FUND*, da qual faziam parte os Professores Feridun Turkman, Vic Barnett, Clive Anderson, Marion Scott e Richard Smith. Tais encontros originaram a publicação de vários livros, editados pela Wiley, dedicados a problemas emergentes na área da Estatística Ambiental, à atribuição de bolsas e de prémios, entre outras actividades. Foi talvez também graças a esta ligação com o Professor Vic Barnett que surgiu o interesse do Professor Feridun Turkman pela Estatística Espacial que se traduziu mais tarde na criação de um grupo de investigação no CEAUL dedicado a esta temática e ao seu posterior reconhecimento nacional e internacional.

A internacionalização do Professor Feridun Turkman é inquestionável, sendo membro de várias associações de Estatística de renome internacional, tais como a *Royal Statistical Society*, a *Bernoulli Society* e o *International Statistical Institute*.

Os interesses de investigação do Professor Feridun Turkman são muito abrangentes, desde o seu primeiro envolvimento no estudo de Séries Temporais e Teoria de Valores Extremos, até aos modelos Bayesianos Hierárquicos Espaço-temporais, passando por modelação de recursos humanos, taxas de desemprego, problemas ambientais de diversa natureza, entre muitos outros. As teses de doutoramento que orientou reflectem bem esta variedade de interesses. Nos últimos anos, o Professor Feridun Turkman tem colaborado intensamente com o Professor José Pereira, do Instituto Superior de Agronomia e Centro de Estudos Florestais da Universidade de Lisboa, e com o Professor Carlos da Câmara, da FCUL e do Instituto D. Luiz, especialmente na temática da modelação de dados de incêndios florestais, que se reveste de grande relevância para Portugal.

A sua actividade científica tem sido muito profícua. Desde 1980, data de conclusão do seu doutoramento, tem publicado os seus trabalhos em numerosas revistas científicas de renome internacional, tais como *Journal of the Royal Statistical Society, Series B e C*, *Environmetrics*, *The Annals of Probability*, *Extremes* e *Spatial Statistics*, entre muitas outras.

Foi membro do Comité Editorial *do Journal of Applied Mathematics – Open Access Journal* em 2010 e 2011, membro do Comité Editorial *do Journal of Statistical Theory and Practice* desde 2008, e Editor Associado da revista *Revstat – Statistical Journal* de 2014 a 2019. Em 2014 tive o privilégio de ser co-autora do livro *Non-Linear Time Series: Extreme Events and Integer Value Problems*, publicado pela editora Springer, juntamente com o Professor Feridun Turkman e com o Professor Manuel González Scotto. Também gostaria de mencionar o livro, escrito em Português, intitulado *Análise de Sucessões Cronológicas*, do qual o Professor Feridun Turkman foi co-autor com os Professores Bento Murteira e Daniel Müller.

O Professor Feridun Turkman foi ainda investigador principal de diversos projectos de investigação financiados pela Fundação para a Ciência e para a Tecnologia (FCT) e pela *Windsor Treaty* (Reino Unido).

O Professor Feridun Turkman dedicou grande parte do seu tempo, esforço e dedicação ao DEIO e à FCUL, tendo sido presidente do DEIO nos períodos de 1997-2001 e 2009-2012, coordenador do CEAUL de 1992 a 1999, e Presidente da Escola de 2014 a 2017.

A sua capacidade de visualizar oportunidades destinadas a enaltecer e divulgar a Estatística em Portugal, no CEAUL e no DEIO são de realçar. Nos projectos de consultoria conseguidos por sua iniciativa envolveu diversos membros do CEAUL.

Poderia continuar a enumerar mais contribuições do Professor Kamil Feridun Turkman Contudo, termino esta minha exposição realçando as suas qualidades humanas e a sua capacidade de compreensão dos problemas apresentados pelos seus colaboradores. Agradeço ao meu grande amigo, o Professor Feridun Turkman, o seu incondicional apoio nos bons e maus momentos da minha vida.

Patrícia de Zea Bermudez

Prémio Carreira SPE 2019 - Carlos Braumann

Carlos Braumann, Professor Emérito da Universidade de Évora, foi galardoado com o Prémio Carreira da Sociedade Portuguesa de Estatística (SPE), em reconhecimento pelas suas relevantes contribuições no desenvolvimento científico, pedagógico e de divulgação da Estatística em Portugal.

Nascido em Lisboa em 1951, cursou o ensino secundário no Liceu Camões e os estudos universitários na Universidade de Luanda, Angola. Obteve o grau de doutor em 1979 na *State University of New York at Stony Brook* e a agregação em Processos Estocásticos na Universidade de Évora em 1988. É membro eleito do *International Statistical Institute* desde 1992. Foi Presidente da *European Society for Mathematical and Theoretical Biology* (2009-12), Presidente da Sociedade Portuguesa de Estatística (2006-09 e 2009-12), membro do *European Regional Committee da Bernoulli Society* (2008-12), membro do Conselho Superior de Estatística em representação do CRUP (1989-2002) e membro do CEIES (*Comité Consultatif Européen de l' Information Statistique dans les Domaines Économique et Social*, órgão consultivo da Comunidade Europeia, 2005-08).

Na Universidade de Évora, onde desenvolveu quase toda a sua carreira académica desde 1975, Carlos Braumann assumiu vários cargos, entre eles o cargo de Reitor (2010-2014) e Vice-Reitor (1987-94). Foi Presidente do Conselho Científico (1999-2001), Presidente do Conselho do Departamento de Matemática e da Área Departamental de Ciências Exatas (1991, 2005-07) e Diretor do Centro de Investigação em Matemática e Aplicações (1994-99). Lecionou mais de 30 disciplinas, de todos os ciclos de ensino e nas mais variadas áreas de aplicação, em várias instituições de ensino superior, mas principalmente na U. Évora. Coordenou e participou na criação e restruturação de vários planos curriculares nos diferentes ciclos de estudo e foi perito da A3ES para a avaliação institucional. É membro do Centro de Investigação em Matemática e Aplicações da Universidade de Évora e as suas publicações (cerca de uma centena foram sujeitas a arbitragem científica) têm incidido especialmente nas equações diferenciais estocásticas e suas aplicações, particularmente biológicas (dinâmica de populações, pescas, crescimento individual de animais, demografia, etc.) e financeiras. Participou e coordenou vários projetos de investigação e orientou pós-doutoramentos, teses de doutoramento e de mestrado, trabalhos de fim de curso e estágios.

Na sua qualidade de Presidente da SPE, presidiu a congressos anuais e suas Comissões Científicas e foi coeditor das Atas publicadas pela Sociedade em português e das primeiras Atas publicadas em inglês (no âmbito de um acordo com a Springer de várias sociedades nacionais de Estatística). Integrou a organização da *57th Session of the International Statistical Institute-ISI* em Lisboa (2007) e promoveu, em parceria com a Associação Brasileira de Estatística, a elaboração do Glossário Estatístico em Língua Portuguesa (já na "webpage" do ISI). Foi fundador da *Federation of European National Statistical Societies*, que a SPE integrou. Promoveu a criação da jSPE (secção de jovens estatísticos da SPE) e de novas iniciativas (com financiamento do Ciência Viva) de divulgação junto dos jovens (a exposição interativa itinerante Explorística, que foi galardoada com um Prémio da *International Association for Statistical Education*, e a iniciativa Radical Estatística).

Russell Alpizar Jara

O MUNDO DA ESTATÍSTICA

ORGANIZAÇÃO PARTICIPANTE

Índice

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR