Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *Outono de 2010*

Estatística Espacial

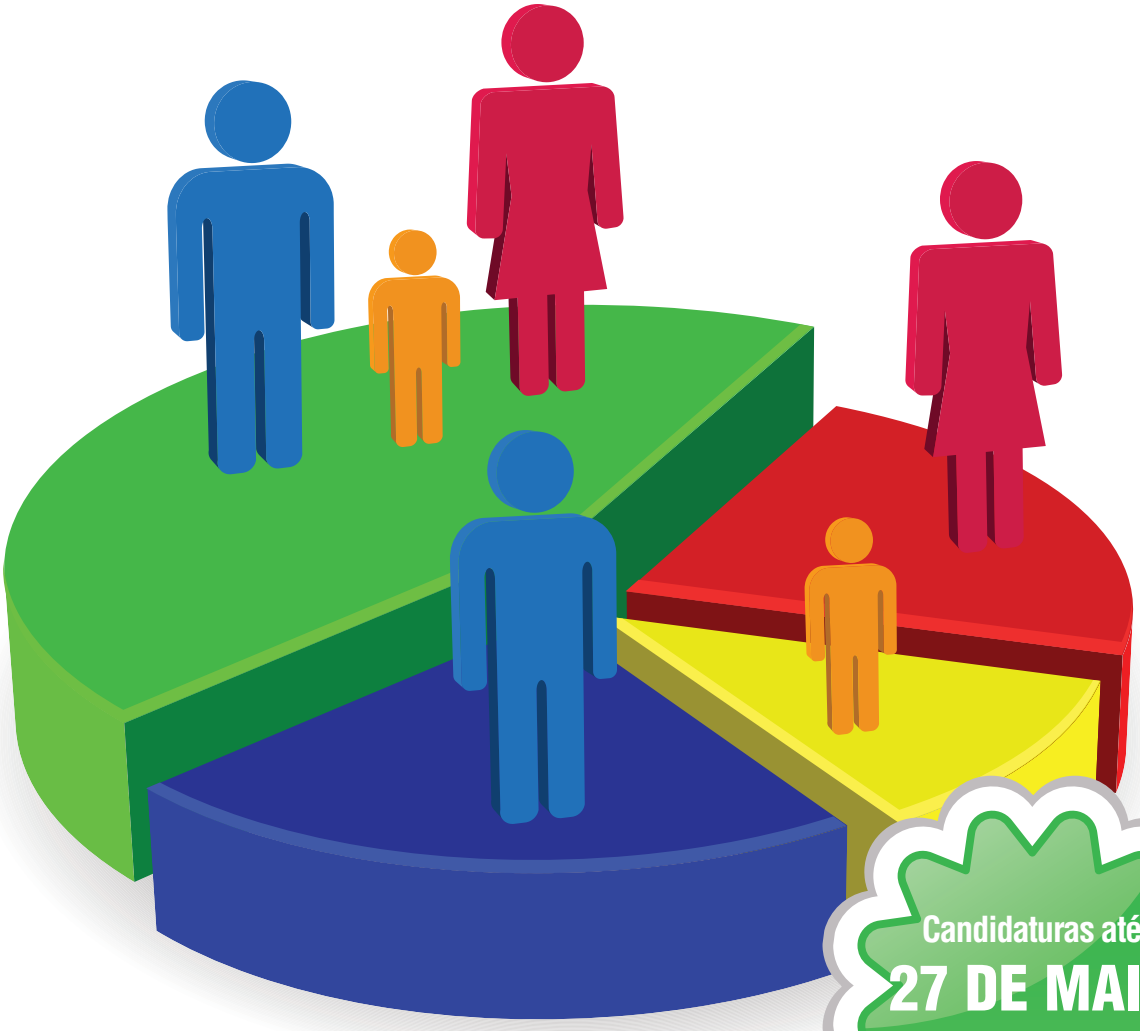
Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe@fc.ul.pt*
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 1000 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

Este Boletim tem o apoio da **FCT** Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO ESTATÍSTICO JÚNIOR 2011



Candidaturas até 27 DE MAIO DE 2011

CONTACTOS

Sociedade Portuguesa de Estatística
Bloco C6, Piso 4 – Campo Grande
1749-016 Lisboa

Telef./Fax 21 750 01 20

www.spestatistica.pt

spe@fc.ul.pt

Com o apoio:

Porto Editora

COMEMORAÇÕES DO 1º DIA MUNDIAL DA ESTATÍSTICA

20 DE OUTUBRO DE 2010

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DE LISBOA
EDIFÍCIO C3 | ANFITEATRO 3.2.14

www.ceaul.fc.ul.pt

Programa

14.30 - 15.00 Sessão de Abertura

Prof. Carlos Braumann, Presidente da SPE
Prof. João Sentieiro, Presidente da FCT
Drª Alda Carvalho, Presidente do INE
Prof. Pinto Paixão, Director da FCUL
Profª Maria Antónia Amaral Turkman, Coordenadora do CEAUL

15.00 -16.00 Palestra convidada

Prof. Anthony O'Hagan, University of Sheffield, UK
"Statistical methods for cost- effective health care"

16.00 - 18.00 Átrio do Edifício C6 da FCUL

Apresentação de painéis da responsabilidade de Grupos de Investigação em Estatística e outras áreas científicas.
PORTO DE HONRA

Inscrições:
E-mail: ceaul@fc.ul.pt
Data limite: 15 de Outubro de 2010
A inscrição, embora gratuita, é obrigatória por motivos de organização.
Será dado um certificado de participação.

Centro de Estatística e Aplicações
Universidade de Lisboa

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

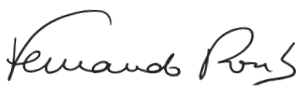
Editorial

… sobre a Estatística em 2010; em ano bom …

1. O International Statistical Institute (ISI) completou 125 anos de existência, no passado dia 24 de Junho de 2010. De acordo com a história, tudo começou em 1853 com uma primeira série de congressos internacionais de Estatística que constituíram um ponto de partida e a motivação suficiente para que trinta anos mais tarde fosse constituído o ISI. Um encontro inicial em Londres juntou estatísticos seniores vindos de organismos estatais de produção de estatísticas e também de academias.
Com os estatísticos seniores - membros eleitos - e com os membros inscritos nas suas 7 secções, o ISI congrega profissionais oriundos de todas as mais importantes universidades e institutos de investigação, das diversas agências governamentais e ainda de algumas especializadas organizações internacionais e comerciais.
A existência do ISI inclui, quase na totalidade, a história da Sociedade Portuguesa de Estatística. Na realidade, a SPE desde o início se ligou aos projectos internacionalistas daquele Instituto, principalmente através do seu líder inicial e grande promotor da Estatística em Portugal. O Prof Tiago de Oliveira desde muito cedo inscreveu a SPE nos projectos do ISI.
A íntima ligação da SPE com o ISI foi aprofundada com a reestruturação levada a cabo no início da década de 90. A SPE tornou-se membro institucional do ISI em 1994. Actualmente, a nível individual, Portugal tem 14 membros eleitos. A Espanha 45 e o Brasil 18.
No passado mais recente, a SPE teve uma forte afirmação no âmbito das fundamentais actividades do ISI com o qual, empenhadamente, colaborou na organização da 56ª Sessão do Congresso Internacional bienal que reuniu em Lisboa um número recorde de estatísticos de todo o mundo. Culminou assim uma participação a nível institucional e também através de muitos dos seus membros mais activos que integraram as diversas comissões (Cf. *Boletim SPE Outono de 2007*, para mais detalhes sobre essa actividade).
No âmbito do ISI existem várias Comissões especializadas. De entre elas podemos salientar aquelas que, mais de perto, se dedicam às questões específicas da Estatística em determinadas zonas do mundo. Depois das comissões da Ásia Oriental e da região Árabe está a ser dinamizada uma Comissão para a América Latina protagonizada por alguns colegas brasileiros. É uma boa notícia, desde logo pela proximidade de tantos estatísticos portugueses com estreita ligação ao Brasil e à nossa congénere ABE - Associação Brasileira de Estatística. Deste elã, pelas notícias sabemos, já surgiu como fruto a realização do Congresso Internacional do ISI no Rio de Janeiro em 2015, após Dublin 2011 e Hong-Kong 2013.
2. Como sabemos, (Cf. Notícia no *Boletim SPE Primavera de 2010 e o cartaz em destaque neste Boletim Outono de 2010*) no próximo dia 20 de Outubro comemora-se o I Dia Mundial da Estatística. Para além da simbologia de tal efeméride é fundamental salientar-se a importância da sua criação, isto é, de um culminar de crescimento e reconhecimento da Estatística e da sua projecção aos mais diversos níveis da sociedade actual. Várias actividades se programam e de que daremos notícias no próximo *Boletim* SPE.
3. Um passo final para um ano bom: j SPE - a secção de Jovens Estatísticos da SPE (Cf. Notícia neste *Boletim*). Esta é mais uma excelente iniciativa da SPE e que, obviamente, necessita do maior empenho dos seus membros: os juniores e os seniores.
Apenas com uma activa colaboração de todos esta iniciativa de jovens e para jovens terá o seu melhor sucesso!
O *Boletim* SPE dá todo o apoio e espera publicar boas notícias o mais breve possível.

Conforme sugestão recebida, na contra – capa deste *Boletim* inclui-se uma página com um Índice Geral de autores e respectivos textos.

O tema central do próximo *Boletim* será *Sondagens e Censos.*



Mensagem do Presidente

Caros Colegas:

Realizou-se nas Termas de S. Pedro do Sul, de 29 de Setembro a 2 de Outubro de 2010, o nosso XVIII Congresso Anual. A Comissão Organizadora do XVIII Congresso, presidida pelo nosso Colega Paulo Oliveira, da Universidade de Coimbra, e vice-presidida pela nossa Colega Carla Henriques, do Instituto Politécnico de Viseu, fez um excelente trabalho que muito contribuiu para o sucesso do Congresso. Foi uma inauguração auspiciosa da nova modalidade de organização conjunta dos Congressos da SPE por uma instituição universitária e uma instituição politécnica, a assinalar a importância da cooperação entre os dois sistemas de ensino superior no desenvolvimento do ensino e da investigação em Estatística. Tivemos cerca de 180 participantes, que apresentaram 88 comunicações orais e 43 comunicações em poster. Tivemos cinco conferências convidadas de alto nível, um excelente minicurso do Colega Carlos Tenreiro e o correspondente manual "Uma Introdução à Estimação Não-Paramétrica da Densidade", a entrega do Prémio SPE ao Colega Gonçalo Jacinto e a sua apresentação do trabalho premiado, uma sessão organizada pela Sociedade Italiana de Estatística que dá continuidade ao intercâmbio entre os congressos das duas sociedades que iniciámos o ano passado, bem como a entrega dos Prémios Estatístico Júnior 2010 (com o apoio da Porto Editora) e a apresentação em poster dos trabalhos vencedores, uma escolha difícil perante uma participação record. O interessante programa social deu também um excelente contributo para o sucesso deste Congresso. A todos os que participaram, colaboraram ou patrocinaram o XVIII Congresso e as actividades que nele decorreram, o nosso muito obrigado.

No Congresso, foi ainda apresentada pelos seus coordenadores (Paulo Canas Rodrigues e Miguel de Carvalho) a nova Secção dos Jovens Estatísticos na SPE, a jSPE, veículo para que os jovens estatísticos desempenhem um papel de relevo no futuro da sociedade.

Os "Selected Papres" do XVII Congresso, realizado em Sesimbra em 2009, vão ser brevemente publicados pelo grupo editorial Springer numa colaboração entre este grupo e as Sociedades Portuguesa, Italiana, Espanhola e Francesa de Estatística. Os nossos agradecimentos a todos os intervenientes neste Congresso, que foi mais um grande sucesso.

A colaboração internacional entre sociedades estatísticas europeias está em franco desenvolvimento, tendo havido um encontro de presidentes ou representantes de 14 sociedades estatísticas europeias em Paris há poucos dias atrás. Nessa reunião, em que a SPE foi representada pela Colega Luísa Loura, foram decididas algumas formas de cooperação, designadamente nem termos de informação mútua e de uma proposta de tratamento idêntico dos sócios nas inscrições dos Congressos (a SPE já aprovou esta proposta). Outras formas de cooperação irão ser estudadas.

Contamos com a vossa participação no evento comemorativo do Dia Mundial da Estatística que a SPE organiza conjuntamente com o Centro de Estatística e Aplicações da Universidade de Lisboa a 20 de Outubro de 2010 e onde será feita uma apresentação da investigação que decorre nos centros de investigação portugueses que trabalham na área da Estatística.

E, não tarda muito, teremos o XIX Congresso. O Instituto Superior Técnico e o Instituto Politécnico de Leiria são os anfitriões, sendo a Comissão Organizadora presidida pelo Colega António Pacheco, com a Colega Alexandra Seco na vice-presidência. Será na Nazaré e esperamos vê-los todos lá.

Até ao Boletim da Primavera de 2011, com uma saudação muito cordial.

Notícias

• XVIII Congresso SPE

Termas de S. Pedro do Sul: Mente sã em corpo são!

Cumprindo um ritual de finais de Setembro, a comunidade estatística volta a reunir-se para uma profícua troca de conhecimento e convívio, sob o mote de mais um Congresso Anual da SPE, desta feita o XVIII.

Na noite de 28 de Setembro, rumamos ao interior centro de Portugal e eis-nos em terra de Lafões, mais precisamente nas Termas de S. Pedro do Sul, cujas nascentes de águas termais brotando a 68,7ºC são a força motriz da região. De facto, o poder terapêutico destas águas é reconhecido faz mais de 2000 anos. Já os romanos, grandes apreciadores de banhos quentes, no século I da era cristã edificaram neste local um *Balneum* (denominação da qual proveio a palavra Banho), do qual ainda podem ser apreciados vestígios. Séculos mais tarde, sob as suas ruínas, D. Afonso Henriques manda edificar um novo balneário e concede foral à *Vila do Banho* (nome por que ficou conhecido o local após a retirada dos romanos), elevando-a desta forma a concelho, mais tarde a "Couto do Reino" e, posteriormente a "Couto de Honra". Reza a história que, em 1169, o Rei Fundador, aproveitando as propriedades sulfúreas, sódicas, carbonatadas e silicatadas das águas das Caldas Lafonenses, ter-se-á vindo nelas curar da fractura na perna direita que sofreu após uma retirada precipitada da Batalha de Badajoz.

Desde então as termas ganharam fama e foram frequentadas por nobres, plebeus e ilustres figuras da corte portuguesa. Destas destaca-se a estadia nas termas do Rei D. Manuel I, e da Rainha D. Amélia, a fim de curarem os seus males. Em 1500, aquando da estada do Rei Venturoso na terra, o balneário sofre várias alterações e passa a chamar-se Hospital Real das Caldas de Lafões. Este é substituído séculos mais tarde, por um novo balneário mandado construir pela câmara municipal em 1884. Por Decreto Real de 1895, assinalando a passagem da soberana Rainha D. Amélia nas termas, o nome da soberana é atribuído ao novo balneário (o qual subsiste até à data) e a terra passa a denominar-se Caldas da Rainha D. Amélia, nome que passou à denominação actual de Termas de S. Pedro do Sul aquando do advento da República. Mais recentemente, em 1987, inaugura-se um novo e moderno balneário termal – o balneário D. Afonso Henriques – e procede-se à restauração do balneário Rainha D. Amélia. Não admira que os dois balneários da cidade, actualmente em perfeitas condições, tenham os nomes destes personagens tão ligados à história da região.

Mas porque veio este ano o congresso até às termas? Estará a comunidade estatística doente ou a envelhecer? Nada disso, como notaremos no decorrer deste relato!

Comecemos por dar uma olhadela à Comissão Organizadora Local deste Congresso, este ano pela primeira vez a cargo de duas instituições: a Universidade de Coimbra (UC) e o Instituto Politécnico de Viseu (IPV). Se há quem se lembre (nós não que à data ainda nem suspeitávamos que iríamos entrar nestas lides da estatística), há 16 anos atrás a UC organizou o II Congresso Anual da SPE. Imaginam onde? Numas termas, claro! Nas do Luso mais precisamente. E, longe de ter sido uma banhada, reza a história que esse foi um dos mais animados Congressos da SPE com "direito" a banhos de piscina de madrugada! Estaria a sempre jovem Comissão Organizadora Local, presidida pelo colega Paulo Eduardo Oliveira (UC), e composta ainda pelas colegas Carla Henriques (IPV), Cristina Martins (UC), Maria da Graça Temido (UC) e Madalena Malva (IPV), a reservar-nos algo do género? Se algo se passou este ano desconhecemos, pelo menos nada nos foi feito chegar pelas más línguas…

Um bem haja à Comissão Organizadora Local pela excelente escolha do Hotel do Parque, nas Termas de S. Pedro do Sul, onde, sob um sol ameno de Outono, nos foi proposto trabalhar desfrutando da calma e beleza deslumbrante das serras envolventes e das águas límpidas do rio Vouga.

No dia 29 de Setembro, iniciámos os trabalhos com o mini-curso "Uma Introdução à Estimação Não-Paramétrica da Densidade", ministrado pelo colega Carlos Terreiro (UC) que nos convenceu não haver razão para preferirmos o estimador do histograma ao estimador do núcleo. Terminado o mini-curso, seguiu-se a habitual Sessão de Abertura do Congresso, com votos de boas vindas e de um proveitoso congresso proferidos pelos colegas Paulo Oliveira (Presidente da Comissão Organizadora), Maria Paula de Oliveira Carvalho (Vice-Presidente do Instituto Politécnico de Viseu), Carlos Braumann (Presidente da SPE) e Nazaré Lopes (Departamento de Matemática da FCTUC). Soubemos que este ano o congresso conta com a participação de colegas da Sociedade Italiana de Estatística, no âmbito do acordo de cooperação entre as duas sociedades. Recordou-se ainda o II Congresso Anual da SPE e o que a SPE evoluiu desde essa data.

Refira-se que o congresso deste ano contou com a presença de cerca de 180 participantes, cujo contributo activo se traduziu num programa com 87 comunicações orais, divididas em 28 sessões e 42 comunicações em poster, divididas em 4 sessões. Isto sem contar com as 5 sessões dos oradores convidados.

Num dia dedicado à estimação não–paramétrica, era chegada a hora de assistir à primeira Sessão Plenária, intitulada "Nonparametric Regression with Nonparametrically Generated Covariates" sob a responsabilidade do Professor Enno Mammen (Universidade de Mannheim - Alemanha).

Terminamos o dia brindados com o Lafões de Honra e com uma interessante visita guiada ao Balneário Rainha D. Amélia. Partimos da fonte próxima da nascente onde, munidos dos nossos sentidos, podemos constatar a quentura e composição sulfúrea destas águas (o pivete a enxofre não deixou margem para dúvidas…). Procedemos então para o Museu do Balneário no qual observamos curiosos aparelhos usados nos antigos tratamentos termais. Seguiu-se a passagem de um filme no auditório e uma lição exaustiva sobre os benefícios destas águas termais, proferida pelo Dr. Aires Mendes Leal (Director da Estância Termal). Mas o momento alto da visita, que exigia calçado a rigor, ainda estava por vir. Animados, visitamos as instalações onde se realizam os tratamentos, munidos de caricatos sapatinhos azuis de todos os tamanhos e feitios, que quase lançavam moda.

Cheios de energia, iniciámos o segundo dia ascendendo ao 5º piso onde decorreriam as sessões de comunicações orais. Do leque de interessantes temas oferecidos nas 4 sessões paralelas matinais, difícil foi decidir quais delas escolher. Seguiu-se uma Sessão de Posters, dominada pelos extremos e na qual, analisando os trabalhos expostos, aproveitamos para repor o teor de cafeína. Após nova ronda de comunicações orais, teve início a segunda Sessão Plenária, subordinada ao tema "Clustering Longitudinal Multivariate Observations", da responsabilidade do colega Maurizio Vichi (Universidade La Sapienza – Roma), presidente da Sociedade Italiana de Estatística. No início da palestra, Maurizio Vichi referiu a recente cooperação entre as duas sociedades e o acordo de várias sociedades europeias de estatística na publicação de uma série internacional de Actas dos Congressos.

Após um merecido almoço, era chegada a hora do desejado passeio. Desta vez, a escolha entre as duas alternativas possíveis já havia sido feita em Julho.

Num dos passeios rumou-se a Viseu. Saídos do autocarro, uma curta e agradável viagem de funicular levou-nos ao largo da Sé. Aí, visitámos o afamado Museu Grão Vasco, instalado no antigo Paço Episcopal mesmo adossado à Sé Catedral. Sob o olhar atento de *São Pedro*, nele pudemos apreciar obras-primas do grande pintor quinhentista Vasco Fernandes (Grão Vasco) e de seus colaboradores e

contemporâneos da Escola de Viseu. Prosseguimos com a visita à Sé Catedral e seus claustros, terminando o passeio com uma prova de vinhos na Adega de Lafões.

No outro passeio, visitámos as Caves da Murganheira. Aí, à medida que penetrávamos nas caves escavadas em rocha granítica azul, nas quais se divisavam garrafas de espumante empilhadas, foram-nos desvendando os segredos da fabricação do seu champanhe. Conhecido o processo de fabrico, passamos a aferir da qualidade do mesmo submetendo o espumante a um apertado processo de prova. De volta ao autocarro, rumámos à Igreja do Convento de Tarouca onde fomos cordialmente recebidos pelo Sr. António Caetano, defensor acérrimo deste monumento. Natural de S. João de Tarouca, após 12 anos de trabalho como alfaiate em Lisboa voltou à terra onde começou a sua luta contra o abandono a que tinha sido sujeito o Convento e a Igreja. Verdadeiro historiador deste mosteiro, informou-nos que este foi o primeiro mosteiro da Ordem de Cister da Península Ibérica, cuja primeira pedra foi colocada por D. Afonso Henriques no dia 1 de Julho de 1152. O interior da igreja é um verdadeiro museu, do qual se destacam o retábulo de D. Pedro (filho bastardo de D. Dinis) da escola de Grão Vasco, o sarcófago de D. Pedro, e os azulejos no altar alusivos à fundação do mosteiro.

De regresso às termas, e já com o cansaço a apertar, ainda houve tempo para apreciar, já por nossa conta e risco, alguma doçaria de eleição da região.

Chegava a sexta-feira. Como de costume, o dia mais intenso do congresso que contava com 16 sessões de comunicações orais, 2 de posters e 2 convidadas. Novidade foi a Sessão 14, toda ela proferida em Inglês, onde um cluster de colegas da SIEDS (Sociedade Italiana de Economia, Demografia e Estatística) apresentou os seus trabalhos.

A manhã terminou com a Sessão Plenária do colega Mário Figueiredo (Instituto de Telecomunicações – IST) que nos falou sobre "Inferência de Redes a Partir de Co-ocorrências".

Pouco passava das 14h30m, quando retomámos os trabalhos. Desta feita, optamos pela Sessão de Modelos Espaciais, sessão com uma adesão tal que foram necessárias mais cadeiras e uma nova ocupação espacial da sala.

Depois da Sessão de Posters, foi a vez do colega Wenceslao González-Manteiga (Universidade de Santiago de Compostela – Espanha) apresentar a Sessão Plenária da tarde, intitulada "General views of the goodness-of-fit tests for statistical models. Applications in finance and environmental problems".

Sem aviso prévio, deparamo-nos com a surpreendente notícia que a foto de grupo ia ser tirada ali mesmo na sala. Sem tempo para retocar a maquilhagem, e desprovidos da nossa fatiota de gala, fizemo-nos o melhor que pudemos à foto de grupo. Pena foi que alguns colegas só souberam da foto após esta ter decorrido! Faltar às plenárias tem destas coisas…

Depois de um dia em cheio, ganhámos o direito a um bom jantar, para o qual não tivemos de ir longe. O momento era de festa e todos estavam com a boa disposição de quem anda nestas lides dos congressos SPE. Após uns aperitivos servidos na entrada do Hotel do Parque, passamos ao salão onde, num acolhedor ambiente à luz de velas, fomos presenteados com a actuação do Orfeão do Instituto Politécnico de Viseu. Dirigidos pelo expressivo maestro Nuno Garrido, professor desta escola, belas vozes interpretaram um repertório nas duas línguas mais faladas no congresso, Italiano e Português. A actuação terminou em grande com um "O Sole Mio" no qual se destacou a belíssima voz do Diniz Coutinho, aluno do IPL. Após o jantar ainda houve tempo para um pé de dança, não nosso que estávamos ainda atarefadas a dar luta às sobremesas, mas de alunos de uma escola de tango.

No quarto e último dia, acordar cedo foi o que mais custou. Não fossem as onze comunicações orais divididas por quatro sessões - Estimação, Aplicações, Distribuições e Modelos - e a confortável cama do hotel teria levado a melhor.

Passou-se de imediato à última Sessão de Posters na qual, entre os demais trabalhos, foram ainda expostos os trabalhos vencedores dos Prémios Estatístico Júnior 2010.

Seguiu-se a sessão reservada à recém-criada Secção de Jovens Estatísticos da SPE - a jSPE - estabelecida com o objectivo de promover, cultivar e incentivar um intercâmbio de informação e conhecimento entre estatísticos em início de carreira. Discursando para a audiência maioritariamente sénior que resistiu até ao final do congresso, Paulo Canas Rodrigues e Miguel de Carvalho apelaram à adesão de Jovens Estatísticos à jSPE, e à divulgação da mesma por parte dos Estatísticos Seniores, incentivando os seus alunos a integrarem e contribuírem para este projecto.

Procedeu-se à entrega do Prémio SPE, este ano atribuído a Gonçalo Jacinto (Universidade de Évora) com o trabalho intitulado "Modelação dos tempos de duração de rotas em redes móveis Ad-Hoc", que nos falou da utilidade destas redes em diferentes cenários e da sua modelação através de processos de Markov determinísticos por troços.

Na última Sessão Plenária deste congresso, intitulada "A transição censitária em Portugal", Fernando Casimiro (Director do gabinete dos censos do INE) prendeu a atenção de todos, palestrando com profundo conhecimento sobre a importância dos censos, do seu enquadramento a nível internacional, das práticas censitárias em Portugal até 2011 e das alterações impostas pelas recomendações da UNECE.

O congresso terminou com a Sessão de Encerramento, onde merecidos agradecimentos foram proferidos aos colegas da Organização Local pelo congresso que nos proporcionaram e onde Paula Brito, Russel Alpizar-Jara e Luisa Loura entregaram os prémios Estatístico Júnior 2010.

Finalmente, passou-se o testemunho aos organizadores do próximo Congresso SPE, que será fruto da cooperação entre o Instituto Superior Técnico e do Instituto Politécnico de Leiria. Em nome da Comissão Organizadora Local, os colegas António Pacheco (IST) e Alexandra Seco (IPL), desvendaram que de 28 de Setembro a 1 de Outubro de 2011, nos levarão de volta ao litoral, à bela cidade da Nazaré, onde, se o tempo ajudar, poderemos desfrutar da bela vista de mar do hotel Miramar Sul. Esperemos que tais distracções não nos levem pelos caminhos aleatórios das areias dessas praias durante o decorrer dos trabalhos científicos.

Sandra Dias, Fátima Ferreira e Irene Oliveira

(DM-UTAD)

• Sobre o novo glossário estatístico inglês-português

Encontra-se nas suas últimas fases o processo de actualização do glossário inglês-português de termos estatísticos pela Comissão Luso-Brasileira constituída no seio da Sociedade Portuguesa de Estatística (SPE) e Associação Brasileira de Estatística (ABE).
Este processo consistirá nomeadamente na eliminação do glossário corrente de entradas dispensáveis e adição de um número significativo de novos verbetes. O glossário resultante tomará já como referência o novo acordo ortográfico da Língua Portuguesa e incluirá notas explicativas que clarificam o sentido de algumas traduções e as opções seguidas pela referida comissão perante itens de tratamento menos incontroverso. Além disso, este documento será complementado com uma listagem de acrónimos, de uso mais ou menos generalizado, e seu significado. Uma novidade do maior interesse respeita ao suporte prioritário de divulgação do novo glossário, baseado numa mais aliciante e amigável plataforma informática, embora se pense ainda disponibilizá-lo numa versão impressa a quem estiver interessado.
Não é demais recordar que a utilidade deste precioso instrumento para a comunidade de estatísticos e tradutores técnicos varia na razão directa da participação de seus utilizadores através de suas sugestões e interrogações dirigidas a qualquer das comissões *ad-hoc* constituídas no âmbito da SPE e ABE.

O Presidente da Comissão Especializada de Nomenclatura Estatística (CENE) da SPE

Carlos Daniel Paulino

• jSPE – a Secção de Jovens Estatísticos

A Sociedade Portuguesa de Estatística tem o prazer de anunciar a criação da jSPE – a Secção de Jovens Estatísticos da SPE. A jSPE foi estabelecida com o objectivo de promover, cultivar e incentivar um intercâmbio de informação e conhecimento entre estatísticos em início de carreira. Não é obrigatório ser "jovem" para pertencer à jSPE, basta somente ser "jovem" neste interesse.
Além da Direcção da jSPE será criado um Comité de Representantes com cerca de 15 Jovens Estatísticos de diferentes Universidades/Politécnicos/Empresas. Este órgão permitirá uma divulgação mais vasta da jSPE e dos eventos organizados pela Secção perante circuitos académicos e profissionais.
Para que este projecto seja um sucesso é necessário o apoio de todos os sócios da SPE. Aos Jovens Estatísticos apelamos que se juntem à jSPE. Aos Estatísticos Seniores, bem como a Seniores de áreas afins, pedimos que divulguem a informação e que incentivem os vossos alunos a integrarem e contribuírem para este projecto.

Caso pretendam fazer parte da jSPE e/ou do Comité de Representantes não hesitem em contactar-nos.

Pela Comissão Instaladora,
Paulo Canas Rodrigues, *paulocanas@gmail.com*
Miguel de Carvalho, *mb.carvalho@fct.unl.pt*

• JOCLAD 2011

Caro(a) colega
De 6 a 9 de Abril de 2011, a Universidade de Trás-os-Montes e Alto Douro (UTAD) recebe as XVIII Jornadas de Classificação e Análise de Dados (JOCLAD 2011).
Informamos que o prazo limite para a submissão de trabalhos é 23 de Janeiro de 2011.
Mais informações estarão brevemente disponíveis em *www. joclad2011.utad.pt*.
Contamos com a sua presença nas Jornadas!!

Pela Comissão Organizadora Local,
Fátima Ferreira

• Prémios "Estatístico Júnior 2010"

A atribuição de prémios "Estatístico Júnior 2010" é promovida pela Sociedade Portuguesa de Estatística, com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística. Neste ano recebemos 17 candidaturas na categoria Ensino Básico, envolvendo um total de 38 alunos, e 38 candidaturas na categoria Ensino Secundário, envolvendo um total de 102 alunos. Não recebemos candidaturas dos Cursos de Educação e Formação de Adultos (EFA), modalidade que foi considerada pela primeira vez nos prémios deste ano.

A cerimónia de entrega dos Prémios Estatístico Júnior 2010, conforme estipulado no Regulamento, decorreu na Sessão de Encerramento do XVIII Congresso Anual da Sociedade Portuguesa de Estatística, no dia 2 de Outubro de 2010, às 12h30 horas, nas instalações do Hotel do Parque nas Termas de São Pedro do Sul.

O Júri foi constituído pelos professores: Doutora Maria Eugénia Graça Martins (Presidente) e Doutora Luísa Canto e Castro de Loura do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa e Doutor Russell Alpizar-Jara do Departamento de Matemática da Universidade de Évora.

No final deste Boletim são apresentados os premiados

A Direcção

• 2ª Competição Internacional de Literacia Estatística

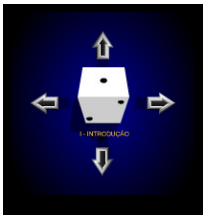
POSTER COMPETITION ISLP 2ª Competição Internacional de Literacia Estatística 2010-2011

A 2ª Competição Internacional de Literacia organizada pelo ISLP (International Statistical Literacy Project) já começou. O objectivo da Competição Internacional de Literacia Estatística é aumentar o reconhecimento da Estatística entre professores e alunos, bem como promover o desenvolvimento recursos de ensino e aprendizagem em todo o mundo.
Mais informações (em várias línguas e em português) em:
http://www.stat.auckland.ac.nz/~iase/islp/competition-second

Pedro Campos

• Modelos de Probabilidade no ALEA

O curso das noções de probabilidade no ALEA tem um novo capítulo: Modelos de probabilidade discretos e contínuos.
Utilizando exemplos, ilustrações e aplicações interactivas, são estudados neste capítulo alguns modelos de probabilidade discretos (modelo uniforme, modelo binomial, modelo geométrico e modelo Poisson) e alguns modelos de probabilidade contínuos (modelo uniforme e modelo normal).



Consulte já em: *http://www.alea.pt/html/probabil/html/probabilidades.html*

Pedro Campos

• A SPE no "45th Scientific Meeting of the Italian Statistical Society"

A Sociedade Italiana de Estatística (SIS), fundada em 1939, promove bianualmente um encontro científico internacional, em que convida Sociedades congéneres de Países da União Europeia. O "45th *Scientific Meeting of the Italian Statistical Society*" (SIS 2010) decorreu em Pádua, Itália, de 16 a 18 de Junho. Várias Sociedades Europeias de Estatística responderam ao convite da SIS para integrarem o programa científico, entre as quais a Sociedade Portuguesa de Estatística (SPE). A convite do Presidente da SPE, Professor Carlos Braumann, avançámos então com a organização de uma sessão temática, ainda acessível em http://homes.stat.unipd.it/mgri/SIS2010/Program/13-SSXIII_SPS/.

A organização e presidência da sessão, intitulada "*Statistics of Extremes in Today's World*", ficaram a cargo de M. Ivette Gomes (Universidade de Lisboa), e foram os seguintes os conferencistas convidados e os temas em discussão:

- M. Isabel Fraga Alves (Universidade de Lisboa, Faculdade de Ciências), com a comunicação intitulada "*How Far can Man go*?", em co-autoria com Laurens de Haan e Cláudia Neves.
- Ana Ferreira (Universidade Técnica de Lisboa, Instituto Superior de Agronomia), com a comunicação intitulada "*Spatial Extremes and High Quantile Estimation of Aggregated Rainfall*".
- Lígia Henriques Rodrigues (Instituto Politécnico de Tomar), com a comunicação intitulada "*Refined Estimation of a Light Tail: an Application to Environmental Data*", em co-autoria com M. Ivette Gomes e Frederico Caeiro.

A língua oficial do encontro foi o Inglês, regra essa seguida por todos os oradores na generalidade. Para além de quatro conferências plenárias, proferidas por James J. Heckman (Universidade de Chicago, USA), George Casella (Universidade de Flórida, USA), Paul Boyle (Universidade de St. Andrews, UK) e Enrico Giovannini (ISTAT, Roma), pudemos assistir a várias das vinte sessões convidadas, em temas diversificados, de que mencionamos unicamente as sessões organizadas por Sociedades Estatísticas Europeias: a sessão da Sociedade Portuguesa de Estatística, intitulada "*Statistics of Extremes in Today's World*", foi organizada por M. Ivette Gomes (Universidade de Lisboa), a sessão da "Royal Statistical Society", intitulada "*A Non-random Sample of Statistics in the U.K.*", foi organizada por Alan Kimber (Universidade de Southampton), a sessão da "Societé Française de Statistique", intitulada "*A Short Glance on French Statistic*", foi organizada por Avner Bar-Hen (Universidade de Paris 5), a sessão da "Sociedad de Estadística e Investigación Operativa", intitulada "*Skewing Symmetry 25 Years on*", foi organizada por Arthur Pewsey (Universidade da Estremadura), e a sessão da "Deutsche Statistische Gesellschaft", intitulada "*Copulas: Theory and Applications*", foi organizada por Friedrich Schmid (Universidade de Colónia).

Foram três dias repletos de um programa científico deveras interessante, num ambiente de saudável e alegre convívio entre os participantes, numa cidade pequena e deveras simpática, sob os auspícios do Santo António de Lisboa, que afinal também pode ser considerado de Pádua.

M. Ivette Gomes e M. Isabel Fraga Alves

A Geoestatística e o Ambiente: um acompanhamento contínuo

Raquel Menezes, *rmenezes@math.uminho.pt*

Universidade do Minho

1. Enquadramento

Na actualidade, a Estatística Espacial ocupa um lugar importante, graças ao desenvolvimento tecnológico que permite a fácil obtenção de dados espaciais. Para além da existência das fontes tradicionais de dados espaciais, tal como mapas, material recolhido nos censos, fotos aéreas, surgem agora novas fontes com excelente fiabilidade, nomeadamente com os dados obtidos por satélite. Estas novas tecnologias encontram-se associadas à disponibilidade de uma maior capacidade computacional e software específico, tal como software de processamento de imagem e sistemas de informação geográfica.

Os Modelos Espaciais trabalham com dados recolhidos em distintas localizações espaciais. Estes modelos medem a relação entre observações obtidas em diversos locais, devendo representar a noção intuitiva da existência de uma correlação entre dados situados na proximidade espacial. Tipicamente, esta correlação diminui com o aumento da distância. Estes modelos devem igualmente reflectir a presença de erros espacialmente correlacionados. A análise de correlação espacial permite então observar a forma como variáveis, tal como leituras de algum indicador de poluição, obtidas em pontos distintos do espaço se relacionam. O conhecimento destas relações é particularmente relevante de um ponto de vista prático, uma vez que permite estimar valores para as localizações nas quais não foram efectuadas medições.

Uma amostra de dados pode ser formada com base em observações obtidas em uma, duas ou três dimensões. Medições da qualidade da água efectuadas ao longo de um percurso fluvial são unidimensionais. Por outro lado, medições de pluviosidade e outras variáveis meteorológicas são obtidas em pontos particulares, mas colectivamente formam um campo aleatório bidimensional. Por último, medições de concentrações de minerais no solo são por vezes tratadas como problemas tridimensionais dado que podem ocorrer em diversas profundidades.

Um processo espacial é um processo estocástico. Pode ser representado por um conjunto de variáveis aleatórias (ou vectores) $Y(x)$, indexado sobre x definido num conjunto $A \subset R^d$, um espaço euclidiano d-dimensional com $d = 1,2,3$. A sua notação usual é $\{Y(x) : x \in A\}$. Considerem-se as localizações espaciais $\{x_1,\cdots,x_n\}$, então $Y(x_1),\cdots,Y(x_n)$ identificam dados observados nestas localizações. Estas observações podem ser obtidas a partir de uma, ou mais, variáveis discretas ou contínuas.

De acordo com Cressie (1993), a natureza de A permite identificar três processos espaciais principais, nomeadamente processos reticulados, processos pontuais e processos contínuos. Estes últimos são geralmente referidos como Geoestatística.

Nos processos contínuos, ao contrário dos processos reticulados, entre quaisquer dois pontos espaciais associados a observações existentes, existe sempre um outro ponto onde a variável aleatória poderia ser observada. Os processos reticulados permitem modelar dados agregados por áreas; enquanto que, no caso dos processos pontuais, o principal foco de interesse científico é a própria localização.

A Geoestatística refere-se, então, ao ramo da Estatística Espacial para o qual os dados consistem numa amostra finita de valores relacionados com um fenómeno espacialmente contínuo. Exemplos incluem:

altitudes num levantamento topográfico; medições de poluição à custa de uma rede finita de estações de monitorização; caracterização de propriedades do solo à custa de uma amostra representativa; e contagens de insectos num conjunto de localizações pré-seleccionadas.
A Geoestatística nasceu, na década de 60, da necessidade de modelização de recursos geológicos como fenómenos espaciais (Matheron, 1963). Inicialmente associada à escola Francesa *Centre de Geostatistique de Fontainebleau - École des Mines* fundada por G. Matheron, tem na sua curta história conhecido várias fases de evolução impulsionadas pelas particularidades dos diferentes domínios de aplicação. Actualmente, a aplicação da Geoestatística abrange diversas áreas disciplinares, nomeadamente geologia petrolífera, hidrogeologia, meteorologia, oceonagrafia, geoquímica, geografia, agricultura, ecologia e ciências ambientais e da saúde.

2. Aplicações Ambientais

Na sociedade moderna, existe uma consciencialização cada vez mais forte dos problemas ambientais que temos de enfrentar, estando estes tipicamente associadas à intervenção do homem sobre a natureza. Algumas das principais causas estão relacionadas, por exemplo, com o aquecimento global, o aumento do nível dos rios e oceanos, a desarborização das florestas, a poluição dos solos, águas e ar, podendo originar preocupações de saúde ambiental. Torna-se, então, necessário monitorizar o que se passa no Ambiente que nos rodeia, coleccionando dados que permitam compreender o cenário de mudança. Por conseguinte, torna-se muito importante descrever formalmente o Ambiente através de modelos plausíveis, de modo a analisar e interpretar os dados recolhidos, fundamentando a tomada de decisões (ver por exemplo Barnett, 2003; Barnett & Turkman, 1993, 1994, 1997; e Barnett, Stein & Turkman, 1999).
A Geoestatística é uma ferramenta essencial para o trabalho desenvolvido pelos cientistas do ambiente e investigadores afins (Webster & Oliver, 2007). As condições climáticas variam de local para local, as características dos solos variam dependendo da região de abrangência em que estes são examinados, e até atributos da responsabilidade do homem, tal como a poluição, variam de acordo com a localização. Idealmente, as técnicas adoptadas da Geoestatística são ajustadas às necessidades dos investigadores, que as utilizam para fazer predição à custa dos dados recolhidos, permitindo-lhes planear futuros *surveys* quando os recursos são limitados.

2.1. Amostragem Preferencial

As redes de monitorização de indicadores ambientais, em particular indicadores de poluição, são muitas vezes construídas de forma gradual não sistemática respondendo a necessidades que surgem ao longo do tempo, não sendo geralmente claras as razões da escolha das localizações dos monitores. Os monitores podem ser colocados onde os níveis de poluição são esperados ser mais baixos ou elevados, no primeiro caso melhorando as hipóteses da legislação em vigor ser cumprida, e no segundo caso monitorizando o pior cenário. Qualquer uma das situações anteriores é caracterizada por um processo de amostragem denominado por preferencial.
Os métodos geoestatísticos tradicionais, ver por exemplo Cressie (1993) e Diggle & Ribeiro (2007), assumem que as localizações amostradas são seleccionadas independentemente dos valores da variável espacial debaixo de estudo. A Amostragem Preferencial surge quando o processo associado às localizações dos dados e o processo objecto de modelação são estocasticamente dependentes. Em Diggle, Menezes & Su (2010), é proposta uma abordagem baseada ao modelo para este problema, demonstrando-se através de exemplos simulados que ignorar esta dependência pode afectar negativamente a estimação de parâmetros (para a covariância e média do processo aleatório) e a capacidade de predição.

Passe-se, então, à breve apresentação deste modelo. Considere-se que os dados amostrados são da forma $(x_i, y_i) : i = 1, \cdots, n$, onde $x_1, \cdots, x_n$ identificam as localizações dos dados dentro da região de observação $A \subset R^2$ e $y_1, \cdots, y_n$ as medições associadas a estas localizações, e assuma-se que:

- y_i é uma realização de $Y_i = Y(x_i)$, onde $Y(x)$ com $x \in A$ é denominado *processo de medição*;
- x_i são realizações do processo estocástico X denominado *desenho amostral* (note-se que este pressuposto exclui um desenho determinístico, como localizações sobre uma grelha regular);
- e $S(x)$ com $x \in A$ é o *processo espacialmente contínuo não-observado*, que se pretende modelar.

Vejamos o exemplo apresentado em Diggle *et al* (2010) sobre bio-monitorização de poluição na região da Galiza em Espanha. Os dados em causa foram obtidos em 1997, resultando da análise da concentração de metais pesados em musgo. Suspeita-se que tenha ocorrido amostragem preferencial, uma vez que a maioria dos dados se encontram mais próximos das fontes de poluição por questões de redução de custos. Neste exemplo, A identifica a região da Galiza, x_i as localizações do musgo, y_i as medições do indicador de poluição (por exemplo, chumbo) e $S(x)$, ou mais precisamente $\mu(x)+S(x)$, a superfície de poluição espacialmente contínua que se pretende estimar. Um modelo simples para estes dados pode assumir a forma

$$Y_i = \mu(x_i) + S(x_i) + Z_i, \quad i = 1,\cdots,n \qquad (1)$$

onde Z_i são variáveis aleatórias mutuamente independentes com média zero. Caso se considere $E[S(x)] = 0 \quad \forall x$, então $E[Y_i] = \mu(x_i)$ envolvendo possíveis variáveis explicativas na média do processo, exemplos como distância à indústria mais próxima, algum índice de industrialização, ou a elevação do terreno.

Suponha-se que, por simplificação da notação, [.] significa "distribuição de", $S = \{S(x) : x \in R^2\}$ e $Y = \{Y_1,\cdots,Y_n\}$. Então, tradicionalmente, o modelo (1) trata X como sendo determinístico e tem a estrutura $[S,Y] = [S][Y \mid S(X)] = [S][Y_1 \mid S(x_1)]\cdots[Y_n \mid S(x_n)]$ (ver Diggle, Tawn & Moyeed, 1998). Adicionalmente, debaixo do pressuposto da gaussianidade, no modelo (1), $[Y_i \mid S(x_i)]$ seguem distribuições Gaussianas univariadas com média $\mu(x_i)+S(x_i)$ e variância comum τ^2.

O modelo para Amostragem Preferencial, proposto em Diggle *et al* (2010), envolve a distribuição conjunta de S, X e Y, assumindo-se $[S,X] \neq [S][X]$. Por conseguinte, o modelo completo tem a forma

$$[S,X,Y] = [S][X \mid S][Y \mid S(X)], \qquad (2)$$

uma vez que $[Y \mid X,S]$ se pode reduzir à dependência funcional $[Y \mid S(X)]$.

Uma classe específica de modelos é definida através dos seguintes pressupostos adicionais:

1. S é um processo estacionário Gaussiano com média 0, variância σ^2 e função de correlação $\rho(u;\phi) = corr\{S(x_i),S(x_j)\}$ para qualquer x_i e x_j afastados uma distância u (ϕ está relacionado com a distância a partir da qual deixa de haver correlação espacial).
2. Condicionado a S, X é um processo de Poisson não-homogéneo com intensidade $\lambda(x) = \exp\{\alpha + \beta S(x)\}$. (3)
3. Condicionado a S e X, Y é um conjunto de variáveis Gaussianas independentes com Y_i seguindo $N(\mu(x_i) + S(x_i), \tau^2)$.

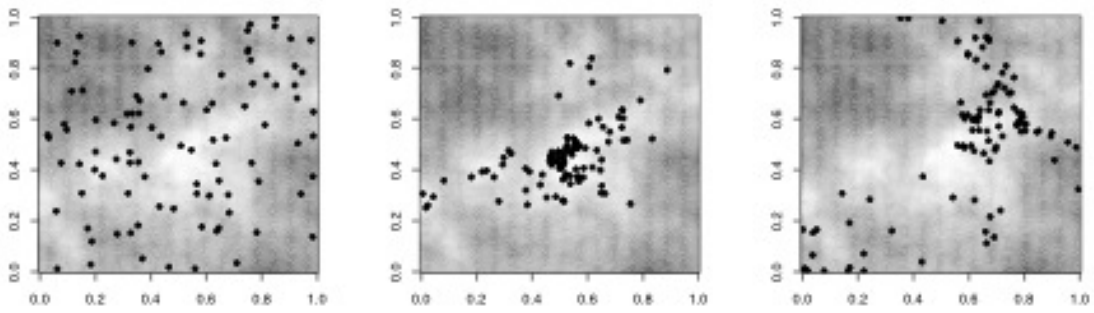


Figura 1. Localizações amostradas e processo espacialmente contínuo $S(x)$ (numa escala de cinzentos). Do painel mais à esquerda para a direita, tem-se um desenho completamente aleatório, uma amostra preferencial e uma amostra formando *clusters*, respectivamente.

O tipo de desenho amostral representado no painel da esquerda da Figura 1 pode ser obtido assumindo-se $\beta = 0$ na equação (3). Por outro lado, valores positivos do parâmetro β estarão associados a uma maior concentração de pontos amostrados onde o campo subjacente $S(x)$ apresenta valores mais elevados, assim como, valores negativos de β associados a valores mais baixos de $S(x)$.

O problema da amostragem preferencial (ver exemplo simulado para $\beta = 2$, no painel do centro da Figura 1) é por vezes confundido com o problema de desenho amostral formando *clusters* (ver painel da direita da Figura 1). No entanto, no último caso não existe dependência estocástica entre os processos X e S, sendo este tipo de desenho amostral por vezes justificado pela necessidade de melhor caracterizar a variabilidade espacial a curtas distâncias, ou mesmo por factores externos ao processo S, tal como a existência de sub-áreas de interesse por questões demográficas ou de acessibilidade geográfica.

Em Menezes, Garcia-Soidán & Febrero-Bande (2008), propõe-se um estimador de variograma tipo-núcleo robusto a *clusters*, tendo em conta um peso inverso à densidade numa dada vizinhança como compensação para as áreas menos populadas. Debaixo de isotropia, esta proposta pode ser representada por

$$\hat{\gamma}(u) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(u)[Y(x_i) - Y(x_j)]^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(u)}, \quad u \geq 0, \qquad (4)$$

onde $w_{ij}(u) = \frac{1}{\sqrt{n_i n_j}} \times K\left(\frac{u - \|x_i - x_j\|}{h}\right)$ e $n_i = \sum_k I_{\{\|x_i - x_k\| \leq \delta\}}$, sendo $K(.)$ e $I_{\{.\}}$ as funções núcleo e indicatriz, respectivamente. Note que n_i representa o número de pontos que caem dentro do círculo de raio δ e centro x_i. A ideia é ajustar a contribuição de qualquer ponto amostrado x_i de acordo com o seu número de vizinhos n_i; como o estimador do variograma soma contribuições de pares de valores observados nos pontos (x_i, x_j), o produto $n_i n_j$ representa a contribuição conjunta do par; a raiz quadrada é adoptada como corrector de escala.

O estimador (4) requer a selecção de dois valores: h o parâmetro janela da função núcleo e o raio de vizinhança δ. O primeiro é obtido minimizando o respectivo erro quadrático médio e o segundo resulta da análise da estimação da densidade das distâncias entre pontos na região de observação. Este estimador de variograma goza de boas propriedades, tal como ser assimptoticamente centrado e consistente (Menezes *et al*, 2008).

Em Menezes & Tawn (2009), é analisada a possível associação entre o processo de medição Y e as respectivas localizações espaciais, debaixo de um desenho amostral multi-etapa. O exemplo de motivação diz respeito a níveis de radioactividade medidos em Rongelap, uma ilha do Oceano Pacífico usada para testar armas nucleares nos anos 50. Os dados foram recolhidos em duas etapas, debaixo de um desenho amostral uniforme e formando *clusters*, o que poderá afectar as conclusões de uma análise standard que não tenha em conta estas características. Neste trabalho propõe-se um método para detectar amostragem multi-etapa enviesada, na qual os valores observados em etapas posteriores dependem dos valores observados em etapas anteriores.

2.2. Outros Contributos Não-Paramétricos

Conforme já referido, a necessidade de reconstruir um fenómeno sobre toda a região de observação, à custa de um conjunto finito de dados, pode ocorrer num largo espectro de áreas, nomeadamente nas ciências ambientais. Na maioria dos casos, apenas um única realização está disponível em cada uma das localizações espaciais observadas, por conseguinte é comum exigir-se pressupostos adicionais sobre o processo aleatório envolvido para permitir prosseguir com a inferência. A estacionaridade é um pressuposto típico e, debaixo desta condição, diferentes técnicas têm sido desenvolvidas para a predição espacial.

As abordagens não-paramétricas podem ser directamente aplicadas com o objectivo de predição, tal como é proposto em Menezes, Garcia-Soidán & Ferreira (2010). Neste trabalho, apresenta-se o seguinte preditor núcleo não-paramétrico,

$$\hat{Y}(x;h) = \frac{\sum_{i=1}^{n} w_i Y(x_i)}{\sum_{i=1}^{n} w_i},$$

onde $w_i = K_d((x - x_i)/h)$, K_d representa uma função núcleo d-dimensional, h o parâmetro janela e x é a nova localização onde se pretende obter um valor de predição. Debaixo do pressuposto de um desenho amostral estocástico para as localizações, prova-se que o erro quadrático médio do preditor tende para negligível com o aumento do tamanho amostral. O cálculo da janela óptima teórica obriga, no entanto, à estimação de quantidades desconhecidas. Consequentemente, são sugeridas técnicas por validação-cruzada para a selecção de uma janela empírica quer local quer global.
Uma opção alternativa para a predição, bastante adoptada na prática, passa por prosseguir via as técnicas de *kriging*, que produzem predições precisas debaixo de diversos pressupostos; ver, por exemplo, Christakos (1992) e Cressie (1993). Em particular, a implementação dos preditores *kriging* exige que a estrutura de segunda ordem seja adequadamente estimada.

Tipicamente, as funções de covariância ou variograma (Menezes, Garcia-Soidán & Febrero-Bande, 2005) disponibilizam uma medida para a dependência espacial, quando se assume estacionaridade de segunda ordem ou intrínseca do processo aleatório, respectivamente. Numa primeira etapa, os métodos não-paramétricos podem ser utilizados para a aproximação das funções anteriores, incluindo o estimador tradicional proposto em Matheron (1963) ou alternativas mais robustas, apresentadas em Cressie & Hawkins (1980) e em Genton (1998). Mais recentemente, propostas tipo-núcleo aparecem em Garcia-Soidán *et al* (2004) e Menezes *et al* (2008), para dados uniformemente distribuídos ou formando *clusters* na região de observação, respectivamente. A validade dos vários estimadores anteriores, contudo, não é obrigatoriamente satisfeita, uma vez que estes podem não obedecer às propriedades de condicionalmente definido-positivo ou negativo dependendo de se tratar da função de covariância ou variograma, respectivamente. Consequentemente, estes estimadores não podem ser directamente adoptados nas equações de *kriging*.
O problema anterior poderá ser ultrapassado pela selecção de um modelo paramétrico apropriado e, de seguida, pelo cálculo de estimativas óptimas dos parâmetros; embora, o problema da incorrecta especificação do modelo seja a principal desvantagem desta abordagem. A adequabilidade de um modelo paramétrico pode ser avaliada através de diagnóstico gráfico ou através de testes de ajuste de bondade como o proposto em Maglione & Diblasi (2004) debaixo da gaussianidade do processo aleatório.
Alternativamente, para a obtenção de uma estimação válida pode se recorrer a algum método não-paramétrico que transforme um estimador em um estimador válido. Hall, Fisher & Hoffman (1994) propõe primeiro truncar e de seguida inverter a transformada de Fourier de um dado estimador, embora não seja sugerida uma estratégia para a escolha óptima do termo de truncagem. Em Garcia-Soidán & Menezes (2007), uma abordagem por séries ortonormais é sugerida para a obtenção de um estimador válido, debaixo de algumas hipóteses tal como a escolha de bases válidas e também daqueles termos que correspondem a coeficientes de Fourier positivos.

2.3. Conclusões

A título de conclusão reforça-se a ideia da importância das aplicações ambientais para os desenvolvimentos recentemente ocorridos na Geoestatística, embora as contribuições daí resultantes possam ser adoptadas em outras áreas de aplicação.
Os exemplos de motivação aqui apresentados, assim como as respectivas soluções, apenas envolveram a referência espacial dos dados. No entanto, a referência temporal pode ser igualmente importante, tendo despoletado num passado recente uma área de investigação bastante promissora, tipicamente denominada por modelação espaço-temporal. Os recentes avanços tecnológicos têm permitido que, em diversos contextos, os dados sejam recolhidos ao longo do espaço e do tempo simultaneamente. Um exemplo particular surge rotineiramente na monitorização ambiental onde medições periódicas (horárias, diárias, ...) são feitas numa rede de estações de monitorização espacialmente distribuídas. Um dos muitos problemas em aberto passa por estender o modelo de Amostragem Preferencial, apresentado na secção 2.1, para um contexto espaço-temporal; tal justifica-se pelo facto do grau de amostragem preferencial se poder alterar ao longo de períodos de tempo, caso as estratégias de amostragem sejam sazonais ou dependentes de motivações políticas.

Referências

Barnett V. (2003). *Environmental Statistics: Methods and Applications*. Wileys's series in Probability and Statistics.

Barnett V., Stein A. & Turkman K.F. (1999). *Statistics for the Environment, Volume 4, Statistical Aspects of Health and the Environment*. John Wiley and Sons, UK.

Barnett V. & Turkman K.F. (1993, 1994, 1997). *Statistics for the Environment: Volumes 1-3*. John Wiley and Sons, UK.

Cressie N. (1993). *Statistics for spatial data*. J. Wiley.

Cressie N. & Hawkins D. (1980). Robust estimation of the variogram. *Mathematical Geology*, 12, 115-125.

Cristakos G. (1992). *Random fields models in earth sciences*. Academic Press. San Diego.

Diggle P. & Ribeiro P.J. (2007). Modelbased geostatistics. Springer series in Statistics.

Diggle P., Menezes R. & Su T.L. (2010). Geostatistical Inference under Preferential Sampling. *Journal of Royal Statistics Society*, series C, vol 59, part 2, 191-232 (RSS read paper, with discussion).

Diggle P., Tawn J. & Moyeed R. (1998). Model-based geostatistics. *Applied Statistics*, 47, 299-350 (with discussion).

Garcia-Soidán P. & Menezes R. (2007). Covariance estimation via the orthonormal series. Proceedings of the 56th Session of the International Statistics Institute 2007, Lisboa.

Garcia-Soidán P. González-Manteiga W. & Febrero-Bande M. (2004). Local linear regression estimation of the variogram. *Statistics and Probability Letters*, 64, 169-179.

Genton M. (1998). Highly robust variogram estimation. *Mathematical Geology*, 30, 213-221.

Hall P., Fisher I. & Hoffmann B. (1994). On the nonparametric estimation of covariance functions. *Annals of Statistics*, 22, 2115-2134.

Maglione D.S. & Diblasi A.M. (2004). Exploring a valid model of the variogram of an intrinsic spatial process. *Stochastic Environmental Research and Risk Assessment*, 18, 366-376.

Matheron G. (1963). Principles of Geostatistics. *Economic Geology*, 58, 1246-1266.

Menezes R., Garcia-Soidán P. & Febrero-Bande M. (2005). A comparison of approaches for valid variogram achievement. *Computational Statistics*, 20 (4), 623-642.

Menezes R., Garcia-Soidán P. & Febrero-Bande M. (2008). A kernel variogram estimator for clustered data. *Scandinavian Journal of Statistics*, vol 35, issue 1, 18-37.

Menezes R. & Tawn J. (2009). Assessing the effect of biased and clustered sampling on variogram estimation. *Environmetrics*, vol 20, issue 4, 445-459.

Menezes R., Garcia-Soidán P. & Ferreira C. (2010). Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics*, vol 20, issue 4, 445-459.

Webster R. & Oliver M.A. (2007). *Geostatistics for Environmental Scientists*, 2nd Edition. Wiley.

Processos Max-Estáveis: uma Caracterização Simples e Exemplos

Ana Ferreira, *anafh@isa.utl.pt* Laurens de Haan, *ldhaan@ese.eur.nl*

ISA – UTL e CEAUL *Erasmus University Rotterdam, Tilburg University e CEAUL*

1. Introdução

A teoria de valores extremos para processos estocásticos tem re-surgido recentemente com algum impacto, bastante proporcionado pelo interesse actual na modelação espacial. No entanto, ainda se mostra complexa, encontrando-se ainda e de certa forma algo disconexa.

Os processos max-estáveis são os processos que, teoricamente, se justificam os adequados para modelar o comportamento de níveis elevados de processos estocásticos (note-se que estamos interessados em processos contínuos). Logo, ao admitir um processo estocástico a representar determinado fenómeno e.g. no espaço, os processos max-estáveis devem ser a base para a modelação dos extremos no espaço. Por exemplo, encontram-se aplicações destes processos na modelação da precipitação em dada região, em Coles e Tawn (1996) e Buishand *et al* (2008).

No que se segue, caracterizam-se de uma forma simples os processos max-estáveis, i.e. processos limite do máximo de processos independentes e identicamente distribuídos, convenientemente normalizados (Secção 3). Daremos alguns exemplos na Secção 4, em particular o processo limite do máximo de certos processos gaussianos, e os processos móveis mistos. Complementa-se a exposição na Secção 5, com algumas referências adicionais, nomeadamente a abordagens alternativas que se encontram na literatura recente, na modelação de extremos no espaço, bem como algumas aplicações.

2. Definições Elementares

No que se segue, S representa um subconjunto de $\mathbb{R}^d$, $d \in \mathbb{N}$, $C(S)$ o espaço das funções reais e contínuas em S e, $=^d$ e $\xrightarrow{d}$ representam igualdade e convergência em distribuição respectivamente.

Definição 2.1. *Um processo estocástico* $Y = \{Y(s)\}_{s\in S}$ *diz-se* Max-Estável *se, para* $Y_1, Y_2, \ldots, Y_n$ *cópias independentes e identicamente distribuídas (i.i.d.) de* Y*, existirem funções reais contínuas* $A_s(n) > 0$ *e* $B_s(n)$ *tais que,*

$$\left\{ max_{1\le j\le n} \frac{Y_j(s) - B_s(n)}{A_s(n)} \right\}_{s\in S} =^d \{Y(s)\}_{s\in S}, \qquad \text{para todo o } n \in \mathbb{N}. \qquad (2.1)$$

Note-se que o máximo é obtido para cada s separadamente, e que as "constantes" normalizadoras podem depender de s. Assume-se que $Y(s)$, $\forall\, s \in S$, têm distribuição não degenerada, o que implica que as distribuições marginais de Y são necessariamente do tipo de valores extremos. Sem perda de generalidade consideramos que as marginais univariadas são generalizada de valores extremos standartizada, $P(Y(s) \leq x) = \exp\{-(1+\gamma(s)\,x)^{-1/\gamma(s)}\}$, $1+\gamma(s)x > 0$, $\gamma(s) \in \mathbb{R}$. Salienta-se a importância do parâmetro $\gamma(s) \in \mathbb{R}$ que, para cada s, representa o usual índice de valores extremos. Interpretado em termos de função, $\gamma(.)$ deve ser uma função contínua em S.

Definição 2.2. *O processo*

$$\{Z(s)\}_{s\in S} := \left\{(1+\gamma(s)\,Y(s))_+^{1/\gamma(s)}\right\}_{s\in S}, \qquad (2.2)$$

com Y de (2.1), é um processo Max-estável simples*, no sentido de as suas distribuições marginais serem Fréchet unitárias, $P(Z(s) \leq x) = exp(-1/x)$, $x > 0$, $\forall\, s \in S$. Obviamente que (2.1) verifica-se para Z e, note-se, com $A_s(n) = n$ e $B_s(n) = 0$, $\forall\, s \in S$.*

É conveniente, na definição seguinte, considerar processos $X = \{X(s)\}_{s\in S} \in C(S)$, isto é, $X(s)$ está bem definido para todo o $s \in S$ e, a função $X(s)$ é uma função contínua quase certamente.

Definição 2.3. *$X = \{X(s)\}_{s\in S} \in C(S)$ diz-se pertencer ao* domínio de atração de *$Y = \{Y(s)\}_{s\in S}$ se, para $X_1, X_2, \ldots$ cópias i.i.d. de X e, $a_s(n) > 0$ e $b_s(n)$ funções reais contínuas,*

$$\left\{max_{1\leq j\leq n}\frac{X_j(s)-b_s(n)}{a_s(n)}\right\}_{s\in S} \xrightarrow{d} \{Y(s)\}_{s\in S}, \qquad n \to \infty, \qquad (2.3)$$

em $C(S)$, admitindo que $Y(s)$ é não degenerada, $\forall\, s \in S$.

Note-se que, nas aplicações o que mais naturalmente se observa não é o processo max-estável, mas processos que se supõem estar no domínio de atracção de algum processo max-estável.
É consequência das definições anteriores que o processo limite em (2.3) é necessariamente um processo max-estável. Na Secção 4 refere-se o caso dos processos gaussianos.
Para uma exposição mais detalhada dos conceitos anteriores, nomeadamente do domínio de atracção, refere-se de Haan e Lin (2001) e de Haan e Ferreira (2006).

3. Processos Max-Estáveis: uma caracterização

Um modo simples de introduzir processos max-estáveis é recorrer à teoria de valores extremos univariada, da seguinte forma menos habitual. Assim, começaremos por discutir o caso univariado e posteriormente passaremos ao caso funcional, com recurso a propriedades elementares dos processos pontuais de Poisson. Por clareza da exposição introduzem-se sucintamente estes processos.

Definição 3.1. *Um* processo pontual de Poisson *em $E \subset \mathbb{R}^d$, de* intensidade (ou média) *$\nu(.)$, é uma medida pontual aleatória, que denotaremos por $N_V(.) := \sum_{i=1}^{\infty}\delta_{V_i}(.)$, $V_i \in E$, verificando: (i) $N_V(A)$ é uma variável aleatória (v.a.) com distribuição de Poisson, de média $\nu(A)$, para todo o Boreliano $A \subset E$, (ii) $N_V(A_1), N_V(A_2), \ldots, N_V(A_k)$ são v.a.´s independentes sempre que $A_1, A_2, \ldots, A_k \subset E$ são Borelianos disjuntos dois a dois, $\forall\, k \in \mathbb{N}$.*

Proposição 3.1. *(*Representação de uma v.a. com distribuição Max-Estável). *Considere-se o processo pontual de Poisson $N_V(.) := \sum_{i=1}^{\infty}\delta_{V_i}(.)$ em $(0,\infty)$ e com medida de intensidade dv/v^2. A v.a.,*

$$\tilde{V} := max_{1\leq i<\infty} V_i \qquad (3.1)$$

tem distribuição (max-estável) Fréchet unitária.

Daremos duas demonstrações deste resultado. Uma primeira mais directa e natural. A segunda mais instrutiva e introdutória ao que se segue relativamente aos processos max-estáveis.

Demonstração 1. Calculando directamente a distribuição de probabilidades de $\widetilde{\mathrm{V}}$. Para $x > 0$,

$$P\left(\widetilde{\mathrm{V}} \le x\right) = P(V_i \le x, \text{ para todo o } 1 \le i < \infty) = P(\mathrm{N_V}(x,\infty) = 0) = \exp\left(-\int_x^\infty \frac{dv}{v^2}\right) = \exp\left(-\frac{1}{x}\right),$$

de acordo com a distribuição de Poisson. Assim, $\widetilde{\mathrm{V}}$ tem distribuição Fréchet unitária i.e., de valores extremos ou max-estável. ■

Demonstração 2. Verificando que a distribuição de $\widetilde{\mathrm{V}}$ é max-estável via (2.1). Ou seja, que

$$max_{1\le j\le k}\widetilde{\mathrm{V}}_j =^d k\widetilde{\mathrm{V}}, \qquad \text{para todo o } k \in \mathbb{N}, \tag{3.2}$$

onde $\widetilde{\mathrm{V}}_1, \widetilde{\mathrm{V}}_2, \ldots$ denotam cópias i.i.d. de $\widetilde{\mathrm{V}}$.

Considere-se em primeiro lugar o lado esquerdo de (3.2). Cada $\widetilde{\mathrm{V}}_j$ envolve a realização do processo pontual $\mathrm{N_{V_j}}$, i.e. $\widetilde{\mathrm{V}}_j = max_{1\le i<\infty} V_{i,j}$. Assim, no total tem-se o máximo de entre todos os elementos de k processos pontuais i.i.d. Um vez que $\sum_{j=1}^{k} \mathrm{N_{V_j}}$, com $\mathrm{N_{V_j}}$ processos i.i.d., resulta num processo pontual de Poisson com medida de intensidade dada pela soma das medidas de intensidade individuais, de facto, do lado esquerdo de (3.2) tem-se, o máximo de entre todos os pontos de um processo pontual de Poisson com medida de intensidade $k\, dv/v^2$.
Considere-se agora o lado direito de (3.2). Tem-se $max_{1\le i<\infty} kV_i$ ou, o máximo dos pontos do processo pontual $\mathrm{N}_{k\mathrm{V}}$. Sendo,

$$P(\mathrm{N}_{k\mathrm{V}}(a,b) = r) = P\left(\mathrm{N_V}\left(\frac{a}{k},\frac{b}{k}\right) = r\right) = e^{-q}\frac{q^r}{r!}, \quad \text{onde } q = \int_{a/k}^{b/k} \frac{dv}{v^2} = \int_a^b \frac{k}{v^2}\, dv\,.$$

Logo, $\mathrm{N}_{k\mathrm{V}}$ é um processo pontual de Poisson com medida média $k\, dv/v^2$. Como processos pontuais com iguais medidas médias (finitas) são iguais em distribuição, (3.2) verifica-se. ■

Veremos agora a generalização do resultado anterior para processos max-estáveis.

Proposição 3.2. (*Estrutura dos processos Max-Estáveis simples). Considere-se um processo pontual de Poisson* $N_V(.) := \sum_{i=1}^{\infty} \delta_{V_i}(.)$ *definido em* $(0,\infty)$ *e com medida de intensidade* dv/v^2. *Independentemente de* $N_V(.)$, *sejam* $W_1, W_2, \ldots$ *processos estocásticos i.i.d. contínuos em S, verificando* $W_1(s) \ge 0$, $EW_1(s) = 1$, $\forall\, s \in S$, *e* $E sup_{s\in S} W_1(s) < \infty$. *O processo estocástico*

$$Z(s) \coloneqq max_{1\le i<\infty} V_i\, W_i(s), \qquad s \in S \tag{3.3}$$

é max-estável simples.

Demonstração. Sejam Z_1 , Z_2 , ... cópias i.i.d. de Z. Tal como na segunda demonstração da Proposição 3.1 tem-se,

$$max_{1\le j\le k}\, Z_j = \; max_{1\le j\le k} \; max_{1\le i<\infty}\, V_{i,j}\, W_{i,j}(s), \quad s \in S, \; k \in \mathbb{N}, \tag{3.4}$$

onde, $\{V_{i,j}\}_{i,j}$representam os pontos de um processo pontual de Poisson obtido da junção de pontos provenientes de processos pontuais de Poisson independentes. Assim, (3.4) tem a mesma estrutura que (3.3), mas com medida de intensidade diferente, igual a $k\, dv/v^2$.

De seguida, considere-se $k\,Z$. Novamente pelo raciocínio na segunda demonstração da Proposição 3.1, $k\,Z = max_{1\le i<\infty} V_i^*\, W_i(s)$ onde V_i^* são os pontos do processo de Poisson com medida de intensidade $k\,dv/v^2$. Logo, $max_{1\le j\le k}\, Z_j =^d k\,Z$, i.e. Z é max-estável simples. ■

Nota: As condições $EW_1(s) = 1,\ \forall\, s \in S$, e $Esup_{s\in S} W_1(s) < \infty$ garantem, por um lado que as marginais univariadas são Fréchet unitárias e, que o processo max-estável é finito quase certamente em intervalos limitados, como se pode ver das distribuições marginais de dimensão finita. Estas, podem obter-se do seguinte modo: para cada $s_1, \dots, s_n \in S,\ \ x_1, \dots, x_n > 0, \forall\, n \in \mathbb{N}$,

$$P(Z(s_1) \le x_1, Z(s_2) \le x_2, \dots, Z(s_n) \le x_n)$$

$$= P(max_{1\le i<\infty} V_i\, W_i(s_1) \le x_1, max_{1\le i<\infty} V_i\, W_i(s_2) \le x_2, \dots, max_{1\le i<\infty} V_i\, W_i(s_n) \le x_n)$$

$$= P\left(V_i \le min_{1\le j\le n} \frac{x_j}{W_i(s_j)},\ \text{para todo o}\ 1 \le\ i < \infty\right)$$

$$= \exp\left(-\iint_{min_{1\le j\le n}\, x_j/W_1(s_j)}^{\infty} \frac{dr}{r^2}\, dF_{W_1}\right) = \quad \exp\left(-E_{W_1}\left(max_{1\le j\le n}\ \frac{W_1(s_j)}{x_j}\right)\right).$$

Inversamente, prova-se que todos os processos simples max-estáveis admitem a representação (3.3) (Corolário 9.4.5 em de Haan e Ferreira (2006)). Desta forma, (3.3) caracteriza os processos max-estáveis simples. A distribuição de probabilidades de Z é determinada pela distribuição de probabilidades de W_1, designada de distribuição espectral. Em particular, a estacionaridade de W_1 implica a estacionaridade de Z.

4. Exemplos de Processos Max-Estáveis

4.1. O caso Gaussiano

Processos max-estáveis que se têm revelado bastante interessantes são, quando se toma para o processo W em (3.3) alguns processos Gaussianos específicos. Por exemplo, dos mais simples mas já suficientemente rico é o dado por (Brown e Resnick, 1977),

$$W(s) \coloneqq e^{B(s)-|s|/2},\quad s \in\ \mathbb{R}, \qquad (4.1)$$

sendo B um movimento Browniano em $\mathbb{R}$. Note-se que o processo max-estável obtido é estacionário, embora W em (4.1) não o seja. Alternativas mais gerais para W são dadas em Kabluchko *et al* (2009), cf. Teorema 2.

No contexto do dominio de atracção, o processo obtido de (3.3) com (4.1) é o processo max-estável limite que se obtém do máximo de certos processos Gaussianos independentes, convenientemente normalizados e com uma transformação de escala no espaço (i.e. em s) conveniente. Note-se que, da teoria de valores extremos multivariada sabe-se que "máximos de v.a.´s normais são assimptóticamente independentes", daí a transformação de escala em s para se obter um limite não trivial. Mais concretamente (Brown e Resnick (1977), Kabluchko *et al* (2009) Teorema 17):

Sejam $X_1, X_2, \dots$ processos gaussianos contínuos e independentes de média zero, variância 1 e função de covariância admitindo uma determinada condição "natural" de variação regular, envolvendo uma função L de variação lenta em 0. O processo,

$$b_n\,(X_i(t_n\,s) - b_n), \qquad s \in D \subset \mathbb{R}^d,$$

com D numa vizinhança de zero, $b_n = (2\log n)^{1/2} - (2\log n)^{-1/2}\left(\left(\frac{1}{2}\right)\log\log n + \log\left(2\sqrt{\pi}\right)\right)$ e $t_n = \inf\{t > 0 : L(t)t^{\alpha} = b_n^{-2}\}$, $\alpha \in (0,2]$, pertence ao domínio de atracção de um processo do tipo (3.3) com (4.1).

Quanto a aplicações, Buishand *et al* (2008) aplicaram uma extensão do modelo obtido de (3.3) com (4.1) para duas dimensões, na modelação de extremos de precipitação no espaço.

4.2. Processos Max-Móveis

Talvez a classe de modelos mais referida na literatura em aplicações de processos max-estáveis, seja a dos *processos max-móveis* (de Haan e Pickands, 1986). São processos que admitem a representação,

$$Z(s) \coloneqq max_{1\le i<\infty} V_i\, f(s - T_i), \qquad s \in \mathbb{R}^d, \tag{4.2}$$

onde: (i) $\{(V_i, T_i)\}$ são relativos ao processo pontual $N_{V,T}(.) := \sum_{i=1}^{\infty} \delta_{(V_i,T_i)}(.)$ em $(0,\infty)\times\mathbb{R}^d$ com medida de intensidade $v^{-2}\,dv\,dt$; (ii) $f: \mathbb{R}^d \to [0,\infty)$ é uma função densidade de probabilidade unimodal contínua.

São processos max-estáveis simples, estacionários e contínuos quase certamente (Balkema e de Haan, 1988). As distribuições marginais de dimensão finita são dadas por: para cada $s_1, \dots, s_n \in S$, $x_1, \dots, x_n > 0$, $\forall\, n \in \mathbb{N}$,

$$P(Z(s_1) \le x_1, Z(s_2) \le x_2, \dots, Z(s_n) \le x_n)$$

$$= P(max_{1\le i<\infty} V_i\, f(s_1 - T_i) \le x_1, max_{1\le i<\infty} V_i\, f(s_2 - T_i) \le x_2, \dots, max_{1\le i<\infty} V_i\, f(s_n - T_i) \le x_n) =$$

$$P\left(V_i \le min_{1\le j\le n} \frac{x_j}{f(s_j - T_i)}, \text{ para todo o } 1 \le i < \infty\right) = \exp\left(-\int_{\mathbb{R}^d}\int_{min_{1\le j\le n}\, x_j/f(s_j-t)}^{\infty} \frac{dr}{r^2}\, dt\right) =$$

$$\exp\left(-\int_{\mathbb{R}^d} max_{1\le j\le n}\, \frac{f(s_j - t)}{x_j}\, dt\right) .$$

O muitas vezes designado de "modelo de Smith (1990)", consiste no modelo (4.2) com f a função densidade de probabilidade normal, por exemplo aplicado na modelação de valores elevados de precipitação em Coles e Tawn (1996). Este modelo encontra-se implementado no programa "Spatial Extremes"- software R, por Ribatet (2009).

Teoricamente, o caso de f normal e outros casos específicos foram analisados em de Haan e Pereira (2006).

5. Referências Adicionais

Referências adicionais para a base teórica de processos max-estáveis são de Haan (1984) e Giné *et al* (1990).

Um exemplo de um processo alternativo aos anteriormente apresentados e recente na literatura é o designado processo M4 ("máximos móveis de máximos multivariados"), resumidamente consistindo em misturas de processos max-moveis discretos (Zhang e Smith, 2004,2009), com aplicações em séries financeiras. Outros exemplos são abordados em Davis e Mikosch (2008) e Davis e Resnick (1993).

Para ainda outras prespectivas refere-se, por exemplo, M.-Shamseldin *et al* (2010) ou, numa prespectiva Bayesiana, e.g. Cooley *et al* (2006), Fawcett e Walshaw (2006), Sang e Gelfand (2010), todas estas com aplicações a dados ambientais.

Obviamente que a lista de referências dada é uma pequena selecção do que se pode encontrar na literatura.

6. Bibliografia

Balkema, A.A. e de Haan, L. (1988) Almost sure continuity of stable moving average processes with index less than one. *Annals of Probability* **16**, 333—343

Brown, B. e Resnick, S. (1977) Extreme values of independent stochastic processes. *J. Appl. Probab.* **14**, 732—739

Buishand, A. de Haan, L. e Zhou, C. (2008) On spatial extremes; with application to a rainfall problem. *Annals of Applied Statistics* **2**, 624—642

Coles, S.G. e Tawn, J.A. (1996) Modelling extremes of the areal rainfall process. *J. R. Statist. Soc*. B **58**, 329—347

Cooley D., Nychka, D. e Naveau Ph. (2006) Bayesian spatial modeling of extreme precipitation return values. *J. Amer. Statist. Assoc*. **102**, 824—840

Davis, R.A. e Mikosch, T. (2008) Extreme value theory for space-time processes with heavy-tailed distributions. *Stochastic Processes and their Applications* **118**, 560—584

Davis, R.A. e Resnick, S.I. (1993) Prediction of stationary max-stable processes. *Annals of Applied Probability* **3**, 497—525

Fawcett, L. e Walshaw, D. (2006) A hierarchical model for extreme wind speeds. *Applied Statistics* **55**, 631—646

Giné E., Hahn M. G. e Vatan P. (1990) Max-infinitely divisible and max-stable sample continuous processes. *Probab. Th. Rel. Fields* **87**, 139—165

Kabluchko, Z., Schlather, M. e de Haan, L. (2009) Stationary max-stable fields associated to negative definite functions. *Annals of Probability* 37, 2042—2065

de Haan, L. (1984) A spectral representation for max-stable processes. *Ann. Prob*. **12**, 1194—1204

de Haan, L. e Fereira, A. (2006) *Extreme Value Theory: An Introduction*. Springer, Boston.

de Haan, L. e Lin, T. (2001) On convergence toward an extreme value distribution in C(0,1). *Annals of Probability* **29**, 467—483

de Haan, L. e Pereira, T.T. (2006) Spatial Extremes: the stationary case. *Annals of Statistics* **34**, 146—168

de Haan, L. e Pickands, J. (1986) Stationary min-stable stochastic processes. *Probab. Th. Rel. Fields* **72**, 477—492

Ribatet, M. (2009) Spatial Extremes: An R package for modeling spatial extremes. Curso: Max-stable processes, theory and practice, CEAUL, FCUL, Lisboa, 14 de Out., 2009.

M.-Shamseldin E., Smith R.L., Sain S.R., Mearns, L.O. e Cooley D. (2010) Downscaling extremes: a comparision of extreme value distributions in point-source and gridded precipitation data. *Annals of Applied Statistics* **4**, 484-502.

Sang, H.Y. e Gelfand, A.E. (2010) Continuous Spatial Process Models for Extreme Values. Journal of Agricultural Biological and Environmental Statistics **15**, 49—65

Smith, R.L. (1990) Max-stable processes and spatial extremes. Unpublished manuscript.

Zhang, Z. e Smith, R.L. (2004) The behavior of multivariate maxima of moving maxima processes. *J. Appl. Probab.* **41**, 1113—1123

Zhang, Z. e Smith, R.L. (2009) On the estimation and application of max-stable processes. *J. Statist. Plann. Inference* 1**40**, 1135—1153

Análise Bayesiana de Modelos Espaço-temporais Aditivos

Giovani Loiola da Silva, *gsilva@math.ist.utl.pt*

Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL)

Departamento de Matemática – IST, Universidade Técnica de Lisboa

1. Introdução

Os métodos estatísticos para a análise de dados espaço-temporais têm aumentado consideravelmente nas últimas décadas devido sobretudo ao avanço de métodos computacionais bayesianos e à acessibilidade aos sistemas de informação geográfica (GIS). A utilização desses métodos está frequentemente ligada ao mapeamento de taxas de doenças que visa a prevenção e monitorização das mesmas. A análise espacial aqui considerada pode ser encarada como uma análise exploratória para descrever visualmente a distribuição espacial de taxas ou proporções sobre uma determinada região ao longo do tempo.

Apesar de modelos lineares generalizados (vide *e.g.* Amaral-Turkman e Silva, 2000) e modelos aditivos generalizados (Hastie e Tibshirani, 1990) poderem ser usados na análise de dados de contagem espaciais, opta-se aqui por adaptações desses modelos, concretizadas em modelos (mistos) com introdução de efeitos aleatórios para ter em conta dados observáveis não independentes, nomeadamente, uma classe de modelos mistos proposta por MacNab e Dean (2001) para capturar padrões espaço-temporais. Os modelos espaciais com efeitos aleatórios são em geral complexos devido quer à grande dimensionalidade do vector de efeitos aleatórios quer à consideração de modelos não gaussianos. Uma forma natural de lidar com esse tipo de modelos é através de modelos hierárquicos bayesianos (vide e,g, Banerjee, Carlin e Gelfand, 2004). Para maiores detalhes sobre inferência bayesiana e modelos espaço-temporais para taxas e proporções, veja-se *e.g.* Paulino, Amaral-Turkman e Murteira (2003) e Silva e Dean (2006), respectivamente.

Numa perspectiva epidemiológica, um objectivo comum ao estudo de uma doença é estimar o risco de um indivíduo contrair uma doença dentro de um determinado período de tempo. A determinação do risco da doença faz-se, por exemplo, com a estimação da taxa ou da proporção de doentes da população. As taxas de incidência de doenças em uma região são em geral representadas em mapas para identificar tendências espaciais (ou temporais) e/ou factores de risco ligados às doenças. Daí o nome mapeamento de taxas de doenças ou simplesmente mapeamento de doenças. O mapeamento de doenças tem sido abordado em vários trabalhos usando diferentes métodos de estimação, nomeadamente métodos bayesianos empíricos, quasi-verossimilhança penalizada (PQL) e Monte Carlo via cadeia de Markov (MCMC) (vide *e.g.* Besag, York e Mollié, 1991; Breslow e Clayton, 1993; Bernardinelli, Clayton e Montomoli, 1995; MacNab eDean, 2001; Lawson, Browne e Vidal Rodeiro, 2003; Silva *et al.*, 2008). Breslow e Clayton (1993) popularizaram o uso de métodos baseados em PQL para inferir em modelos mistos lineares generalizados (GLMM), incluindo um exemplo com o mapeamento de taxas de doenças raras, onde os modelos mistos controlam a variação extra-Poisson através da introdução de efeitos aleatórios. Silva *et al.* (2008) apresentam a abordagem bayesiana de modelos mistos binomiais espaço-temporais, usando métodos MCMC para obter inferências sobre os

parâmetros de interesse e fazer comparação e selecção de modelos. Uma revisão da análise de dados espaciais para mapeamento de doenças encontra-se nomeadamente em Lawson, Browne e Vidal Rodeiro (2003), com o uso de *software* específicos, e Waller e Gotway (2004), com a descrição detalhada dos principais modelos espaciais para dados de Saúde Publica.

2. Modelos espaço-temporais aditivos

Considere uma população em risco de uma região subdividida em *r* sub-regiões e *s* períodos de tempo. Seja Y_{it} o número de "sucessos" (observável) da sub-região i e tempo t em n_{it} "ensaios", $i = 1, \dots, n$, $t = 1, \dots, s$. Em dados de saúde, n_{it} poderia ser o tamanho de uma população de potenciais utilizadores de um serviço de saúde ao nível individual, enquanto Y_{it} seria o número total de indivíduos que utilizaram o tal serviço de saúde durante um determinado período de tempo. Condicional em θ_{it}, supõe-se que Y_{it} tem uma distribuição Binomial com parâmetros n_{it} e θ_{it}, onde θ denota a probabilidade de sucesso. Uma aproximação da distribuição Poisson à distribuição Binomial pode ser apropriada se as taxas de ocorrência do evento de interesse forem pequenas (doenças raras), o que ocorre usualmente em análise de dados espaciais. Nesse caso, condicional em μ_{it}, Y_{it} tem distribuição de Poisson com média $\mu_{it} = n_{it}\,\theta_{it}$.
MacNab e Dean (2001) propuseram uma classe de modelos (Poisson) mistos aditivos generalizados (espaço-temporais), utilizando por conveniência a função de ligação logaritmica, definida por

$$\log \mu_{it} = \alpha_0 + S_0(t) + b_i + S_i(t) \qquad (1)$$

onde α_0 pode representar a taxa média sobre o tempo e a região, $b = (b_1, \dots, b_r)'$ é o vector de efeitos aleatórios espaciais, enquanto $S_0(t)$ e $S_i(t)$ podem ser funções suavizadas arbitrárias de *t*, $i = 1, \dots, n$, $t = 1, \dots, s$.
Todavia, tal aproximação pela distribuição Poisson não é válida para algumas situações na utilização em saúde, onde as proporções de ocorrência do sucesso são moderadas ou mesmo altas (*e.g.* doenças não raras ou contagiosas). Nesse sentido, a classe de modelos (1) pode ser definida em termos do logaritmo das chances, dando origem à seguinte classe de modelos (Binomiais) espaço-temporais (Silva *et al.*, 2008)

$$\operatorname{logit} \theta_{it} \equiv \log(\theta_{it}/1 - \theta_{it}) = \alpha_0 + S_0(t) + b_i + h_i + S_i(t) \qquad (2)$$

onde b_i e h_i denotam aqui componentes dos vectores de efeitos aleatórios espacialmente correlacionados $\boldsymbol{b}$ e não estruturados $\boldsymbol{h}$, respectivamente, $i = 1, \dots, n$. Para esses modelos, pode-se considerar uma base B-*spline* cúbica sem intercepto e um único nó interno quer para $S_0(t)$ quer para $S_i(t)$. Por exemplo, no primeiro caso, $S_0(t) = \sum_{k=1}^{4} \alpha_k B_k(t)$, onde $B_k(t)$ é uma das funções da base B-*spline*, com correspondente coeficiente α_k, $k = 1, \dots, 4$. Note-se que as funções da base B-*spline* são obtidas por um algoritmo recursivo (Silva e Dean, 2006) implementado *e.g.* na rotina *bs* do *software* R. Por vezes, considera-se $S_i(t) = \delta_i t$ em (1) e (2) para estimar o efeito temporal linear δ_i da sub-região i. Isso pode ser particularmente útil na análise exploratória de tendências temporais, onde o período de tempo sob observação é inferior a 10 anos e se a ocorrência do evento ao nível de pequenas áreas não indicar picos ou quedas bruscas.

3. Análise bayesiana

Os métodos espaciais bayesianos de mapeamento de doenças supõem usualmente que as regiões são condicionalmente independentes, dados os parâmetros do modelo, e incorporam subsequentemente a dependência no segundo nível como parte de uma distribuição *a priori* dos parâmetros. O desenvolvimento recente de modelos bayesianos quer empíricos quer hierárquicos em mapeamento de doenças tem como objectivo a obtenção de estimadores estáveis para as taxas de mortalidade em pequenas áreas, usando a informação de todas as sub-regiões para inferir sobre a taxa de mortalidade de cada sub-região. Banerjee, Carlin e Gelfand (2004) fornecem uma revisão detalhada de modelos hierárquicos bayesianos para dados espaciais e espaço-temporais, incluindo a análise de dados espaço-temporais multivariados. Por questão de simplicidade, faz-se agora o desenvolvimento da abordagem

bayesiana somente para os modelos (1) considerando B-*spline* cúbico para $S_0(t)$ e $S_i(t)$, com $S_i(t) = \sum_{k=1}^{4} \beta_{ik} B_k(t)$.

Distribuições *a priori*: Para o vector $\boldsymbol{h}$ de efeitos aleatórios não estruturados espacialmente, considera-se uma distribuição Normal multivariada, com distribuições univariadas (independentes) dadas por $h_i|\sigma_h^2 \sim Normal\ (0, \sigma_h^2)$, onde o hiperparâmetro σ_h^2, controla a quantidade de heterogeneidade adicional, $i = 1, \ldots, n$. Em relação ao vector $\boldsymbol{b}$ de efeitos aleatórios espacialmente correlacionados, considera-se o modelo autoregressivo condicional (CAR) intrínseco (Besag, York e Mollié, 1991), com representação em termos de distribuições condicionais univariadas Normais de médias e variâncias dadas, respectivamente, por

$$E(b_i|\boldsymbol{b}_{-i}, \sigma_b^2) = \sum_{j \in \mathrm{M}_i} \frac{b_j}{m_i} \quad e \quad Var(b_i|\boldsymbol{b}_{-i}, \sigma_b^2) = \frac{\sigma_b^2}{m_i} \tag{3}$$

onde M_i denota o conjunto de índices dos vizinhos da sub-região i e m_i o número dos seus vizinhos, com σ_b^2 a controlar a quantidade de variação espacial. Para os coeficientes das funções B-*spline* cúbicas consideram-se frequentemente distribuições *a priori* Normais independentes, *i.e.*, $\alpha_k|v_k^2 \sim N\ (0, v_k^2)$ e $\beta_{ik}|\sigma_k^2 \sim N\ (0, \sigma_k^2)$. Por fim, para as componentes de variâncias atribui-se uma distribuição *a priori* Gama invertida *IG(c,d)* com parâmetro de forma *c* e parâmetro de escala *d*. Ou seja, $\sigma_b^2 \sim IG(c_1, d_1)$, $\sigma_h^2 \sim IG(c_2, d_2)$, $\sigma_k^2 \sim IG(c_{3\mathrm{k}}, d_{3\mathrm{k}})$ e $v_k^2 \sim IG(c_{4\mathrm{k}}, d_{4\mathrm{k}})$, $k = 1, \ldots, 4$. Na prática, as distribuições *a priori* dos coeficientes B-*spline* e das componentes de variância são geralmente escolhidas, numa base não informativa, com elevada dispersão mas mantendo uma natureza própria.

Distribuições *a posteriori*: A partir da verosimilhança Poisson associada à classe de modelos (1), com B-*spline* cúbico para $S_0(t)$ e $S_i(t)$, e das distribuições *a priori* referidas acima, obtêm-se a correspondente distribuição conjunta *a posteriori* que é proporcional a

$$\begin{aligned}
&\prod_{i=1}^{n} \prod_{t=1}^{s} \left[e^{-\mu_{it} + y_{it} \log(\mu_{it})}\right] \times \left(\frac{1}{\sigma_b^2}\right)^{\frac{n}{2}+c_1+1} \mathrm{e}^{-\frac{1}{\sigma_b^2}\left(d_1 + \sum_{i=1}^{n} \frac{n_i (b_i - \bar{b}_i)^2}{2}\right)} \\
&\times \left(\frac{1}{\sigma_b^2}\right)^{\frac{n}{2}+c_1+1} \mathrm{e}^{-\frac{1}{\sigma_b^2}\left(d_1 + \sum_{i=1}^{n} \frac{n_i (b_i - \bar{b}_i)^2}{2}\right)} \times \left(\frac{1}{\sigma_h^2}\right)^{\frac{n}{2}+c_2+1} \mathrm{e}^{-\frac{1}{\sigma_h^2}\left(d_2 + \sum_{i=1}^{n} \frac{h_i^2}{2}\right)} \\
&\times \prod_{k=1}^{4} \left(\frac{1}{v_k^2}\right)^{\frac{1}{2}+c_{3k}+1} \mathrm{e}^{-\frac{1}{v_k^2}\left(d_{3k} + \frac{\alpha_k^2}{2}\right)} \left(\frac{1}{\sigma_k^2}\right)^{\frac{n}{2}+c_{4k}+1} \mathrm{e}^{-\frac{1}{\sigma_k^2}\left(d_{4k} + \sum_{i=1}^{n} \frac{\beta_{ik}^2}{2}\right)} \times \pi(\alpha_0)
\end{aligned} \tag{4}$$

onde $\mu_{it} = e^{\log n_{it} + \alpha_0 + \sum_{k=1}^{4} \alpha_k B_k(t) + b_i + h_i + \sum_{k=1}^{4} \beta_{ik} B_k(t)}$ e $\pi(\alpha_0)$ é uma distribuição a priori (vaga mas própria) para α_0.

É difícil de lidar com a distribuição *a posteriori* (4) visto que não é possível obter explicitamente as suas distribuições *a posteriori* marginais. Não obstante, essas distribuições marginais podem ser avaliadas usando métodos MCMC (vide e.g. Chen, Shao e Ibrahim, 2000). Em particular, a amostragem Gibbs trabalha com amostras geradas iterativamente para cada parâmetro a partir da sua distribuição *a posteriori* condicional completa, obtida da distribuição conjunta (4) fixados todos os outros parâmetros do modelo (Silva e Dean, 2006). A implementação desses métodos não é difícil em muitos pacotes estatísticos com possibilidade de programação. Alternativamente, a amostragem de Gibbs através do *software* GeoBUGS (Thomas *et al*., 2004) poderia ser usada. Nesse caso, podem-se obter estimativas dos parâmetros de interesse do modelo em causa como, *e.g*., o risco relativo espacial da sub-região i, definida aqui por $e^{b_i + h_i}$, $i = 1, \ldots, n$. Entretanto, chama-se a atenção para o facto de a implementação prática dessas técnicas requerer um certo cuidado neste cenário, visto que os modelos vigentes envolvem muitos parâmetros que são fracamente identificados.

4. Ilustração

Os dados desta ilustração dos modelos espaço-temporais aditivos (1) reportam-se a crianças, com idade inferior a 1 ano e viva à nascença, no Brasil entre 1991 e 2003. A quantidade y_{it} denota o

número de mortalidade infantil observado no i-ésimo estado brasileiro (sub-região) e no ano t, enquanto n_{it} o correspondente total de crianças, $i = 1, \ldots, 27$, $t = 1, \ldots, 13$ (fonte: DATASUS - Ministério da Saúde, Brasil).

A Tabela 1 lista alguns modelos aditivos (1) que foram considerados úteis na análise desses dados espaço-temporais, tendo um nível crescente de complexidade, *e.g.*, o modelo *M2* é o modelo de tendência linear simples *M1* com os efeitos aleatórios estruturados espacialmente (b_i) e não estruturados (h_i). O modelo *M3* considera B-*spline* para $S_0(t)$, enquanto o modelo *M4* adopta B-*spline* para $S_0(t)$ e $S_i(t)$. A fim de comparar estes modelos em competição, calcularam-se as suas medidas sumárias de ajustamento *Desvio* e *DIC* (Spiegelhalter *et al*., 2002), sendo p_D uma estimativa do número de parâmetros do modelo que penaliza a sua complexidade. Com base nesses valores, o modelo seleccionado é *M4*, evidenciando um efeito temporal global não linear $e^{\alpha_0 + S_0(t)}$ e um efeito espacial $e^{b_i + h_i}$ nas taxas de mortalidade infantil, representados graficamente nas Figuras 1 e 2, respectivamente. Note-se que amostras de tamanho 10000 foram obtidas pelos métodos MCMC para todos os modelos, tomando o 5° valor de cada iteração da sequência simulada, após 5000 iterações do período de aquecimento. Um estudo de convergência das amostras simuladas foi realizado usando alguns métodos diagnósticos implementados no *software* BOA (Smith, 2007). Nenhum desses indicou cuidados especiais sobre as quantidades simuladas.

Modelo definido pelo $\log \mu_{it}$	p_D	*Desvio*	*DIC*
$M1$: $\alpha_0 + \beta\, t$	1.99	82623	82625
$M2$: $\alpha_0 + \beta\, t + \delta_i\, t + b_i + h_i$	27.93	25392	25420
$M3$: $\alpha_0 + S_0(t) + \delta_i\, t + b_i + h_i$	56.99	10143	10200
$M4$: $\alpha_0 + S_0(t) + S_i(t) + b_i + h_i$	134.59	5980	6114

Tabela 1: Medidas de comparação e seleção dos modelos espaço-temporais (1).

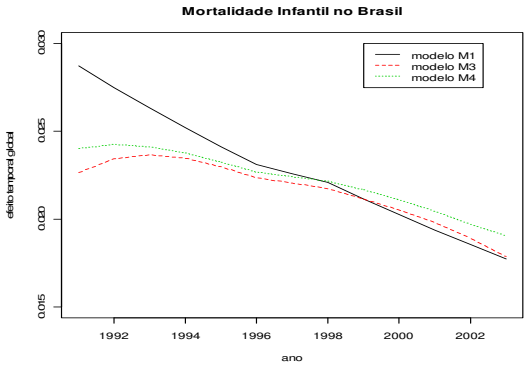


Figura 1: Estimativas do efeito temporal global nas taxas de mortalidade infantil no Brasil, 1991-2003

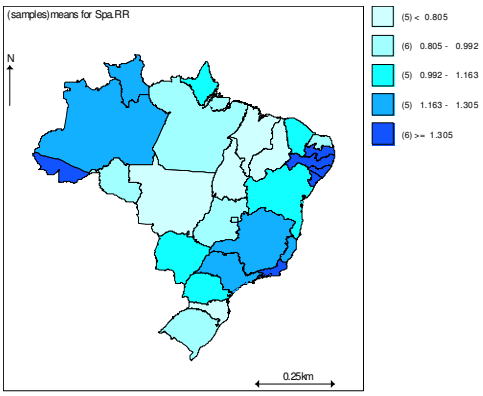


Figura 2: Mapa dos efeitos espaciais (direita) nas taxas de mortalidade infantil no Brasil, 1991-2003

Apesar de haver efeito temporal global decrescente nesse período (veja-se Figura 1), há ainda estados com grande variação espacial em especial no nordeste brasileiro tais como Pernambuco, Paraíba e Alagoas (veja-se Figura 2). Esses também apresentam grandes estimativas do efeito espaço-temporal, denotado aqui por $e^{\alpha_0+S_0(t)+b_i+h_i}$, fugindo à tendência comum de diminuição do risco de mortalidade infantil de 1991 a 2003 nos demais estados brasileiros (veja-se Figura 3). Além disso, as estimativas das componentes de variância σ_b^2 e σ_h^2 (intervalos de credibilidade a 95%) baseadas no modelo *M4*, respectivamente $0.4414\ (0.2296,0.7851)$ e $0.0094\ (0.0002,0.0594)$, confirmam uma heterogeneidade espacial não observada devido sobretudo a efeitos espacialmente correlacionados. Para uma análise de sensibilidade das suposições *a priori* utilizadas nas distribuições dos hiperparâmetros do modelo, veja-se Silva *et al*. (2008).

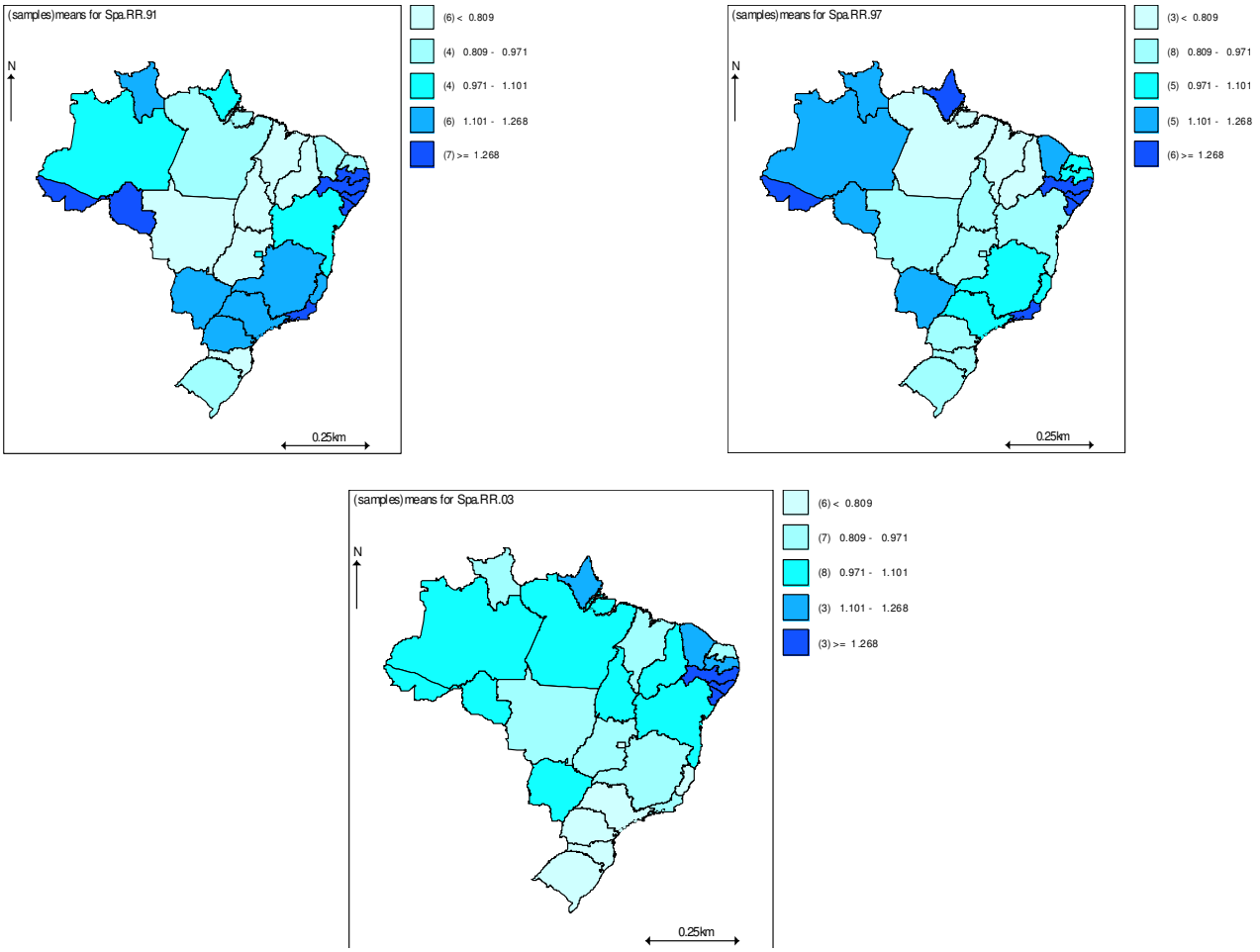


Figura 3: Mapas das estimativas das taxas de mortalidade infantil em 1991 (esquerda), 1997 (direita) e 2001 (centro) no Brasil baseadas no modelo *M4*.

5. Comentário final

Os modelos (1) e (2) proporcionam estimativas suavizadas dos efeitos temporal global e espaço-temporal no mapeamento de taxas e proporções ao longo do tempo e produzem informação interpretativa dos dados. Na análise de mortalidade infantil no Brasil, esses oferecem importantes mecanismos para isolar *e.g*. sub-regiões com maior tendência de risco ("hotspot").

Bibliografia

Amaral Turkman MA, Silva GL. *Modelos Lineares Generalizados - da teoria à prática*. Sociedade Portuguesa de Estatística, Peniche, 2000.

Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, 2004.

Bernardinelli L, Clayton D, Montomoli C. Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine*; **14**, 2411-2431, 1995.

Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1-59, 1991.

Breslow NE, Clayton DG. Approximate inference in generalized linear models. *Journal of the American Statistical Association*, **88**, 9–25, 1993.

Chen M-H, Shao Q-H, Ibrahim, JG. *Monte Carlo Methods in Bayesian Computation*. Sprnger-Verlag, 2000.

Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman and Hall, 1990.

Lawson AB, Browne WJ, Vidal Rodeiro CL. *Disease Mapping with WinBugs and MLWin*. Wiley, 2003.

MacNab YC, Dean CD. Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*; **57**, 949-956, 2001.

Paulino CD, Amaral-Turkman MA, Murteira B. *Estatística Bayesiana*. Fundação Caloute Gulbenkian, Lisboa, 2003.

Silva GL, Dean CB. *Uma Introdução à Análise de Modelos Espaço-temporais para Taxas, Proporções e Processos de Multi-estados*. Associação Brasileira de Estatística, Caxambú, 2006.

Silva GL, Dean CB, Niyonsenga T, Vanasse A. Hierarchical Bayesian spatiotemporal analysis of revascularization odds using smoothing splines. *Statistics in Medicine*, **27**, 2381-2401, 2008.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*; **64**, 583-639, 2002.

Smith BJ. *BOA User Manual* (version 1.1.6). Department of Biostatistics, College of Public Health, University of Iowa, 2007 (http://www.public-health.uiowa.edu).

Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D. *GeoBUGS User Manual* (version 1.2). Department of Epidemiology and Public Health, Imperial College, St Mary's Hospital London, 2004 (http://www.mrc-bsu.cam.ac.uk/bugs/).

Waller LA, Gotway CA. *Applied Spatial Statistics for Public Health Data*. Wiley, 2004.

A Análise de Dados Espaciais em duas áreas das Ciências Biológicas

Teresa Sampaio, *tsampaio@isa.utl.pt*
CEF, Instituto Superior de Agronomia, UTL

Juan Zwolinski, *juan.zwolinski@noaa.gov*
NOAA NMFS Southwest Fisheries Science Center, USA

Manuela Neves, *manela@isa.utl.pt*
CEAUL e Instituto Superior de Agronomia, UTL

O desafio…

Muitos fenómenos que são objecto do nosso estudo ocorrem no espaço e/ou no tempo. Saber "onde" e/ou "quando" ocorrem é muito importante e nalguns casos essencial para a compreensão do problema. A recolha e tratamento de características de interesse de um dado fenómeno, considerando explicitamente a sua localização espacial e incorporando essa informação na análise a realizar, é o cerne da Análise de Dados Espaciais. Esta é uma área da Estatística relativamente recente e por isso com uma intensa actividade de investigação. São diversas as áreas do conhecimento onde a Análise de Dados Espaciais tem dado um contributo considerável para uma melhor compreensão de alguns dos seus fenómenos tais como saúde, ecologia, ambiente, agronomia, geologia, sociologia, entre outras.

Aceitar **o desafio** que o Prof. Fernando Rosado nos lançou para escrever sobre Análise de Dados Espaciais foi um "risco" que decidimos correr, pelo gosto que esta área da Estatística nos desperta. Muitos outros colegas, alguns dos quais citaremos neste texto, estarão muito mais habilitados do que nós para escrever sobre este tema. E, como existe já excelente material didáctico, as lacunas ou falhas que aqui surjam poderão ser colmatadas pela consulta dessas obras. Desde o clássico livro de Cressie (1991), referência obrigatória para quem envereda por estes "caminhos", a um recente e excelente livro preparado pelas nossas colegas Lucília de Carvalho e Isabel Natário (Carvalho & Natário, 2008) para um mini-curso realizado aquando do XVI Congresso da SPE em Vila Real, podemos citar algumas outras obras como Goovaerts (1997), Chilès & Delfiner (1999), Webster & Oliver (2001), Diggle (2003) e Soares (2006), umas com uma vertente mais teórica, outras com predominância nas aplicações. Aceitámos o desafio porque gostaríamos de trazer aqui a nossa experiência de aplicação da Análise de Dados Espaciais no tratamento de situações reais no contexto das Ciências Biológicas. A utilização de metodologias da Estatística Espacial em duas áreas científicas que em Portugal encontram uma grande expressão - a área dos recursos pesqueiros e a área do melhoramento genético do sobreiro - surgiu na sequência do trabalho desenvolvido na realização das teses de mestrado de dois dos autores, Zwolinski (2003) e Sampaio (2007).

Notas Introdutórias

Os dados espaciais são constituídos por observações ou medições de uma ou mais características de um fenómeno em n localizações, $\boldsymbol{s}_1,\boldsymbol{s}_2,\ldots\boldsymbol{s}_n$, de um domínio espacial contínuo limitado ou ainda em regiões específicas. Além das observações daquelas características, os dados espaciais podem ainda referir-se às próprias localizações. As localizações são na maioria das vezes coordenadas no plano $\mathbf{R}^2$ ou no espaço $\mathbf{R}^3$**.** Enquanto os modelos de séries temporais procuram captar as correlações entre respostas obtidas em diferentes pontos no tempo, nos dados espaciais é a estrutura de correlação espacial que necessita de ser incorporada e modelada.

Um exemplo pioneiro, onde intuitivamente a componente espaço foi incorporada nas análises realizadas, surgiu num trabalho de John Snow (1854) a propósito de uma epidemia de cólera trazida da Índia e que alastrou em Londres. O mapa da zona de Londres onde grassava a epidemia, localizando precisamente as residências dos óbitos ocasionados pela doença e as bombas de água que abasteciam a cidade, permitiu identificar o local de uma destas bombas como o epicentro da epidemia. Estudos posteriores confirmaram as hipóteses avançadas por Snow e o seu trabalho é apontado como o primeiro exemplo de um estudo espacial de um problema de saúde pública, ver mais pormenores em Carvalho & Natário (2008). A importância da dependência espacial foi também compreendida no início do séc. XX quando alguns dos métodos de delineamento experimental (repetição, aleatorização, blocos,…) foram introduzidos em estudos agrícolas no sentido de controlarem a correlação espacial, ver Webster & Oliver (2001) para uma boa resenha histórica.

A Análise de Dados Espaciais compreende o estudo de três tipos de dados:

- **Os dados referentes a pontos** ou **dados geoestatísticos –** são observações registadas em localizações fixas de um espaço contínuo. A Geoestatística começou com aplicações no domínio da pesquisa mineira e o prefixo "geo" significa que os assuntos tratatos eram referentes à "Terra". Mas a geoestatística vai agora muito para além da sua origem, tendo aplicações em áreas ambientais como climatologia, monitorização ambiental, geologia, topografia, ecologia, genética agrícola e florestal, etc.
- **Os dados referentes a processos pontuais –** são as localizações aleatórias do acontecimento de interesse num domínio fixo. Uma questão fundamental na análise deste tipo de dados é saber se as observações de interesse ocorrem aleatoriamente ou apresentam algum padrão de regularidade ou de agrupamento no espaço. Como exemplos deste tipo de dados podemos referir as localizações de crimes, ocorrências de doenças, localizações de espécies vegetais, etc.
- **Os dados referentes a áreas –** são dados espaciais indexados a sub-regiões que constituem uma partição de um domínio contínuo. As regiões podem ser regular ou irregularmente espaçadas. Observações obtidas por detecção remota em satélites constituem um exemplo de dados deste tipo. A superfície da terra é dividida em pequenos rectângulos (*píxeis*) e os dados são registados num reticulado regular em $\mathbf{R}^2$.

Neste trabalho faremos uma breve introdução à Geoestatística, pois é esta a área onde foram desenvolvidas as aplicações que serão referidas.

Conceitos básicos em Geoestatística

Um conceito chave na análise de dados espaciais é o conceito de *dependência espacial*. A maior parte das ocorrências, sejam naturais ou sociais, apresentam entre si relações que dependem da distância (e.g., o crescimento de uma árvore num local afecta o crescimento de outras próximas). A dependência espacial é expressa pela *autocorrelação espacial.* Este termo, derivado do termo correlação que avalia a relação entre duas variáveis, avalia a correlação da mesma variável, mas em locais distintos do espaço.

Se $\boldsymbol{D}$ é a região de interesse (uma floresta, um campo, uma área mineira, etc), cada ponto $\boldsymbol{s}$ em $\boldsymbol{D}$ pode ser descrito pelas coordenadas x e y no plano, $\boldsymbol{s}=(x,y)$. Nas aplicações pretende-se registar os valores de uma variável Z em localizações dentro da região $\boldsymbol{D}$. $z(\boldsymbol{s})$ será então o valor de uma variável observado na localização $\boldsymbol{s}$. Na prática as medições são obtidas num número finito n de pontos, sendo um conjunto de dados geoestatísticos apresentado como $z(\boldsymbol{s}_1), z(\boldsymbol{s}_2),\ldots, z(\boldsymbol{s}_n)$.
Um modelo muito simples para dados geoestatísticos é $Z(\boldsymbol{s})=\mu+\varepsilon(\boldsymbol{s})$, no qual se admite que $E[Z(\boldsymbol{s})]=\mu$, $Var[Z(\boldsymbol{s})]=\sigma^2$ e $Cov[Z(\boldsymbol{s}_1),Z(\boldsymbol{s}_2)]=C(\boldsymbol{s}_1-\boldsymbol{s}_2)$, onde C é designada por *função de covariância* ou *covariograma*. Quer dizer, a covariância da resposta em duas localizações distintas depende apenas do vector das distâncias entre as localizações. Um processo que verifique as condições indicadas acima diz-se ser um processo *estacionário de segunda ordem*. Quando a função covariância depende apenas da distância, diz-se que o processo é *isotrópico*. A chave do sucesso na modelação da variação a pequena escala de dados espacialmente correlacionados reside na adequada identificação da relação

entre a distância entre pontos (dados) e a correlação entre eles. Em vez de usar a função de covariância, é costume usar-se uma função relacionada, o *variograma.* O variograma considera o quadrado das diferenças entre pontos, i.e., $[Z(\boldsymbol{s_1})-Z(\boldsymbol{s_2})]^2$. Se a variância das diferenças depender apenas da diferença $\boldsymbol{s_1}-\boldsymbol{s_2}$, chama-se *variograma* a $\mathrm{Var}[Z(\boldsymbol{s_1})-Z(\boldsymbol{s_2})]$ e costuma representar-se por $2\gamma(\boldsymbol{s_1}-\boldsymbol{s_2})$. À função γ chama-se *semivariograma*. Veremos a utilização e interpretação destes conceitos no que se segue.

Vários modelos teóricos de semivariogramas foram desenvolvidos e a estimação adequada dos parâmetros desses modelos é uma peça chave para a fase da inferência. A metodologia base para a predição espacial é o *kriging*, baseado na minimização do erro quadrático médio. O termo *kriging* deve-se a Matheron (1963) que o criou em honra ao Engenheiro de Minas sul africano D.G. Krige, cujo trabalho em 1951 apresenta as equações básicas para a interpolação linear óptima num domínio espacialmente correlacionado. Diferentes procedimentos de *kriging* foram sendo desenvolvidos, sendo o seu objectivo prever o valor da resposta Z em localizações não observadas no domínio em estudo. Os processos de estimação e predição necessitam de pressupostos, tendo os *processos gaussianos* um papel central na modelação espacial.

Aplicação da análise de dados espaciais no estudo de um ensaio genético de sobreiro

O sobreiro (Quercus suber L.) é uma espécie florestal de interesse nacional, não só pelo seu contributo na economia mas também pelo seu papel ecológico e social. Portugal é líder mundial na produção e exportação de cortiça. Produzimos cerca de 53% do total de cortiça e de acordo com dados de 2005 do *United Nations Statistics* ocupamos o primeiro lugar das exportações com cerca de 60%. Aproximadamente 90% da cortiça transformada no nosso País tem como destino o mercado internacional. O valor gerado por estas transacções representa aproximadamente 0,7% do PIB (preços de mercado), 2,3% do valor das exportações totais nacionais e cerca de 30% do total das exportações portuguesas de produtos florestais (APCOR, 2009).
Portugal faz parte de uma rede internacional de ensaios genéticos de sobreiro (ensaios de proveniências e descendências[1]), nos quais estão representadas populações provenientes de toda a área de distribuição natural da espécie - Portugal, Espanha, França, Itália, Marrocos, Argélia e Tunísia (Varela, 2003). O objectivo destes campos experimentais (testes genéticos) é permitir aos melhoradores florestais por um lado quantificar a variabilidade genética existente e por outro identificar, seleccionar e multiplicar os melhores genótipos de forma a garantir um aumento na produção de cortiça em quantidade e qualidade (Almeida, 2005). A identificação dos melhores genótipos é feita com base na avaliação do fenótipo, isto é, com base no valor observável da característica em estudo e que resulta de duas componentes: a componente genética que é transmitida à descendência e a componente ambiental. Para seleccionar os melhores genótipos é necessário ordená-los, tanto quanto possível, pelo seu mérito genético controlando a influência ambiental. Mas esta tarefa nem sempre é fácil, pois os factores que caracterizam a maioria dos ensaios genéticos – grande número de genótipos (tratamentos), grandes áreas experimentais (em geral, superiores a 1ha), topografia irregular, heterogeneidade espacial da fertilidade e estrutura do solo, que pode reflectir-se em gradientes ou manchas, afectando os tratamentos de forma sistemática – conduzem a uma "camuflagem" dos factores genéticos.

Uma forma de contornar o problema da dependência espacial das avaliações efectuadas em ensaios de campo florestais passa por utilizar um delineamento experimental adequado. O clássico delineamento em blocos casualizados completos, frequentemente usado nos testes genéticos florestais (Wrigth, 1976;

[1] Os ensaios de proveniências são ensaios de campo onde árvores da mesma espécie mas de diferentes origens geográficas/áreas de distribuição (proveniências) são plantadas e o seu comportamento, em termos adaptativos e de crescimento, é avaliado. Estes ensaios permitem avaliar a variabilidade genética disponível para o melhoramento, maximizar a adaptabilidade das proveniências e avaliar a existência de interacção genótipo-ambiente (Wright, 1976; Zobel & Talbert, 1984). Nos ensaios de descendências as famílias são as entidades genéticas avaliadas. Estes ensaios são estruturas experimentais que permitem avaliar a superioridade genética de um indivíduo como progenitor, isto é, comparar o valor da sua descendência com o valor das descendências de outros progenitores. Assim, os principais objectivos destes ensaios, no caso de famílias de meios-irmãos, são determinar o valor reprodutivo dos progenitores por intermédio do valor médio das respectivas descendências, avaliar o ganho genético e seleccionar novos indivíduos para a geração seguinte.

Zobel & Talbert, 1984), embora apresente várias vantagens, dificilmente consegue remover a correlação espacial existente entre árvores vizinhas (*autocorrelação espacial*) pelo que, diminui a precisão das estimativas das componentes da variância genotípica e ambiental. Tal situação compromete o rigor e interpretação das estimativas dos parâmetros genéticos e as consequentes decisões do melhorador. Uma vez que a variância genotípica traduz a variabilidade genética da população em estudo e influencia os resultados da selecção genética, os melhores preditores lineares não enviesados dos efeitos genotípicos dependem directamente das componentes de variância envolvidas no modelo de análise (Gonçalves, 2008). Fica evidente que a avaliação destes ensaios será tanto mais rigorosa quanto mais adequado for o delineamento experimental e a capacidade das técnicas estatísticas controlarem a variabilidade espacial. Vários estudos de ensaios genéticos florestais (Magnussen, 1990; Anekonda & Libby, 1996; Fu *et al.,* 1999; Kusnadar & Galwey, 2000; Costa e Silva *et al.,* 2001; Saenz-Romero *et al*., 2001; Joyce *et al*., 2002; Hamann *et al.,* 2002; Dutkowski *et al*., 2002; Dutkowski *et al.,* 2006; Williams, 2006 e Zamudio *et al.,* 2007) sugerem que a precisão das estimativas das componentes da variância dos modelos em análise pode aumentar, quando aplicada uma análise estatística espacial que considere a variabilidade espacial do ensaio e a modele utilizando princípios da geoestatística.

A aplicação do modelo linear misto é frequente no contexto dos ensaios genéticos, pois permite incorporar as diversas causas da variabilidade – genética e ambiental. Por outro lado, a possibilidade de considerar as entidades genéticas como factores de efeitos aleatórios, permite ao melhorador prever, com base nos melhores preditores lineares não enviesados, o ganho genético de selecção. A incorporação da localização espacial das árvores (coordenadas espaciais) tem permitido ajustar modelos mistos espaciais aos dados provenientes destes ensaios. Estes, tal como o modelo misto clássico, permitem modelar a tendência global (processo não estacionário) desde que sejam considerados no modelo termos apropriados tais como os factores do delineamento experimental. A grande vantagem dos modelos espaciais resulta da possibilidade de modelar a tendência local que se admite como um processo estacionário de segunda ordem. Tal modelação só é possível graças à decomposição do vector de erros aleatórios na soma de um vector de erros aleatórios espacialmente correlacionados, $\boldsymbol{\xi}$, e que portanto reflecte a variação espacial em pequena escala, e um vector de erros aleatórios sem correlação espacial, $\boldsymbol{\eta}$. Mais ainda, o facto de se considerar $\boldsymbol{\xi}$ como um processo estacionário de segunda ordem permite modelar a matriz de variância-covariância do erro de diferentes modos dado que a correlação entre unidades experimentais apenas depende da distância que as separa. É neste contexto que surge a geoestatística como ferramenta de apoio à análise de grandes ensaios experimentais. O estudo do processo espacial, em particular da existência de correlação entre observações sucessivas, pode ser feito recorrendo a instrumentos geoestatísticos como por exemplo o semivariograma empírico que permite modelar a variabilidade espacial do campo experimental. O semivariograma experimental dos resíduos de um modelo que considere os erros independentes (modelo clássico) é a ferramenta de diagnóstico mais usada dado que a sua representação gráfica permite identificar a presença, magnitude e padrão da correlação espacial bem como averiguar se tal correlação depende da direcção considerada (anisotropia). Este, além de ser um instrumento de avaliação da existência de correlação espacial, permite ainda obter valores iniciais para as estimativas dos parâmetros - *range, sill* e *nugget*. A partir destas estimativas é possível, de modo iterativo, ajustar um modelo teórico de semivariograma aos dados em estudo, e portanto identificar a estrutura da matriz de correlação espacial do erro (isotrópica ou anisotrópica com correlação esférica, gaussiana, exponencial, potência, etc.) que melhor caracterize o processo e proporcione estimativas mais precisas dos efeitos em estudo.

O trabalho desenvolvido teve como objectivo, avaliar por um lado se, à semelhança de outros estudos, existe correlação espacial entre a altura das árvores do ensaio de proveniências de sobreiro da Herdade do Monte Fava e por outro, averiguar se uma análise espacial é capaz de remover o efeito de tal correlação e produzir estimativas mais precisas do efeito genotípico das proveniências, relativamente ao modelo clássico. Para tal, iniciou-se o estudo com uma avaliação da dependência espacial da altura das árvores, seguiu-se a modelação espacial da estrutura de variância-covariância do erro e por fim comparou-se a eficiência do modelo linear misto com a eficiência de vários modelos mistos espaciais para este ensaio genético delineado classicamente.

O ensaio de proveniências da Herdade do Monte Fava, foi instalado em Portugal em 1998 e localiza-se em Ermidas do Sado, concelho de Santiago do Cacém (80º7''W, 38º00'N). Nele, estão representadas 35 proveniências de sobreiro: nove portuguesas, sete espanholas, uma luso-espanhola, cinco italianas, quatro francesas, duas tunisinas, seis marroquinas e uma argelina. O delineamento experimental em blocos casualizados completos (30 blocos), com compasso de plantação de 6mx6m, permitiu fazer o estudo da variabilidade espacial considerando uma grelha regular de 98 colunas (*Ci* com *i = 1,...98*) e 41 linhas (*Lj* com *j = 1, ..., 41*). Cada árvore foi identificada pela coordenada espacial (*Ci, Lj*) que corresponde a uma distância real no campo experimental.

Do ajustamento do modelo linear misto clássico verificou-se que existe variabilidade significativa associada aos factores em estudo - bloco e proveniência. A presença de variabilidade espacial foi claramente identificada pela representação gráfica do semivariograma dos resíduos do modelo clássico. O crescimento da função semivariograma com a distância, traduziu a existência de correlação espacial entre árvores vizinhas. É de salientar que quando considerado um modelo clássico sem inclusão do efeito do bloco, a presença de tendência global foi ainda mais notória. Não se registou uma distância a partir da qual as observações deixavam de estar correlacionadas. Daqui se conclui que a inclusão de termos associados ao delineamento experimental consegue controlar a variabilidade espacial a larga escala. A modelação espacial da estrutura de variância-covariância do erro baseada no ajustamento de modelos mistos com estruturas de correlação esférica, gaussiana, potência e potência anisótropica permitiu concluir que:

- existe variabilidade espacial significativa associada aos dados da altura;
- com os modelos mistos espaciais o ajustamento dos dados é significativamente melhor;
- o efeito do bloco deixou de ser significativo na maioria dos modelos espaciais ajustados, o que sugere que o controlo da variabilidade espacial foi assegurado pela modelação espacial;
- existe variância genotípica significativa;
- as previsões dos efeitos genotípicos das proveniências são mais precisas com o ajustamento dos modelos espaciais, logo a selecção das proveniências melhor adaptadas é mais eficiente.

Se é certo que as metodologias espaciais ainda não têm a aplicação desejada no contexto biológico apresentado, com os resultados obtidos neste e noutros estudos de genética espera-se que a estatística espacial seja considerada uma alternativa às clássicas análises de dados, por forma a aumentar a eficiência dos programas de melhoramento genético de espécies florestais.

A Geostatística nas ciências pesqueiras

A distribuição dos recursos vivos marinhos, e em particular dos recursos pelágicos (aqueles que habitam a coluna de água sem dependência directa do substrato), está organizada espacialmente em vários níveis intrincados (Mackinson *et al.,* 1999). Em escalas mais pequenas (distâncias na ordem dos cm aos m) imperam as relações ao nível dos indivíduos, enquanto que em maiores escalas (distâncias medidas das dezenas às centenas de km) os padrões de distribuição são progressivamente mais influenciados pelo ambiente, através da disponibilidade do habitat (Zwolinski *et al.,* 2010). Como os diversos efeitos agregativos se confundem, o grande desafio da ecologia pesqueira marinha reside na adequada obtenção de valores isolados para poder avaliar correctamente a abundância dos recursos e obter relações não enviesadas sobre a sua distribuição e as relações com o ambiente. Neste domínio, a geoestatística tem tido um papel fundamental (Rivoirard *et al.,* 2000; Legendre, 1993). Apesar da utilização dos procedimentos de análise de dados espaciais nas ciências pesqueiras ser relativamente recente, é uma área bastante profícua que hoje em dia se alia a metodologias clássicas como a regressão ou estatística multivariada para distinguir os factores internos (e.g., reprodução, comportamento agregativo) dos factores externos (e.g., ambiente, presença de predadores, disponibilidade alimentar) em estudos de ecologia, distribuição e abundância dos recursos pesqueiros (Petitgas, 1998; Zwolinski *et al.,* 2006; Zwolinski *et al.,* 2009).

Um dos pilares das ciências pesqueiras é o cálculo regular da abundância das populações (Hilborn & Walters, 1992). A abundância é em regra obtida com recurso a campanhas científicas que determinam a densidade das populações na sua região de distribuição, utilizando artes de pesca *standard*, métodos acústicos ou colectores de plâncton para a estimação indirecta da abundância dos adultos (Gunderson, 1993). As estimativas de abundância são posteriormente utilizadas em modelos populacionais que também incorporam séries temporais das capturas comerciais para o acompanhamento da redução dos efectivos nas várias coortes anuais. Os modelos de avaliação fornecem então estimativas de abundância consideradas absolutas a partir das quais se tomam decisões sobre a fracção do recurso a ser explorada no futuro próximo.

No caso de se proceder a uma amostragem aleatória da população, o cálculo da densidade média numa dada região e respectiva variância consideram a hipótese de independência entre as observações (Cochran, 1977). No entanto, por motivos logísticos, a amostragem de grandes populações no mar é feita de forma sistemática ou regular (Simmonds & McLennan, 2005). Nestes casos, se a população amostrada for espacialmente estruturada, o cálculo da variância requer que a dependência espacial seja tomada em conta. Assim, dentro da teoria intrínseca da geoestatística, se for possível admitir a estacionaridade de segunda ordem, a estimativa da variância da média numa região V já não é dada por s^2/n, onde s^2 representa a variância amostral mas por $s_E^2 = 2\bar{\gamma}_{\alpha V} - \bar{\gamma}_{VV} - \bar{\gamma}_{\alpha\beta}$, onde $\bar{\gamma}_{VV}$ é o semivariograma médio para todas as distâncias na região V, $\bar{\gamma}_{\alpha\beta}$ é a semivariância entre todas as amostras e $\bar{\gamma}_{\alpha V}$ é a semivariância entre todas as amostras e a região V (Rivoirard *et al.,* 2000). Na maioria das vezes, a amplitude dos valores para determinação da autocorrelação é bastante inferior às distâncias possíveis na área de amostragem do que resulta $s^2/n < s_E^2$. Há casos porém em que a estacionaridade na média e na variância não são pressupostos realistas e o cálculo do semivariograma pode ficar comprometido (Chilès & Delfiner, 1999). A não estacionaridade na média obriga a que se efectue a decomposição do sinal em tendência (processo determinístico) e correlação espacial (processo estocástico). No entanto esta decomposição não é única e depende muitas vezes do propósito do utilizador ou da existência de variáveis explicativas (Cressie, 1991). Num cenário puramente preditivo, o cálculo da média e da sua variância numa região em que o pressuposto da estacionaridade é posto em causa acarreta uma perda de informação e não gera estimativas de variância mínima. A solução mais corrente passa pela criação de estratos considerados independentes e com estacionaridade interna para os quais se calculam as respectivas médias e variâncias. Em situações em que os limites são difusos e a estacionaridade pode ser posta em causa, Matheron (1971) desenvolveu a abordagem transitiva, que assenta em pressupostos menos rígidos do que a abordagem intrínseca, necessitando apenas que a amostragem seja feita por transectos paralelos e equidistantes. As abundâncias obtidas com base numa amostragem por transectos são utilizadas para calcular a função de covariância transitiva, a partir da qual se pode obter uma estimativa da variância da abundância global (Bez, 2002).

Há situações em que o mapeamento da distribuição espacial é tão importante como o cálculo da abundância. Dentro do quadro da geoestatística intrínseca, mapas de densidade podem ser criados por *krigagem* (Cressie, 1991). A integração numérica das superfícies de densidade fornece estimativas de abundância globais, podendo as variâncias ser estimadas com o recurso a simulações que usam modelos gaussianos (Gimona & Fernandes, 2003). Mas neste caso é necessário que a distribuição dos dados satisfaça certos critérios como a gaussianidade e a estacionaridade intrínseca (implicando a ausência de tendência), que na prática dificilmente se verificam. Para interpolações de dados com tendência, a utilização de *splines thin-plate* (Wahba & Wendelberger, 1980) dentro de um quadro de modelos aditivos generalizados (GAM) surge como a melhor alternativa. As *splines thin-plate* estão formalmente relacionadas com a *krigagem* universal (Cressie, 1991 e Watson, 1984), mas quando usadas nos GAM facilitam a modelação de casos em que a distribuição espacial e a presença de variáveis explicativas coexistem de uma maneira complexa (Wood & Agustin, 2002 e Wood, 2005). A principal vantagem da modelação espacial utilizando *splines* é o facto de a determinação da estrutura espacial (quer da tendência, quer da correlação espacial) não ser feita explicitamente, mas resultar de um processo de optimização que permite uma situação de compromisso entre a qualidade do ajustamento e a "suavidade" da função (Wahba & Wendelberger, 1980). Ainda, num GAM, a variável

resposta pode tomar qualquer distribuição disponível num modelo linear generalizado, pelo que é possível a modelação em duas fases de observações em que os processos que regulam a densidade e a presença podem ser considerados independentes (Zwolinski *et al*., 2009). Os GAM permitem portanto a modelação de estruturas não estacionárias (quer na média, quer na variância), eliminando o processo de pós-estratificação da área de estudo, não necessitando da utilização de transformações estabilizadoras da variância nem o cálculo explícito da correlação espacial.

A utilização da análise de dados espaciais na biologia pesqueira não tem tido apenas relevância na estimação da abundância dos recursos. Hoje em dia, o estudo da autocorrelação espacial é necessário para a compreensão da ecologia e distribuição dos organismos (Diniz *et al.,* 2003). Entre outros propósitos, a modelação da distribuição das espécies em função do ambiente permite quantificar o seu nicho ecológico de forma a fazer previsões sobre deslocações, invasões ou proliferações em função de alterações do ambiente (Guisan & Thuiller, 2005). Mais ainda, quando a distribuição espacial de um recurso pode ser inferida, *a priori,* a partir de modelos de distribuição e informação remota, a amostragem do recurso para efeitos de avaliação pode ser feita de forma mais eficiente. A utilização conjunta de modelos de regressão e análise geoestatística permite procurar as escalas em que os efeitos ambientais e a correlação espacial fundamental (aquela que está exclusivamente relacionada com a atracção de organismos conspecíficos) actuam (Hawkins *et al.,* 2007, Zwolinski *et al.,* 2010). Para compreender esta última, são imperiosos estudos de pequena escala quer espacial quer temporal, onde o efeito das variáveis ambientais possa ser considerado negligenciável (Zwolinski *et al.,* 2006). No outro extremo, a análise da distribuição das espécies em função do ambiente tem necessariamente de integrar grandes regiões, de preferência bem maiores do que o maior alcance da autocorrelação fundamental e durante um intervalo de tempo considerável para que os efeitos da distribuição espacial em pequena escala possam ser amenizados (Guisan *et al.,* 2007). Uma vez que os peixes que formam cardumes e os próprios cardumes apresentam uma densidade no espaço que terá intrinsecamente autocorrelação positiva, (Zwolinski *et al.,* 2009), os valores dos erros padrão são relativamente baixos inflacionando a percepção da significância dos modelos de previsão de distribuição e os erros tipo I (Hawkins *et al*. 2007). Um dos cuidados a ter na modelação de relações ambiente-organismos reside portanto no conhecimento prévio das escalas de correlação espacial fundamentais de forma a criar planos de amostragem adequados ou então utilizando modelos que contemplem explicitamente a correlação espacial (Diniz *et al*., 2003, Pinheiro & Bates, 2000)

Em conclusão, a estatística espacial tem tido no passado recente um papel preponderante na compreensão da distribuição dos organismos marinhos, no seu mapeamento e no cálculo das suas abundâncias. A quantificação dos *habitat* das espécies irá ultimamente reflectir-se na adequação das campanhas de monitorização, permitindo definir melhor os limites de amostragem e levando consequentemente a uma racionalização do esforço amostral. O conhecimento dos vários níveis de agregação dos organismos marinhos permitirá adoptar padrões finos de amostragem adequados a cada espécie e cenário de abundância, ditando ainda o melhor estimador a utilizar (e.g, Barange & Hampton, 1997; Simmonds *et al,* 2009).

Outros desafios se seguem …

A estimação e previsão em Geoestatística, com aplicações de grande interesse nas áreas mais variadas, são na maioria dos casos baseadas na hipótese da gaussianidade. Tal pressuposto é inadequado quando se pretende modelar o comportamento da cauda de uma distribuição. As abordagens baseadas no variograma podem também não ser possíveis porque pode ter-se $E[Z(\boldsymbol{s})]=+\infty$ e $Var[Z(\boldsymbol{s})]=+\infty$.

Outros modelos e metodologias necessitam por isso de ser desenvolvidos e constituem uma área recente de investigação – a Estatística de Extremos Espaciais. Muitas contribuições têm estado a surgir, das quais queremos mencionar o recente *package* SpatialExtremes (Ribatet, 2009), integrado no ambiente **R.**

Referências Bibliográficas

Almeida, M.H. (2005). Importância do controlo genético e da manipulação dos materiais florestais de reprodução na sustentabilidade dos montados de sobro em e.Ciência, *Revista da Ciência, Tecnologia e Inovação em Portugal. Produção em Portugal: Cortiça e Azeite - Inovação e Crescimento*: 24 – 27.

Anekonda, T.S. & Libby, W.J. (1996). Efectiveness of nearest-neighbor data adjustment in a clonal test of Redwood. *Silvae Genetica*, 45(1): 46 – 51.

APCOR, 2009. *Anuário 2009*. Associação Portuguesa de Cortiça. Santa Maria de Lamas.

Barange, M., & Hampton, I. (1997). Spatial structure of co-occurring anchovy and sardine populations from acoustic data: implications for survey design. *Fisheries Oceanography*,6: 94-108.

Bez, N. (2002). Global fish abundance estimation from regular sampling: the geostatistical transitive approach. *Canadian Journal of Fisheries and Aquatic Science,* 59: 1921-1931.

Carvalho, M.L. & Natário, I.C. (2008). *Análise de Dados Espaciais*. Sociedade Portuguesa de Estatística.

Chilès, J-P., & Delfiner, P. (1999). *Geostatistics-modelling spatial uncertainty*. 1st ed. Wiley Series in Probability and Statistics, John Wiley & Sons Inc., New York.

Costa e Silva, J., Dutkowski, G.W. & Gilmour, A.R. (2001). Analysis of early tree height in forest genetic trials is enhanced by including a spatially correlated residual. *Canadian Journal of Forest Research,* 31: 1887 – 1893.

Cochran, W.G. (1977). *Sampling techniques*. 3rd ed. John Wiley & Sons, New York.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley & Sons, New York.

Diniz, J.A.F, Bini, L.M., & Hawkins, B.A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, 12: 53 – 64.

Diggle, P.(2003). *Spatial Point Patterns* -2nd ed. Arnold London.

Dutkowski, G.W., Costa e Silva, J, Gilmour, A.R., Wellendorf, H. & Aguiar, A. (2006). Spatial analysis enhances modeling of a wide variety of traits in forest genetic trials. *Canadian Journal of Forest Research,* 36: 1851 – 1870.

Dutkowski, G.W., Costa e Silva, J., Gilmour, A.R. & Lopez, G.A. (2002). Spatial analysis methods for forest genetic trials. *Canadian Journal of Forest Research,* 32: 2201–2214.

Fu, Y.B., Namkoong, G. & Yanchuk, A.D. (1999). Spatial patterns of tree height variations in a series of Douglas-fir progeny trials: implications for genetic testing. *Canadian Journal of Forest Research,* 29: 714 – 723.

Gimona A. & Fernandes P.G., (2003). A conditional simulation of acoustic survey data: advantages and potential pitfalls. *Aquatic Living Resources,* 16:123–129.

Gonçalves, E.M.F. (2008). *Modelos estatísticos espaciais para ensaios de populações vegetais.* Tese de doutoramento em Matemática e Estatística. Universidade Técnica de Lisboa, Instituto Superior de Agronomia.

Goovaertz, P (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.

Guisan, A., & Thuiller, W. 2005. Predicting species distributions: offering more than simple habitat models. *Ecology Letters*, 8, 993–1003.

Guisan, A., Graham, C.H., Elith, J., Huettmann, F., & NCEAS Species Distribution Modelling Group. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions,* 13: 332-340.

Gunderson, D.R. (1993). *Surveys of fisheries resources*. John Wiley & Sons, Inc., New York.

Hamann, A., Kosky, M.P. & Namkoong, G. (2002). Improving precision of breeding values by removing spatially autocorrelated variation in forestry field experimernts. *Silvae Genetica*, 51(5/6): 210 – 215.

Hawkins, B.A., Diniz-Filho, J.A., Bini, L.M., De Marco, P., & Blackburn, T.M. (2007). Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. *Ecography*, 30: 375-384.

Hilborn, R. & Walters, C.J. (1992). *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty*. Chapman & Hall, New York.

Joyce, D., Ford, R. & Fu, Y.B. (2002). Spatial patterns of tree height variations in a black spruce farm-field progeny test and neighbors-ajusted estimations of genetic parameters. *Silvae Genetica*, 51(1): 13 – 18.

Kusnadar, D. & Galwey, N. (2000). A proposed method for estimation of genetic parameters on forest trees without raising progeny: critical evaluation and refinement. *Silvae Genetica*, 49(1): 15 – 21.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? Ecology 74: 1659-1673

Magnussen, S. (1990) - Application and comparison of spatial models in analyzing treegenetics field trials. *Canadian Journal of Forest Research*, 20(5): 536 – 546.

Matheron, G. (1971). The theory of regionalized variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique* no 5.

Mackinson, S., Nøttestad, L., Guenette, S., Pitcher, T., Misund, O. A., & Ferno, A. (1999). Cross-scale observations on distribution and behavioural dynamics of ocean feeding Norwegian spring-spawning herring (Clupea harengus L.). *ICES Journal of Marine Science*, 56: 613–626.

Petitgas, P. (1998). Biomass-dependent dynamics of fish spatial distributions characterized by geostatistical aggregation curves. *ICES Journal of Marine Science*, 55: 443–453.

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-Effects Models in S and S-Plus*. Statistics and Computing Series. Springer. Verlag, New York.

Ribatet, M. (2009). *A User's Guide to the SpatialExtremes Package*. École Polytechnique Fédérale de Lausanne, Switzerland.

Rivoirard, J., Simmonds J., Foote K. G., Fernandes P., & Bez N. (2000). *Geostatistics for Estimating Fish Abundance*. Oxford: Blackwell Science.

Saenz-Romero, C., Nordheim, E.V., Guries, R.P. & Crump, P.M. (2001). A case study of provenance/progeny testing using trend analysis with correlated errors and SAS PROCMIXED. *Silvae Genetica*, 50(3-4): 127 – 135.

Sampaio, T. (2008). *Aplicação de Modelos Mistos em Ensaios Genéticos Florestais*. Tese de Mestrado em Matemática Aplicada às Ciências Biológicas. Universidade Técnica de Lisboa, Instituto Superior de Agronomia.

Simmonds, J., & MacLennan, D. (2005*). Fisheries Acoustics: Theory and Practice*. Blackwell Science Ltd., Oxford .

Simmonds, J., Gutiérrez, M., Chipollini, A., Gerlotto, F., Woillez, M. & Bertrand, A. (2009). Optimizing the design of acoustic surveys of Peruvian anchoveta. *ICES Journal of Marine Science*, 66: 1341-1348.

Soares, A. (2006). *Geoestatística para as Ciências da Terra e do Ambiente*. Colecção ensino da Ciência e da Tecnologia. IST Press.

Varela, M.C. (2003). *Handbook concerted action Fair 1 CT 95-0202 Evaluation of genetic resources of cork oak for appropriaté use in breeding and gene conservation strategies*. Estação Florestal Nacional.

Watson, G.S. (1984). Smoothing and interpolation by kriging and with splines. *Mathematical Geology* 16: 601-615.

Wahba G., & Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly Weather Review,* 108: 1122–1145

Webster, R. & Oliver, M.A. (2001). *Geostatistics for Environmental Scientists*. Statistics in Practice. John Wiley & Sons.

Williams, E.R., John, J.A. & Whitaker, D. (2006). Construction of resolvable spatial row-column designs. *Biometrics*, 62(1): 103 – 108.

Wood, S.N. (2005*). Generalized additive models: and introduction with R.* 1st ed. Texts in Statistical Sciences, Chapman & Hall/CRC, London, UK.

Wood, S.N., & Augustin, N.H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157: 157–177.

Wright, J.W. (1976). Introduction to Forest Genetics. Academic Press, Inc., New York.

Zamudio, F., Wolfinger, R., Stanton, B. & Guerra, F. (2007). The use of linear mixed model theory for the genetic analysis of repeated measures from clonal tests of forest trees. I. A focus on spatially repeated data. Tree Genetics & Genomes.

Zobel, B. & Talbert, J. (1984). *Applied Forest Tree Improvement*. John Wiley & Sons. Inc. New York.

Zwolinski, J., (2007). *Modelação da distribuição espacial e abundância em pequena escala dos ovos de sardinha* (*Sardina pilchardus*). Tese de Mestrado em Matemática Aplicada às Ciências Biológicas. Universidade Técnica de Lisboa, Instituto Superior de Agronomia.

Zwolinski, J., Mason, E., Oliveira, P.B. & Stratoudakis Y. (2006). Fine-scale distribution of sardine (Sardina pilchardus) eggs and adults during a spawning event. *Journal of Sea Research*, 56: 294-304.

Zwolinski, J., Fernandes, P.G., Marques, V. & Stratoudakis, Y. (2009). Estimating fish abundance from acoustic surveys: calculating variance due to acoustic backscatter and length distribution error. *Canadian Journal of Fisheries and Aquatic Sciences,* 66: 2081-2095.

Zwolinski, J.P., Oliveira, P.B., Quintino, V. & Stratoudakis, Y. (2010). Sardine potential habitat and environmental forcing off western Portugal. *ICES Journal of Marine Science,* 67, *versão on-line.*

Uma Perspectiva Histórica da Estatística Espacial

M. L. Carvalho*, *mlcarvalho@fc.ul.pt*

I. Natário**, *icn@fct.unl.pt*

** Faculdade de Ciências da UL; CEAUL*

*** Faculdade de Ciências e Tecnologia da UNL; CEAUL*

1. Introdução

Neste trabalho pretende-se dar uma breve resenha histórica do surgimento e desenvolvimento de uma área da estatística relativamente recente, actualmente conhecida por estatística espacial, que engloba o estudo dos fenómenos em que a localização no espaço contribui de forma decisiva para a sua variabilidade ou em que a questão de interesse é a própria localização do fenómeno. Começamos por esta última situação em que os dados são modelados através de processos pontuais e, seguidamente, apresentamos separadamente para a primeira situação, o caso em que o efeito espacial tem variação contínua - dados referentes a pontos - e o caso em que a essa variação é discreta - dados referentes a áreas.

Note-se que os modelos espaciais vêm responder à necessidade de alargar o conjunto dos métodos clássicos de análise de dados, sujeitos à hipótese de independência das observações, ao caso de dados espaciais onde se torna evidente, tal como acontece com os fenómenos variando no tempo, que dados que estão mais perto no espaço têm tendência a ser mais parecidos dos que os que estão mais afastados. É na capacidade de incorporar, por diversos métodos, esta variação espacial de interdependência de observações que torna estes modelos especiais e capazes de serem aplicados numa grande diversidade de situações reais.

2. Dados referentes a processos pontuais

Os processos pontuais espaciais são modelos matemáticos que descrevem a disposição de objectos que se encontram aleatoriamente ou irregularmente distribuídos no espaço. A um nível básico os dados consistem simplesmente em coordenadas aleatórias de pontos que descrevem as localizações dos objectos, ainda que características adicionais possam ter interesse, sendo registadas (marcas).

Os primeiros exemplos referentes a este tipo de dados surgem possivelmente da necessidade da resolução de questões em probabilidades geométricas, desde o problema da agulha de Buffon (1707-1788), ao cálculo de probabilidades nas áreas da Física e da Astronomia como, por exemplo, o cálculo

da probabilidade de que um certo número de estrelas se encontre num dado rectângulo, no pressuposto que as estrelas se distribuem de forma aleatória no céu (Newcomb, 1860).

O conhecido problema da agulha de Buffon, em que se lança ao acaso uma agulha sobre uma mesa marcada por linhas paralelas (distando mais do que o comprimento da agulha) e se pretende calcular a probabilidade de a agulha cruzar uma das linhas, levanta questões como o que deve acontecer junto das bordas da mesa ou, se a mesa for tomada como sendo de dimensão infinita para obviar a questão, que significado atribuir ao atirar a agulha aleatoriamente para cima da mesa? A resolução matemática desta dificuldade reside na teoria dos processos pontuais.

Os processos pontuais espaciais, sob a forma de processos de Poisson não assim explicitamente identificados, surgiram inicialmente para responder a questões como a acima apresentada na Física e na Astronomia. As propriedades dos processos de Poisson foram derivadas e usadas como marcos relativamente aos quais os dados astronómicos eram julgados. O processo de Poisson homogéneo é utilizado como modelo para padrões de pontos ocorrendo naturalmente no espaço, servindo de fronteira entre processos espaciais cujas realizações podem ser descritas como sendo mais ou menos regulares que as do processo de Poisson.

Outro exemplo seminal dos processos pontuais espaciais, que mostra que um método adequado de representar dados, eminentemente espacial, pode indicar a solução de um problema complexo, mesmo sem se conhecer exactamente as causas do fenómeno em análise, surge no trabalho de John Snow (1855). Em 1854 surgiu em Londres uma epidemia de cólera que matou centenas de pessoas muito rapidamente. Snow suspeitava que a doença era transmitida através de água contaminada e, para tentar provar essa teoria, registou no mapa da parte de Londres onde a epidemia alastrava, a localização precisa da morada de cada um dos mortos por cólera. O mapa que desenhou mostra claramente uma aglomeração do número de mortes nas imediações de determinado poço. Na sequência do seu trabalho, a manivela da bomba do referido poço foi removida. Em 1886 a descoberta do vibrião da cólera veio confirmar a hipótese avançada por Snow, cujo trabalho é com frequência apontado como o primeiro exemplo de aplicação da representação espacial de dados na descoberta de solução para um problema de saúde pública.

A análise de dados referentes a processos pontuais tem frequentemente como objectivo o estudo do agrupamento das ocorrências no espaço, como as registadas no problema da cólera. O modelo de Poisson não é capaz de capturar bem as agregações locais, sendo a génese de uma classe de modelos para aglomeração espacial devida a Newman (1939) e o primeiro modelo genuíno de aglomerações de pontos espaciais devido a Newman e Scott (1958).

Matérn pode ser considerado o pai da estatística espacial moderna, sendo o seu trabalho de doutoramento (1960) dos documentos mais relevantes nesta área, ainda nos dias de hoje. Apesar de as suas contribuições serem mais reconhecidas na área dos dados referentes a pontos, também nos processos pontuais espaciais ele desenvolveu fundamentos importantes. Destacam-se os modelos para processos pontuais espaciais que apenas impõem uma distância mínima entre quaisquer dois pontos do processo (por exemplo, quando os pontos representam localizações de árvores em florestas densas), lançando as bases do que hoje designamos por processo pontual de Markov.

Foram Ripley e Kelly (1977) que definiram precisamente este processo pontual de Markov, que permanece a família de modelos mais vastamente utilizada para descrever padrões pontuais espaciais.

Strauss (1975) generalizou o modelo de Matérn acima descrito a uma classe de modelos em que a densidade conjunta de uma qualquer configuração de pontos numa dada região do espaço é proporcional a um parâmetro θ levantado ao número de pares de pontos distando menos do que a distância mínima do modelo de Matérn: $\theta = 0$ correspondendo ao modelo de Matérn, $\theta = 1$ ao processo de Poisson homogéneo e os modelos em que $0 < \theta < 1$ sendo úteis para padrões espacialmente mais regulares do que este último, mas sem uma restrição estrita de distância mínima. Strauss ainda pensou que talvez esta classe de modelos pudesse servir para modelar a aglomeração no caso de $\theta>1$, mas os trabalhos de Kelly e Ripley (1976) mostraram que tais valores não correspondiam a processos bem definidos no plano.

Ripley (1977) estabeleceu uma abordagem sistemática para analisar dados de processos pontuais espaciais, juntando o que anteriormente não pareciam ser trabalhos relacionados em modelação e análise de padrões espaciais incluindo modelos de processos pontuais especificados através de interacções locais, estatísticas resumo funcionais usadas como ferramentas de construção de modelos e métodos Monte Carlo para avaliar a bondade de ajustamento. A essência da abordagem de Ripley para o ajuste de modelos reside na comparação entre estatísticas resumo funcionais dos dados com realizações simuladas de um modelo proposto, sendo a ferramenta mais a si associada a função-K, uma estimativa não paramétrica das propriedades do momento de segunda ordem de um processo pontual espacial estacionário.

Nas últimas décadas têm-se assistido a um grande desenvolvimento da metodologia dos processos pontuais, suportado pelo enorme progresso teórico que conheceu esta área e conduzido pelas inúmeras aplicações de vários campos da ciência. São referências incontornáveis desta evolução Ripley (1977, 1981), Diggle (1983, 2003), Cressie (1993), Stoyan e Stoyan (1994), Stoyan, Kendal e Mekle (1995), Van Lieshout (2000), Moller e Waagepetersen (2004), Illian, Penttinen, Stoyan e Stoyan (2008), Gelfand, Diggle, Fuentes e Guttorp (2010) e Gaetan e Guyan (2010). Para uma breve introdução sobre o assunto veja, por exemplo, Carvalho e Natário (2008).

3. Dados referentes a pontos

Este ramo da análise espacial de dados é habitualmente conhecido por geoestatística e os contributos iniciais mais importantes surgiram de dois locais muito afastados um do outro, África do Sul e Suécia, em dois campos de aplicação também bem diferentes, produção mineira e exploração florestal, que passamos a descrever brevemente.

A origem da geoestatística está indissoluvelmente ligada ao método empírico de cálculo apresentado em 1951 pelo engenheiro de minas sul-africano D. J. Krige, para o problema da avaliação da rentabilidade de exploração de uma mina através da estimação da quantidade de minério aí existente feita com base em amostras retiradas de locais igualmente espaçados no terreno.

Este problema da predição espacial, foi mais tarde tratado do ponto de vista mais formal por Mathéron (1962) na École des Mines de Fontainebleau, França, sempre no contexto da exploração mineira, e por Gandin (1963) na União Soviética, no contexto da meteorologia, e desde então não tem deixado de conhecer enormes desenvolvimentos com a utilização, cada vez mais vulgar, de sofisticados meios computacionais e de representação geográfica fornecidos pelos Sistemas de Informação Geográfica (SIG).

Duma forma simples, pode dizer-se que se pretende predizer

$$T = \int_{\mathcal{D}} Y(s)ds,$$

em que $Y(s)$ é um processo estocástico espacial contínuo, que foi observado numa amostra de localizações, $s_i : i = 1, \dots, n$, de dimensão finita, do domínio $\mathcal{D}$. As observações podem ser feitas com ou sem erro sendo que, no primeiro caso, o modelo pode ser estendido de forma simples para os incorporar adicionando ao verdadeiro valor do processo nas diferentes localizações variáveis aleatórias independentes de média zero e variância τ^2, (interpretada no contexto do variograma como o efeito de pepita),

$$S_i = Y(s_i) + Z_i.$$

Os métodos de predição pontual de $Y(s)$ em localizações não conhecidas, são designados por kriging em honra ao homem que primeiro tratou o tema, e abrangem formas simples em que se supõe o processo $Y(s)$, com esperança constante μ e estrutura de covariância completamente conhecida, estimado pelo melhor preditor linear nas observações - kriging simples - e formas mais elaboradas em

que a esperança constante é substituída por um modelo de regressão $\mu(s) = \boldsymbol{X}(s)'\boldsymbol{\beta}$ com um vector de covariáveis também elas espacialmente referenciadas - kriging universal.

Estes métodos de predição óptima superam largamente os métodos de interpolação comuns, como o inicialmente usado por Krige, visto que incorporam a estrutura da correlação e permitem a predição da variabilidade das estimativas, ou seja, a sua precisão.

Na maior parte dos casos estes modelos são gaussianos e a sua completa especificação, desde que o valor médio seja dado por um modelo linear no máximo tão elaborado como o que acabámos de referir, repousa numa estrutura para a covariância espacial, que há que estimar também a partir dos dados.

A estimação da função de covariância ou, o que é equivalente no caso de estacionaridade forte, do semivariograma, $\gamma(\|s - s'\|) = var[Y(s) - Y(s')]$, é o coração da geoestatística. O primeiro estimador empírico do semivariograma foi proposto por Mathéron e é ainda hoje o mais comummente usado, mas existem outros estimadores empíricos cujas propriedades podem ser consultadas por exemplo em Cressie (1993). Contudo, para tratar o problema de predição que expusemos no início da secção, é evidentemente importante ajustar ao semi-variograma empírico uma função que, tal como as distribuições de probabilidade, pode pertencer a diferentes famílias paramétricas.

O problema do ajustamento de modelos teóricos ao variograma empírico tem evoluído bastante e, embora actualmente seja comum a utilização de métodos baseados na verosimilhança (máxima verosimilhança e máxima verosimilhança restrita), incluindo a abordagem bayesiana, que estimam todos os parâmetros do modelo conjuntamente, os métodos livres de distribuição como o dos mínimos quadrados ou mínimos quadrados generalizados continuam a ter grande popularidade por exigirem propriedades de estacionaridade espacial mais fracas.

Neste campo é imprescindível nomear Bertil Matérn (1960), que na sua dissertação de doutoramento no Colégio Real de Silvicultura da Suécia apresentou uma função paramétrica com características especiais para modelar a correlação espacial. Note-se que no caso particular em que os processos são espacialmente estacionários, a função de covariância é dada por,

$$cov[Y(s), Y(s')] = \sigma^2 \rho(d),$$

em que $\sigma^2 = var\,[Y(s)]$ e $\rho(d)$ é a correlação entre duas quaisquer variáveis do processo, $Y(s)$, $Y(s')$, e $d = \|s - s'\|$ é a distância entre s e s' e o modelo de Matérn modela esta função por

$$\rho(d) = \left[2^{\kappa-1}\,\Gamma(\kappa)\right]^{-1} (d/\phi)^{\kappa}\,\mathrm{K}_{\kappa}(d/\phi)\,,$$

em que $\mathrm{K}_{\kappa}(.)$ é a função de Bessel modificada de ordem $\kappa > 0$ e ϕ um parâmetro de escala que tem a dimensão da distância.

Existem muitos outros modelos de famílias de funções de correlação com um só parâmetro, algumas de utilização frequente no contexto desenvolvido pela escola de Mathéron, mas o modelo de Matérn, além de ter como casos particulares algumas dessas famílias, como os modelos exponencial e gaussiano, tem a vantagem da parte inteira do parâmetro κ representar a ordem de diferenciabilidade em média quadrática do processo, controlando desta forma a regularidade das respectivas trajectórias.

O trabalho de Matérn na área dos processos espaciais foi de uma grande riqueza. Na sua tese, além de outros resultados importantes já referidos na secção anterior, tratou a questão da amostragem espacial e da respectiva teoria assintótica fazendo uma clara distinção entre os casos em que um domínio constante de área $|\mathcal{D}|$ vai sendo preenchido por amostras de dimensão crescente que produz estimativas consistentes da média $T/|\mathcal{D}|$, e o caso em que o domínio vai sendo expandido pela observação de cada vez mais pontos que produz estimativas consistentes de μ, valor médio do processo.

Os métodos usados por Mathéron em França e os trabalhos de carácter mais teórico sobre predição em processos estocásticos desenvolvidos tanto por Matérn como por exemplo por Whitle (1963) mantiveram-se separados e a sua ligação só ficou bem clara com o trabalho de Ripley(1981). Mais recentemente, é no campo da abordagem hierárquica a muitos dos modelos gerais que, para além permitirem extensões a distribuições pertencentes á família exponencial, como por exemplo em Diggle *et al*. (1998), são mais facilmente ajustáveis via métodos bayesianos, que o problema central desta área, a predição espacial, mais tem progredido. Outras linhas de investigação situam-se na criação de modelos para processos não estacionários e anisotrópicos como são exemplos, Sampson e Guttorp (1992) e Ecker e Gelfand (2002).

4. Dados referentes a áreas

O terceiro tipo de problema que referimos diz respeito à dependência espacial entre dados agrupados por regiões que foi certamente percebido por R. A. Fisher durante o seu trabalho de agricultura experimental nas décadas 20 e 30 do século XX na Rothamsted Experimental Station em Inglaterra, que é referido como tendo dado origem ao planeamento de experiências.

Tratava-se de modelar a colheita de trigo em cada um dos 20×25 rectângulos contíguos com uma área de cerca de 8,5 m^2 cada. Um modelo simples para esta variável pode ser dado por

$$Y_{ij} = \mu + Z_{ij} \quad , \quad \text{i=1,…20; j=1, …,25,}$$

em que μ é a média da colheita de trigo por rectângulo e Z_{ij} são perturbações independentes com valor médio nulo. No entanto, Fisher apercebeu-se que estas perturbações relacionadas com a fertilidade do solo, exposição solar, inclinação, etc, não eram de facto independentes e que os rectângulos mais próximos tinham características mais similares que os mais afastados. Para combater e neutralizar (embora sem remover) o enviesamento provocado por este efeito espacial externo à experiência, Fisher (1966) advogou o princípio da divisão em blocos e da replicação com aleatorização no planeamento de experiências.

Uma forma moderna de abordar o mesmo problema seria o de adicionar ao modelo anterior a hipótese de que os Z_{ij} têm uma distribuição normal e usar a verosimilhança como base da inferência. O princípio dos blocos pode ser introduzido neste modelo incorporando uma componente indexada pelos blocos, com soma condicionada a zero, que representa a quantidade pela qual se espera que a colheita em cada bloco difira do valor médio global e a sua introdução pode ser olhada como uma forma de ajustamento através de uma covariável representando a variação espacial sistemática suposta constante em cada bloco.

Uma outra forma de resolver mesmo problema foi proposta por Papadakis (1937) através do ajustamento da colheita em cada bloco pela média das colheitas dos blocos vizinhos. Uma formulação moderna deste modelo é hoje bem conhecida sob o nome de Campos Aleatórios de Markov, de que falaremos mais tarde, e seria dada por,

$$Y_{ij}|\{Y_{kl} : (k,l) \neq (i,j)\} \sim \mathcal{N}\left(\mu + \beta\left(\bar{Y}_{\delta(i,j)} - \mu\right), \tau^2\right),$$

em que $\delta(i,j)$ representa o conjunto dos blocos vizinhos de (i,j) e $\bar{Y}_{\delta(i,j)}$ a média das respectivas colheitas.

Embora outros contributos tenham sido dados nos anos seguintes nomeadamente no que se refere à forma como a correlação espacial entre blocos deveria ser modelada, o maior progresso nesta matéria é apresentado o trabalho de Julian Besag (1974) em que foram propostos modelos e respectivos métodos de inferência para analisar dados que, embora conhecidos por dados em "quadrícula", se referem a um conjunto de áreas contíguas de formato regular ou irregular representadas pelas coordenadas geográficas dos seus centróides que constituem os nós da quadrícula.

Os modelos de Besag são conhecidos por autoregressivos condicionais (CAR) pois usam o valor da variável resposta em blocos com determinados espaçamentos como variáveis explicativas embora, ao

contrário dos modelos do mesmo nome para séries cronológicas que impõem restrições de linearidade, a regressão seja neles especificada pelo conjunto das distribuições da variável de interesse em cada um dos blocos do domínio condicionadas pelos valores da variável em todos os outros blocos.

A questão da não exigência da linearidade é a distinção primordial entre este tipo de modelos e os modelos clássicos chamados, no contexto da estatística espacial, modelos autoregressivos simultâneos (SAR). De facto, como refere, Gelfand *et al* (2010), embora sejam equivalentes os modelos unidimensionais temporais

$$Y_t = aY_{t-1} + Z_t\,;\;\; Z_t \sim\; \mathcal{N}(0, \sigma^2)$$

e

$$Y_t|\{Y_s: s < t\} \sim\; \mathcal{N}(a\, Y_{t-1}, \sigma^2),$$

o mesmo não se passa com os modelos unidimensionais espaciais,

$$Y_i = a(Y_{i-1} + Y_{i+1}) + Z_i;\;\; Z_i \sim\; \mathcal{N}(0, \sigma^2)$$

e

$$Y_i|\{Y_j: j \neq\; i\} \sim\; \mathcal{N}(a(Y_{i-1} + Y_{i+1}), \sigma^2),$$

em que a versão condicional do modelo simultâneo depende também de Y_{i-2} e Y_{i+2} .

Baseado em fenómenos da área da Física, Besag apontou neste trabalho que qualquer conjunto de distribuições condicionais completas que sejam mutuamente compatíveis determina exactamente uma distribuição conjunta, o que implica que a distribuição conjunta de um Campo Aleatório de Markov (MRF), conhecida por distribuição de Gibbs, possa ser sempre especificada localmente i.e. à custa das distribuições condicionadas apenas pelos seus vizinhos. Este resultado é de uma importância fundamental visto que permite simular através da amostragem de Gibbs a distribuição conjunta de um MRF que sabemos existir e ser única a partir da especificação local do processo como é, por exemplo referido na secção 3.4 de Carvalho e Natário (2008) e explica a enorme utilização que é actualmente feita destes modelos nas aplicações que envolvem dados espaciais discretos especialmente no contexto da modelação hierárquica de que falaremos seguidamente.

5. Desenvolvimentos recentes

Muitos dos desenvolvimentos recentes na área da estatística espacial foram motivados pelo desenvolvimento da especificação hierárquica de modelos com efeitos aleatórios.
Neste caso o processo estocástico de interesse não é observado directamente mas apenas indirectamente através de variáveis aleatórias que são especificadas condicionalmente ao processo latente construindo-se assim uma estrutura complexa de interdependência através da especificação de dependências locais simples.
Exemplos de aplicação deste tipo de modelos apareceram inicialmente em áreas tão distintas como o mapeamento de doenças (Clayton e Kaldor, 1987) e restauração de imagens (Besag, York e Mollié, 1991), mas a sua utilização é cada vez mais frequente e vasta devido à capacidade de se usarem métodos de Monte Carlo para resolver os problemas de inferência que lhes estão associados. De facto, a interpretação hierárquica é muito útil para ajustar modelos numa perspectiva bayesiana com a vantagem de evitar a inferência assintótica, por vezes totalmente inadequada, que está associada a outras abordagens baseadas na verosimilhança.
O grande desenvolvimento dos computadores pessoais e a disponibilidade de pacotes de software especialmente dirigido a este tipo de abordagem afastou as barreiras que se levantam habitualmente na inferência bayesiana e potenciou muito a sua utilização, embora subsistam problemas com aplicações a grandes conjuntos de dados.

Para terminar não se pode deixar de referir a importância que tem neste momento a modelação de processos que evoluem simultaneamente no espaço e no tempo cujo estudo dominará certamente o futuro da estatística espacial.

Na verdade, os modelos espaciais dão uma fotografia instantânea dum fenómeno, mas o objectivo fundamental de muitas das aplicações é a monitorização de fenómenos, como por exemplo os ligados ao ambiente ou à epidemiologia, que requer uma observação ao longo do tempo. Muitos desses fenómenos estão sujeitos global ou localmente a variações sazonais ou evoluções com tendência que tornam imprescindível a integração de uma componente temporal na estrutura de dependência espacial do modelo se, por exemplo, se pretende predizer o processo em locais não amostrados em instantes futuros.

Presentemente surgem numerosos trabalhos nesta área em qualquer dos três domínios da estatística espacial dos quais apresentamos apenas como exemplo Diggle (2006), Ma (2010) e Richardson (2006).

6. Referências

Besag, J.E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Série B*, **34**:192–236.

Besag, J.E., York, J. e Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of statistical Mathematics,* **4**:49-71.

Carvalho, M.L. e Natário, I. (2008). *Análise de Dados Espaciais*. Sociedade Portuguesa de Estatística.

Clayton, J.M. e Kaldor, D.G. (1987). Empirical bayes estimates of age standardized relative risks for use in disease mapping. *Biometrics*, **43**:671–681.

Cressie, N.A.C. (1993). *Statistics for spatial data*. John Wiley and Sons, New York.

Diggle, P. (1983, 2003). *Spatial point patterns.* Arnold, London.

Diggle,P.J., Tawn, J.A. e Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, **47**: 299-350.

Diggle, P. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical Methods in Medical Research***, 15** (4): 325-336.

Ecker, M.D. e Gelfand, A.E. (2002). Spatial modeling and prediction under stationary non-geometric range anisotropy. *Environmental and Ecological Statistics*, **10** (2):165-178.

Fisher, R.A. (1966). *The design of experiments*. 8th ed., Oliver and Boyd, London.

Gaetan, C. e Guyan, X. (2010). *Spatial Statistics and Modeling*. Springer.

Gandin, L.S. (1963). *Objective analysis of meteorological field.* Technical report, Gidrometeorologicheskoe Izdate'stvo.

Gelfand, A.E., Diggle, P., Fuentes, M. e Guttorp, P. (2010), eds. *Handbook of Spatial Statistics*. Chapman & Hall/CRC.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley.

Kelly, F.P. e Ripley, B.D. (1976). A note on Strauss' model for clustering. *Biometrika*, **63**: 357-360.

Ma, C. (2010). A class of variogram matrices for vector random fields in space and/or time. *Mathematical Geosciences,* **42**. Online First™, 8 September.

Matérn, B. (1960). *Spatial variation*. Technical report, Meddelanden fran Statens Skogsforskninginstitut, Stockholm.

Matérn, B. (1986). *Spatial Variation.* 2nd ed. Springer-Verlag, Berlin.

Mathéron, G. (1962). *Traité de Géostatistique Appliquée*. Editions Technip.

Moller, J. e Waagepetersen, R.P. (2003). *Statistical inference and simulation for spatial point processes*. Chapman & Hall, Boca Raton.

Newcomb, S. (1859-60). Notes on the theory of probabilities. *Mathematical Monthly*, **1** e **2**.

Newman, J. (1939). On a new class of contagious distributions, applicable in entomology and bacteriology. *Annals of Mathematical Statistics*, **10**:35-57.

Newman, J. e Scott, E.L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society, Série B*, **20**: 1-43.

Papadakis, J.S. (1937). Méthode statistique pour des expériences sur champ. *Bull.Inst. Amél. Plantes á Salonique,* Grèce, **23**.

Richardson, S. (2006). Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical Methods in Medical Research*, **15** (4); 385-407.

Ripley, B.D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Série B*, **39**: 172-212.

Ripley, B.D. (1981). *Spatial statistics*. Wiley, New York.

Ripley, B.D. e Kelly, F.P. (1977). Markov point processes. *Journal of the London Mathematical Society*. **15:** 188-192.

Sampson, P. e Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association, **87** (417): 108-119.

Snow, J. (1855). On the Mode of Communication of Cholera. 2nd edition, London :John Churchill, New Burlington Street, England.

Stoyan, D. e Stoyan. H. (1994). *Fractals, random shapes and point fields*. Wiley, New York.

Stoyan, D., Kendal, W.S. e Mekle, J. (1995). *Stochastic geometry and its applications.* 2nd ed., Wiley, New York.

Strauss, D.J. (1975). A model for clustering. *Biometrika*, **62**: 467-475.

van Lieshout, M.N.M. (2000). *Markov point processes and their applications*. Imperial College Press, London.

Whitle, P. (1963). *Prediction and regulation*. 2nd ed., Minneapolis: University of Minnesota.

O INE e os desafios dos sistemas estatísticos nacional e europeu

Jorge M. Mendes, Humberto Pereira, Maria João Zilhão
jorge.mendes@ine.pt, humberto.pereira@ine.pt, mjoao.zilhao@ine.pt

Instituto Nacional de Estatística

1. Um pouco de história

«O Instituto Nacional de Estatística (INE) foi criado em 23 de Maio de 1935. Foram-lhe atribuídas "funções de notação, elaboração, publicação e comparação dos elementos estatísticos referentes aos aspectos da vida portuguesa que interessam à Nação, ao Estado e à ciência" (Lei nº 1911, de 23 de Maio de 1935). A sua criação é entendida como o culminar de um longo processo de centralização do sistema estatístico nacional, cujas raízes remontam até ao século XVIII.

O INE vinha consubstanciar a "Ordem e a Razão" na produção e difusão dos números necessários à "boa governação", ou seja, contribuir para "tirar o Governo do País do empirismo em que tinha caído", beneficiando para o efeito da tradição de centralismo político-administrativo vivido em Portugal desde a sua fundação.» Com efeito, foi no século XVIII, mais precisamente em 1762, que foi criado um primeiro organismo, o Erário Régio, responsável pela produção regular anual das contas públicas. À semelhança do que acontece actualmente, as necessidades não ficaram por aqui, e as competências desse organismo foram complementadas com outras, designadamente na produção das balanças do comércio, e na realização de contagens da população. A título de curiosidade refira-se que o primeiro recenseamento populacional realizado com excepcional rigor ocorreu no século XIX, em 1801, no Continente, e em 1806 na Madeira.

A instabilidade vivida em Portugal durante o século XIX prejudicou a regularidade da produção estatística, embora tivesse sido criada, em 1815, a Comissão de Estatística e Cadastro do Reino. A estabilização da produção e publicação da informação estatística ocorreu apenas no último terço do século XIX, contribuindo para uma ampla divulgação da informação estatística disponível para a generalidade da população (o leitor mais curioso pode complementar esta leitura em http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_publicacoes&PUBLICACOESpub_boui=91259840&PUBLICACOESmodo=2).

Foi um conjunto de disposições legais adoptadas até finais da década de 50 do século XIX que permitiu a criação do modelo de organização dos serviços estatísticos baseado «num órgão técnico que deveria assegurar a coerência e uniformidade das estatísticas e na produção descentralizada de estatísticas, embora com um organismo especializado de produção e difusão de informação estatística, que viria a prevalecer até ao período entre as duas guerras mundiais». Embora com frequentes alterações de pormenor, de denominações e de tutela, o sistema assim continuou até finais da década de vinte do século XX, altura em que fruto de uma reorganização do Ministério da Finanças, o sistema estatístico passou a ser constituído por um Conselho Superior de Estatística, pela Direcção-Geral de Estatística e pelas repartições e serviços estatísticos dos vários ministérios e pelas comissões distritais de estatística. Em 1929 foi tomado um conjunto de medidas importantes de centralização estatística que viria a culminar com a criação do INE em 1935, em substituição da Direcção-Geral de Estatística, e do estabelecimento de um conjunto de princípios sobre os quais deveria assentar o funcionamento do

sistema estatístico. Paralelamente à sua criação e autonomização, o INE foi instalado no edifício, projectado pelo Arquitecto Pardal Monteiro, onde ainda hoje se mantém a sua sede.

O edifício-sede do Instituto Nacional de Estatística, concluído em 1935.

À criação do INE seguiu-se um período de autonomização e afirmação e de resposta à crescente exigência de estatísticas nos mais diversos domínios, pese embora, as alterações de enquadramento institucional, a Segunda Guerra Mundial, as sucessivas reformas e o impacto da integração europeia.
Hoje, o INE faz parte integrante do Sistema Estatístico Europeu, adoptou os princípios do Código de Conduta para as Estatísticas Europeias e goza de independência técnica e administrativa e as estatísticas que produz regem-se por elevados padrões de qualidade e seguindo metodologias conhecidas e definidas de acordo com as melhores práticas internacionais.

2. Enquadramento e organização

Actualmente, o papel do INE na sociedade portuguesa, embora seja fruto da sua missão histórica, está vertido na Lei nº 22/2008, de 13 de Maio, comummente conhecida como Lei do Sistema Estatístico Nacional (SEN). A sua organização interna rege-se pelo Decreto-Lei nº 166/2007, de 3 de Maio e pelas Portarias nº 662-H/2007, de 31 de Maio e nº 839-B/2009, de 31 de Julho, a protecção de dados pessoais (Lei nº 67/98, de 26 de Outubro) e o posicionamento do INE no seio da administração indirecta do Estado (Lei nº 202/2006, de 27 de Outubro).

A actual Lei do SEN mantém, por um lado, o quadro formalizado em 1935 e entretanto estabilizado e enfatiza alguns aspectos, como seja que as estatísticas oficiais são produzidas com independência técnica e consideradas como um bem público, devendo respeitar os padrões nacionais e internacionais de qualidade estatística, bem como satisfazer as necessidades dos utilizadores de forma eficiente e sem sobrecarga excessiva para os fornecedores de informação, nomeadamente através da crescente utilização dos dados administrativos, podendo exigir o fornecimento, com carácter obrigatório e gratuito, a todos os serviços ou organismos, pessoas singulares e colectivas, de quaisquer elementos necessários à produção de estatísticas oficiais e estabelecer a recolha de dados que, ainda que não relevantes para a actividade específica das entidades obrigadas ao seu fornecimento, se revistam de importância estatística.

O SEN compreende o Conselho Superior de Estatística (CSE), órgão do Estado que orienta e coordena o sistema; O INE, IP, órgão central da produção e difusão de estatística oficiais que assegura a supervisão e coordenação técnico-científica do SEN; o Banco de Portugal no âmbito das suas

atribuições de recolha e elaboração de estatísticas monetárias, financeiras, cambiais e da balança de pagamentos, os Serviços Regionais de Estatística das Regiões Autónomas dos Açores e da Madeira que funcionam, em relação às estatísticas oficiais de âmbito nacional, como delegações do INE, IP, mas constituem autoridade estatística na produção de estatísticas oficiais de interesse exclusivo da região; e as entidades públicas produtoras de informação estatística oficial por delegação de competências do INE, IP.

O sistema estatístico português integra-se naturalmente no Sistema Estatístico Europeu (SEE), cujo regime é estabelecido pelo Regulamento (CE) nº 223/2009 do Parlamento Europeu e do Conselho de 11 de Março de 2009. Este Regulamento define o SEE «como uma parceria entre o Eurostat e as autoridades estatísticas nacionais, com vista ao desenvolvimento, produção e divulgação das estatísticas europeias», cabendo ao INE, IP o papel de interlocutor nacional da Comissão (Eurostat) para as questões relacionadas com as estatísticas. As estatísticas europeias são, as estatísticas necessárias para o desempenho das actividades da Comunidade, sendo determinadas pelos Programas Estatísticos Europeu, (quinquenais e anuais).
O desenvolvimento, produção e divulgação das estatísticas europeias rege-se pelos princípios da *Independência Profissional, Imparcialidade, Objectividade, Fiabilidade, Segredo Estatístico e Relação Custo-benefício*. Estes princípios estão desenvolvidos no Código de Conduta para as Estatísticas Europeias, adoptado pela Comissão e aceite pelos Estados-Membros. O Código de Conduta para as Estatísticas Europeias, é um instrumento auto-regulador, cujo objectivo fundamental é melhorar a confiança nas autoridades estatísticas dos Estados Membros e do EUROSTAT, reforçando a sua independência, integridade e responsabilidade e robustecer a qualidade das estatísticas europeias, sendo composto por quinze princípios, repartidos por três áreas:

Áreas	Princípios
Enquadramento Institucional	1. Independência Profissional 2. Mandato para a recolha de dados 3. Adequação de Recursos 4. Compromisso com a Qualidade 5. Confidencialidade Estatística 6. Imparcialidade e Objectividade
Processo Estatístico	7. Metodologia sólida 8. Procedimentos Estatísticos adequados 9. Carga não excessiva sobre os respondentes 10. Eficácia na utilização dos recursos
Produção Estatística	11. Relevância 12. Precisão e Fiabilidade 13. Oportunidade e pontualidade 14. Coerência e comparabilidade 15. Acessibilidade e clareza

3. Os domínios das estatísticas oficiais

A produção de informação estatística em Portugal enquadra-se na moldura de médio prazo estabelecida pelas Linhas Gerais da Actividade Estatística Nacional 2008-2012, que designa os objectivos estratégicos e respectivas prioridades, e pelos planos de actividade anuais do INE e das outras entidades intervenientes na produção estatística nacional

(http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_cont_inst&INST=386983), cuja preparação considera, igualmente, os respectivos Programas Estatísticos Europeu.
O Plano de Actividades anual descreve os desenvolvimentos previstos na cadeia de produção de estatísticas oficiais, nomeadamente aos níveis da Recolha de dados, da Metodologia Estatística e Tecnologias de Informação e Comunicação, Difusão das estatísticas oficias e Avaliação e Gestão da Qualidade, descrevendo, igualmente, o quadro de estatísticas oficiais segundo vários domínios:

- **População e Sociedade**, que engloba cinco subdomínios: População, Trabalho, Emprego e Desemprego, Rendimento e Condições de Vida, Educação, Formação e Aprendizagem, Justiça e Saúde e Incapacidades;
- **Território e Ambiente**, que engloba dois subdomínios: Território e Ambiente;
- **Economia e Finanças**, que engloba quatro subdomínios: Contas Nacionais, Conjuntura Económica e Preços e Empresas;
- **Comércio internacional**
- **Agricultura, Floresta e Pescas**, que envolve dois subdomínios: Agricultura e Florestas e Pescas;
- **Indústria, Energia e Construção**, que engloba dois subdomínios: Indústria e Energia e Construção;
- **Serviços**, que engloba três subdomínios: Comércio Interno, Transportes e Turismo;
- **Inovação e Conhecimento**, que engloba dois subdomínios: Sociedade da Informação e Ciência e Tecnologia;

A difusão da informação estatística privilegia a Internet como principal modo de acesso a dados estatísticos através do Portal do Instituto Nacional de Estatística na internet (www.ine.pt), onde é possível encontrar a informação institucional, informação estatística propriamente dita (sob a forma de destaques, publicações, ou base de dados de indicadores) referente aos domínios elencados, metainformação (classificações, conceitos, etc.) e informação de apoio ao utilizador. Por outro lado, o INE dispõe de bibliotecas na sua sede e respectivas delegações (Porto, Coimbra, Évora e Faro) e, uma vez que a ligação entre o SEN e a Academia tem sido historicamente forte, coloca ao serviço dos utilizadores a Rede de Informação do INE em Bibliotecas do Ensino Superior (RIIBES), disponível em mais de 30 estabelecimentos de ensino, como fonte de difusão para fins de estudo e investigação.

3.1. População e Sociedade

Neste grande domínio convivem temas tradicionais e objecto de produção desde a fundação do INE, temas que foram ganhando gradualmente o seu espaço na actividade estatística, até temas emergentes cuja atenção por parte do SEN se deve ao seu papel fundamental no conhecimento das sociedades contemporâneas. Este domínio encontra-se alinhado com o domínio das estatísticas demográficas e sociais previsto no Programa Estatístico Europeu.

O subdomínio respeitante à População abrange os dados dos recenseamentos da população e da habitação, bem como indicadores sobre a dinâmica demográfica da população. Neste subdomínio assume particular relevância as actividades de preparação dos Recenseamentos da População e da Habitação em 2011.

O subdomínio Trabalho, Emprego e Desemprego abrange as estatísticas do mercado de trabalho produzidas pelo INE, tendo por fonte o Inquérito ao Emprego (com periodicidade trimestral) e cujo modo de recolha se encontra em fase de transição de entrevista pessoal presencial para entrevista telefónica, bem como as estatísticas da responsabilidade delegada no Gabinete de Estratégia e Planeamento do Ministério do Trabalho e Solidariedade Social.

O subdomínio Rendimento e Condições de Vida engloba as estatísticas sobre rendimento, pobreza e desigualdade, tendo por fonte o Inquérito às Condições de Vida e Rendimento (com periodicidade anual), bem como outros indicadores baseados em dados administrativos ou em inquéritos com

periodicidade supra-anual. A este propósito refira-se o Inquérito à Situação Financeira das Famílias, em resultado de uma parceria com o Banco de Portugal, e o Inquérito às Despesas das Famílias, com periodicidade quinquenal, e que é a fonte de construção do cabaz de consumo das famílias que serve de base ao cálculo do Índice de Preços no Consumidor e, por conseguinte, da taxa de inflação.

O subdomínio Educação, Formação e Aprendizagem as estatísticas da educação, formação e aprendizagem produzidas pelo INE com base no Inquérito à Educação e Formação de Adultos (com periodicidade quinquenal), cuja próxima edição está prevista para 2011, bem como as estatísticas da educação elaboradas sob competência delegada pelo INE no Gabinete de Estatística e Planeamento da Educação do Ministério da Educação.

O subdomínio Justiça abrange apenas as estatísticas que são elaboradas sob competência delegada na Direcção-Geral da Política de Justiça do Ministério da Justiça.

O subdomínio Saúde e Incapacidades abrange as estatísticas elaboradas com a intervenção da Direcção-Geral de Saúde, do Ministério da Saúde, que versam sobre indicadores relacionados com o sistema de saúde, e o Instituto de Informática do Ministério do Trabalho e Solidariedade Social, que versam sob protecção social.

3.2. Território e Ambiente

À tradicional produção de indicadores estatísticos sobre o território junta-se, neste domínio, o relativamente emergente tema do ambiente e avaliação da situação ambiental. Enquanto que o subdomínio Território abrange uma quantidade significativa de indicadores físicos sobre o território nacional, bem como o Índice de Desenvolvimento Regional e índice parciais de competitividade, coesão e qualidade ambiental, em resultado de uma parceria com o Departamento e Prospectiva e Planeamento e Relações Internacionais do Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional, o subdomínio Ambiente abrange um conjunto de indicadores sobre água e resíduos que resultam, quer do aproveitamento de dados administrativos junto de organismos da Administração Central, quer dos municípios, quer de inquéritos realizados pelo INE, IP junto de organizações não-governamentais de ambiente (e.g. IONGA), de municípios (e.g. IMPA) e de empresas (e.g. IEPA e IBSA).

3.3. Economia e Finanças

O domínio Economia e Finanças engloba temas tradicionais, como a contabilidade nacional, cujo desenvolvimento ocorreu durante a segunda metade do século XX, as estatísticas do preços, que assumiram um papel instrumental na governação moderna e cuja necessidade se verificou imediatamente após a Primeira Guerra Mundial, bem como as estatísticas das empresas cujo desenvolvimento ocorreu após a década de 30 do século passado, e que são fundamentais hoje em dia no acompanhamento das economias modernas e dos fenómenos como a globalização e o empreendedorismo.

O subdomínio Contas Nacionais engloba as contas nacionais anuais, trimestrais e sectoriais, bem como as contas regionais e as contas satélite. A produção estatística das Contas Nacionais encontra-se centralizada no INE, IP.

O subdomínio Conjuntura Económica e Preços engloba a produção de estatísticas de curto prazo, nomeadamente a produção de índices, quantitativos e qualitativos, de acompanhamento da actividade económica em termos globais e por sectores, bem como a produção do índice que serve de base ao acompanhamento da inflação, o Índice de Preços no Consumidor.

O subdomínio das empresas abrange todas as estatísticas das empresas em geral. É importante referir que uma parte destas estatísticas é baseada no aproveitamento de dados que resultam de actos administrativos cujo desenho incorporou as necessidades do SEN neste domínio.

3.4. Comércio Internacional

Por razões que se prendem com o acompanhamento da evolução das economias, as estatísticas das trocas comerciais com o exterior foram aquelas cujo desenvolvimento de iniciou mas precocemente. Actualmente, num contexto de União Económica e Monetária e abolição de fronteiras, o acompanhamento do comércio com o exterior assume uma importância crescente no contexto das estatísticas macroeconómicas. Este domínio abrange assim todos os indicadores de acompanhamento da actividade comercial com o exterior, União Europeia e resto do Mundo.

3.5. Agricultura, Floresta e Pescas

O domínio da Agricultura, Floresta e Pescas abrange as actividades económicas tradicionalmente ligadas ao sector primário, cuja atenção por parte das estatísticas oficiais se deu a partir dos finais do século XIX, por manifesta importância económica deste sector. Trata-se de um domínio claramente alinhado com o estabelecido no Programa Estatístico Europeu e fundamental ao planeamento da Política Agrícola Comum e das intervenções comunitárias ao nível dos apoios financeiros.

O subdomínio da Agricultura e Floresta, para além das estatísticas agrícolas correntes, integra ainda o Recenseamento Agrícola, realizado decenalmente, cujos resultados preliminares, referentes ao ano de 2009, deverão ser divulgados em 2011. Um outro produto estatístico de grande interesse e fundamental para o conhecimento das disponibilidades alimentares e nutricionais do país, assumindo-se como um quadro alimentar global, expresso em consumos brutos médios diários, traduzidos em calorias, proteínas, hidratos de carbono, gorduras e álcool é a Balança Alimentar Portuguesa que ainda em 2010 deverá divulgar os resultados para o quinquénio 2003-2008.

A produção das estatísticas do subdomínio das Pescas é realizada pelo INE e por competência delegada da Direcção-geral das Pescas e Aquicultura do Ministério da Agricultura, do Desenvolvimento Rural e das Pescas.

3.6. Indústria, Energia e Construção

À semelhança do que aconteceu com o sector primário, também as actividades económicas do sector secundário, receberam uma atenção particular das estatísticas oficiais, que a crescente industrialização do País a partir de finais do século XIX, determinou.

Actualmente o subdomínio da Indústria e Energia compreende as estatísticas produzidas pelo INE e sob competência delegada por parte da Direcção-Geral de Energia e Geologia do Ministério da Economia, Inovação e do Desenvolvimento. Este domínio dedica-se, entre outros, à produção de indicadores de volume da actividade económica no sector da indústria transformadora e à produção de indicadores de produção e consumo dos vários tipos de energia utilizada em Portugal. O domínio das energias alternativas é um tema emergente. Está prevista a realização de o Inquérito ao Consumo de Energia no Sector Doméstico cujo objectivo será não só recolher informação sobre o consumo de energia pelas famílias portuguesas, mas também sobre a eficiência energética e fonte de energia utilizadas.

O subdomínio Construção compreende as estatísticas incluídas no Sistema de Indicadores das Operações Urbanísticas.

3.7. Serviços

O domínio dos Serviços fecha o triângulo abrangendo as actividades económicas do sector terciário, cujo desenvolvimento acompanhou o crescimento deste sector desde a década de 60 do século XX.

O subdomínio Comércio Interno abrange as estatísticas correntes do comércio por grosso e a retalho.

O subdomínio Transportes abrange as estatísticas dos transportes aéreo, marítimo e fluvial, ferroviário e rodoviário, sob a perspectiva quer do transporte de mercadorias, quer de passageiros.

O subdomínio do Turismo compreende a produção das estatísticas correntes sobre a procura turística interna e sobre a capacidade de oferta, bem como as estatísticas produzidas pelo Turismo de Portugal, IP, do Ministério da Economia, da Inovação e do Desenvolvimento.

3.8. Inovação e Conhecimento

O domínio Inovação e Conhecimento engloba um conjunto de temas recentes e emergentes na sociedade portuguesa, em geral, e nas estatísticas oficiais, em particular. No Programa Estatístico Europeu, este domínio surge integrado no domínio heterogéneo das estatísticas multi-temáticas.

O subdomínio da Sociedade da Informação abrange as sucessivas vagas anuais dos inquéritos à utilização das tecnologias de informação e comunicação em diferentes sectores: famílias, empresas e administração pública. Em 2010 está previsto o alargamento ao sector dos hospitais. Nesta área intervém com o INE, sob delegação de competências, a UMIC - Agência para a Sociedade do Conhecimento do Ministério da Ciência, Tecnologia e Ensino Superior.
O subdomínio da Ciência e Tecnologia abrange as estatísticas correntes sobre inscritos e diplomados no ensino superior e as estatísticas da ciência e tecnologia produzidas sob delegação de competências pelo Gabinete de Planeamento, Estratégia, Avaliação e Relações Internacionais (GPEARI) do Ministério da Ciência, Tecnologia e Ensino Superior.

4. O INE: desafios do passado, desafios do futuro

No actual contexto português, europeu e internacional existe com certeza um número infindável de desafios que se colocam ao SEN, em geral, e ao INE, em particular, quer seja num futuro próximo ou longínquo. Seria impossível elencar todos, por razões de espaço. Não poderia, contudo, deixar de se fazer referência a alguns que se assumem como particularmente relevantes: (a) garantir o cumprimento dos princípios de um sistema estatístico moderno; (b) assegurar a produção de informação estatística capaz de dar respostas às exigências contemporâneas; (c) garantir o salutar funcionamento de um sistema estatístico descentralizado e daí retirar partido; (d) assegurar um sistema estatístico eficiente e funcional do ponto de vista do seu financiamento sustentável a longo prazo.

Comecemos pelos desafios de assegurar o cumprimento dos princípios de uma sistema estatístico moderno. A Lei 22/2008, de 13 de Maio, no seu Capítulo II, estabelece um conjunto de princípios fundamentais do SEN, dos quais se destacam três, a saber: autoridade estatística, independência técnica e segredo estatístico. O princípio da autoridade estatística, constante do artigo 4º, estabelece no nº 1 que "*As autoridades estatísticas, no respectivo âmbito de actuação, podem exigir o fornecimento, com carácter obrigatório e gratuito, a todos os serviços ou organismos, pessoas singulares e colectivas, de quaisquer elementos necessários à produção de estatísticas oficiais e estabelecer a recolha de dados que, ainda que não relevantes para a actividade específica das entidades obrigadas ao seu fornecimento, revistam importância estatística*." Trata-se de um princípio caro ao INE, que pretende fazer face à tradicional compartimentação e sentido de propriedade que as autoridades administrativas possuem sobre toda a informação que está sob a sua alçada. Compete ao INE, como autoridade estatística, criar um clima propício a que as resistências existentes sejam realmente ultrapassadas com benefícios claros para o aparelho produtivo da informação estatística oficial e benefícios claros para o erário público e, sobretudo, na redução da carga estatística para os respondentes. Em segundo lugar, no nº 1 do artigo 5º estabelece-se que "*As estatísticas oficiais são produzidas com independência técnica, sem prejuízo do cumprimento das normas emanadas do Sistema Estatístico Nacional ou do Sistema Estatístico Europeu*." No nº 2 do mesmo artigo define-se a independência técnica: "*A independência técnica consiste no poder de definir livremente os métodos, normas e procedimentos estatísticos, bem como o conteúdo, forma e momento da divulgação da informação*". Assegurar cabalmente a independência técnica implica que INE possua e consiga manter um corpo de recursos humanos qualificados que possam livremente discernir e decidir sobre as

metodologias mais adequadas à prossecução da sua missão ou assegurar uma eficiente colaboração com a comunidade científica sempre que seja necessário recorrer à sua *expertise* ou transferir conhecimento. Por fim, mas não menos importante, despindo o princípio dos seus aspectos de maior tecnicidade, o "*segredo estatístico visa salvaguardar a privacidade dos cidadãos e garantir a confiança no SEN*." (nº 1 do artigo 6º) e "*Todos os dados estatísticos individuais recolhidos pelas autoridades estatísticas são de natureza confidencial*,…" (nº 2 do artigo 6º). Assegurar em permanência este princípio face à pressão externa de divulgação extensiva de informação estatística oficial assume-se como talvez o principal desafio futuro do INE e das outras entidades produtoras de informação estatística oficial, sob pena de não conseguir cumprir cabalmente a missão e os objectivos a que se propõe nos contextos português, europeu e internacional.

Ao contrário do que se verificou no passado em que a produção de informação estatística se ficava pelos domínios tradicionais necessários à administração directa do Estado, hoje em dias os domínios são crescentes e determinados por necessidades dos agentes económicos e sociais, de que o Estado é apenas um deles. Será fundamental para o SEN, em geral, e para o INE em particular antecipar as necessidades, quer em termos de domínios emergentes, quer em termos de quantidade e qualidade de informação estatística em domínios já existentes, dando resposta atempada com estatísticas fundamentais à gestão da vida social. A este propósito refira-se o papel construtivo que o Conselho Superior de Estatística, de que o INE é um membro fundamental, pode desempenhar no futuro.

O Instituto Nacional de Estatística, não obstante possuir o ónus da formulação e controle dos padrões a que deve obedecer a actividade de produção estatística em entidades externas ao INE, delega actualmente a produção da informação estatística em organismos do Ministério da Justiça, Educação, Trabalho e Solidariedade Social, Ministério da Agricultura, Desenvolvimento Rural e Pescas, Ministério da Economia, Inovação e Desenvolvimento, Ministério da Saúde e Ministério da Ciência, Tecnologia e Ensino Superior. É de esperar que, no futuro, os domínios da produção se alarguem com efeito imediato nas áreas alvo da delegação de competências. Será um desafio permanente para o INE a operacionalização de um sistema estatístico descentralizado, não descorando a garantia das condições necessárias para que a produção estatística seja de qualidade.

Existe ainda um desafio fundamental no futuro do SEN. Trata-se do aproveitamento de dados administrativos para fins estatísticos.

A administração moderna dos estados proporciona um manancial de informação infindável sobre os agentes sociais e económicos. Numa palavra, muito do que se recolhe ou se utiliza como instrumento de apoio à recolha nos domínios das estatísticas oficiais pode já existir algures numa base de dados como resultado de um qualquer acto administrativo. Obviamente, que conceitos, nomenclaturas e definições utilizados para fins estatísticos são, em muitas situações, divergentes dos congéneres criados para dar resposta aos objectivos administrativos. Todavia, a criação de novas bases de dados ou a adaptação das já existentes para que levem em consideração as necessidades do SEN, ou a aproximação dos conceitos, nomenclaturas e definições estatísticas aos congéneres administrativos poderá atenuar este problema.

O aumento crescente desta apropriação por parte do SEN tem potencialidades e riscos, uma vez que permitindo a recolha de informação estatística sob condições controladas e reduzindo, em simultâneo, a carga estatística sobre os prestadores de informação e o custo de produção, pode potenciar a desconfiança do público sobre um dos princípios básicos do sistema estatístico, o segredo estatístico, pondo em causa a fiabilidade da informação prestada e o funcionamento dos aparelhos administrativo e estatístico.

Nos próximos anos o INE enfrentará o desafio de perseguir este objectivo, como aliás a Lei do SEN prevê já, assegurando que se mantém a confiança do público no sistema estatístico. A este propósito refira-se o exemplo da Informação Empresarial Simplificada (IES) que reuniu num só acto quatro obrigações legais das empresas portuguesas, tendo sido desenhado justamente para dar respostas às necessidade da Direcção Geral de Contribuições e Impostos, Instituto de Registos e Notariado, Banco

de Portugal e Instituto Nacional de Estatística. O sistema criado com a IES permitiu que a fonte administrativa fosse criada (aproveitando o melhor que já existia na DGCI) com o propósito de satisfazer cabalmente, também, as necessidades estatísticas. Com a IES, o INE acede a um manancial de informação estatística de base, adaptada aos requisitos estatísticos, conferindo maior consistência à produção de estatísticas na área das empresas e contas nacionais. As vantagens para as estatísticas e a reacção do tecido empresarial à criação da IES levam a que se considere que faz todo o sentido ser replicado noutros domínios, assegurando as condições necessárias ao bom funcionamento do aparelho estatístico e administrativo. A IES tem sido tratado como um caso de estudo, tratando-se de um exemplo único a nível mundial que começa a ser estudado com vista a ser adoptado por outros países, como é o caso do Canadá e da Bulgária.

Analistas de dados seduzidos pelas potencialidades do R

Ashlee Vance (publicado no New York Times online a 6 de Janeiro de 2009[1])

Tradução e adaptação de Ana M. Pires, *apires@math.ist.utl.pt* e Conceição Amado[2]

O R surgiu em 1996, quando os professores de estatística Robert Gentleman (à esquerda) e Ross Ihaka (à direita)

distribuíram a primeira versão de utilização livre deste software. (Stuart Isett for The New York Times)

Para algumas pessoas R é apenas a 18ª letra do alfabeto. Para um físico pode representar a unidade de medida de radiação ionizante ou a resistência eléctrica, enquanto que para um matemático pode representar o conjunto dos números reais. Mas R é também o nome de uma linguagem de programação muito popular usada por um número crescente de pessoas que se dedicam à análise de dados, quer em ambiente empresarial, quer académico. O R está a transformar-se numa língua franca, em parte porque a exploração de dados (data mining) parece ter atingido a idade de ouro, sendo uma metodologia usada para tarefas que vão desde fixar preços de publicidade, ou descobrir mais rapidamente novos medicamentos até à afinação de modelos financeiros. Empresas tão diversas como a Google, a Pfizer, a Merck, o Bank of America, o grupo InterContinental Hotels e a Shell utilizam o R.

O R encontrou rapidamente adeptos porque estatísticos, engenheiros e cientistas sem grandes conhecimentos de programação conseguem utilizá-lo facilmente.

"O R é realmente importante a um nível que se torna difícil sobrevalorizar," diz Daryl Pregibon, um cientista que faz investigação para a Google e que usa amplamente este software. "O R permite aos estatísticos realizar análises muito complexas sem conhecer as entranhas dos sistemas computacionais."

Além do mais é grátis. O R é um programa de código aberto, e a sua popularidade reflecte também uma mudança de atitude em relação ao tipo de software utilizado pelas empresas. Os programas de código aberto podem ser utilizados e modificados por qualquer pessoa livremente. A IBM, a Hewlett-Packard e a Dell têm lucros anuais de milhões de dólares provenientes da venda de servidores que utilizam o sistema operativo de código aberto Linux, um concorrente do Windows da Microsoft. A maioria dos sítios Web são exibidos usando uma aplicação de código aberto chamada Apache, e as empresas dependem cada vez mais da base de dados de código aberto MySQL para armazenar as suas informações críticas. Muitas pessoas têm acesso aos resultados finais de toda esta tecnologia através do navegador Firefox, também um software de código aberto.

---

[1] Outra versão deste artigo foi publicada a 7 de Janeiro de 2009 na página B6 da edição de Nova Iorque.

[2] Quando o Editor do Boletim nos convidou para escrevermos uma "contribuição sobre o R", a ideia era fazer um artigo técnico que mostrasse as potencialidades do R em modelos avançados/complexos. Acidentalmente, durante as nossas pesquisas, deparámos com este artigo, que demonstra muito melhor do que qualquer uma de nós poderia fazer, e duma forma que diríamos entusiástica, o imenso potencial deste software. Achámos que devíamos partilhar este entusiasmo com todos os sócios da SPE e assim resolvemos substituir parte da nossa contribuição por esta tradução.

Nota do Editor: Este trabalho, programado o Boletim de 2009, apenas é publicado agora devido ao atraso no processo de autorização editorial.

O R é semelhante a outras linguagens de programação, tais como C, Java ou Perl, na medida em que permite aos utilizadores a realização de uma grande variedade de tarefas computacionais através do acesso a vários comandos. Contudo, para os estatísticos, o R é particularmente útil porque contém uma série de mecanismos integrados para organizar dados, executar cálculos e criar representações gráficas. Algumas pessoas familiarizadas com o R descrevem-no como uma versão super equipada da folha de cálculo Excel que permite descobrir tendências nos dados com muito mais clareza do que é possível fazer a partir de informações organizadas em linhas e colunas.

O que torna o R tão útil — e ajuda a explicar a sua rápida aceitação — é que estatísticos, engenheiros e cientistas podem melhorar o código existente ou escrever variantes para tarefas específicas (os chamados pacotes ou packages). Encontram-se assim pacotes específicos para correr algoritmos avançados, ou para gerar gráficos coloridos e texturizados ou ainda para implementar técnicas que permitem escavar mais fundo em bases de dados.

Em apenas um dos sítios web dedicados ao R residem cerca de 1600 pacotes, número que tem vindo a crescer exponencialmente. Existe, por exemplo, um pacote chamado BiodiversityR, que possui uma interface gráfica destinada a facilitar a análise estatística de biodiversidade e ecologia de comunidades. Outro pacote, Emu, analisa padrões de fala, enquanto o pacote GenABEL permite realizar estudos de associação entre características tais como doenças e variações das bases de ADN ao longo do genoma. A comunidade de análise financeira tem uma particular afinidade com o R, contando com dezenas de pacotes especializados (RQuantLib, Rmetrics, portfolio, portfolioSim, tradeCost, backtest, PerformanceAnalytics, etc.)[3].

"A grande beleza do R tem a ver com o facto de o podermos modificar para fazer todo o género de coisas", afirma Hal Varian, economista chefe na Google. "E existe já uma quantidade imensa de material pré-empacotado, de maneira que nos sentimos de pé sobre os ombros de gigantes."

O R surgiu em 1996, quando os professores de estatística Ross Ihaka e Robert Gentleman da Universidade de Auckland na Nova Zelândia, distribuíram o código como um programa de utilização livre. Segundo contam aqueles professores, a ideia de desenvolver alguma coisa do tipo do R surgiu durante uma conversa de corredor. Ambos gostariam de ter uma tecnologia mais adequada para os seus estudantes de estatística, que precisavam de analisar dados e produzir informação em formato gráfico. O software mais próximo do que pretendiam tinha sido projectado por especialistas em computação e era de utilização muito difícil.

Não tendo uma formação específica em tecnologias de programação, aqueles professores consideraram o esforço que fizeram para produzir o código do R, mais um jogo académico do que qualquer outra coisa. Apesar disso, mais ou menos a partir de 1991, trabalharam a tempo inteiro para o R. "Fomos praticamente inseparáveis durante cinco ou seis anos," disse o Prof. Gentleman. "Enquanto um de nós digitava, o outro pensava, e vice-versa".

Alguns estatísticos que deram uma olhadela às versões iniciais do software consideraram-no pouco aperfeiçoado. Mas, apesar das suas deficiências, o R imediatamente conquistou seguidores nas pessoas que viram as possibilidades de personalização daquele software livre.

John M. Chambers, um antigo investigador nos Bell Labs, que agora é professor consultor de estatística na Universidade de Stanford, foi um precursor nesta área. De facto, enquanto trabalhou para os Bell Labs o Prof. Chambers colaborou no desenvolvimento do S, outro projecto de software estatístico, que foi pensado para proporcionar aos investigadores de todas as áreas uma ferramenta acessível de análise de dados. No entanto, o S nunca foi um projecto de código aberto. O software acabou por não conseguir gerar grande interesse e, finalmente, os direitos de S acabaram nas mãos da Tibco Software. Agora o R está a ultrapassar tudo o que o Prof. Chambers tinha imaginado possível com o S. "A diversidade e a excitação à volta do que todas essas pessoas estão a fazer é fantástica," afirmou.

Apesar de ser difícil determinar exactamente qual o número de utilizadores do R, alguns especialistas mais familiarizados com o software estimam que cerca de 250.000 pessoas o utilizem regularmente. Será a popularidade do R nas universidades uma ameaça para o SAS Institute, a empresa privada

[3] A título de curiosidade, realizou-se recentemente uma conferência dedicada a utilizadores do R nesta área: R/Finance 2009: Applied Finance with R April 24 & 25, Chicago, IL, USA.

especializada em software de análise de dados? O SAS, com mais de 2 mil milhões de dólares em receitas anuais, tem sido durante muitos anos o instrumento preferido de académicos e gestores de empresas.

"O R tornou-se realmente a segunda língua para as pessoas que acabam a pós-graduação agora, e há uma incrível quantidade de código que está a ser escrito para ele", disse Max Kuhn, director adjunto de estatística não-clínica da Pfizer. "Podemos olhar para os sítios de mensagens do SAS e descobrir uma diminuição proporcional do número de mensagens."

Do lado do SAS dizem que têm observado uma crescente popularidade do R nas universidades, apesar dos descontos educacionais oferecidos pelo seu próprio software, mas qualificam a nova tecnologia como sendo de interesse para um conjunto limitado de pessoas que trabalham em tarefas muito difíceis. "Acho que se dirige a um nicho de mercado para analistas que usam métodos de alta tecnologia e procuram código gratuito e disponível imediatamente", disse Anne H. Milley, directora de marketing de produtos de tecnologia da SAS. E acrescenta, "Nós temos clientes que constroem motores para a aviação. Quando entro num avião, fico satisfeita por saber que eles não andam a utilizar software que é distribuído gratuitamente."

Mas enquanto a SAS desvaloriza o interesse que o R pode ter para o mundo empresarial, empresas como a Google e a Pfizer dizem que usam este software para praticamente tudo o que é possível. Na Google, por exemplo, usam o R para ajudar a entender as tendências dos preços dos anúncios e para procurar padrões interessantes nos dados relativos às pesquisas efectuadas no seu motor de busca. A Pfizer por seu lado criou pacotes personalizados em R que os seus cientistas usam para explorar os seus próprios dados durante os ensaios não-clínicos de medicamentos em vez de enviar as informações para um consultor estatístico externo.

Os autores do R encaram de forma positiva o facto dessas companhias lucrarem com o fruto do seu trabalho bem como do de centenas de voluntários. O Professor Ihaka continua a ensinar estatística na Universidade de Auckland e gostaria de criar um software mais avançado. O Professor Gentleman aplica um software baseado no R, o Bioconductor, no trabalho de biologia computacional em que se encontra actualmente envolvido, no Fred Hutchinson Cancer Research Center em Seattle.

"O R é uma demonstração real do poder da colaboração, e eu penso que não seria possível construir uma coisa assim de outra maneira," diz o Prof. Ihaka. "Poderíamos ter optado por criar um software comercial, e teríamos vendido cinco cópias desse software."

A Estatística na Gestão do Politécnico

Manuela Figueira Neves, *mfigueira@ipg.pt*
Escola Superior de Tecnologia e Gestão
Instituto Politécnico da Guarda
CEAUL e UDI/IPG

Fernando Neves Santos, *fneves@ipg.pt*
Instituto Politécnico da Guarda. UDI/IPG

1. Introdução

O curso de Gestão foi um dos primeiros a funcionar nos estabelecimentos de ensino superior politécnico estando, actualmente, vulgarizado tanto neste sistema de ensino como no sistema de ensino universitário.

Com a implementação do processo de Bolonha assistiu-se à redução da duração dos cursos com a consequente redução do número de disciplinas do plano de estudos e naturalmente dos conteúdos programáticos leccionados.

Por ser a nossa área de formação base e enquanto docente de disciplinas de estatística do curso de gestão de uma instituição de ensino superior politécnico consideramos relevante analisar as alterações ocorridas com a introdução do processo de Bolonha ao nível do plano de estudos daquele curso e mais especificamente nas disciplinas da área de estatística.

2. O Processo de Bolonha

O processo de Bolonha é um plano de reforma, iniciado em 1999, que visa uniformizar a estrutura do ensino superior nos 45 países europeus participantes. O objectivo é a criação do espaço europeu de ensino superior até 2010 através da criação de um sistema de graus comparável e compreensível por todos, que facilite o reconhecimento dos graus e das qualificações no espaço europeu e a mobilidade dos alunos e diplomados entre instituições, promovendo desta forma a sua empregabilidade.

A transformação mais visível trazida pelo processo de Bolonha é a organização do ensino superior em três ciclos:

- o 1.º ciclo, com duração de três anos, conducente ao grau de licenciado;
- o 2.º ciclo, com duração de dois anos, conducente ao grau de mestre;
- o 3.º ciclo, com duração de três anos, conducente ao grau de doutor.

No caso do ensino politécnico, as alterações levam a que o grau de bacharel deixe de existir, passando a licenciatura com a duração de 6 semestres, equivalentes a 180 créditos e que o mestrado, ou seja, o 2º ciclo de formação, possa ser leccionado tendo a duração de 2 anos equivalentes a 120 créditos.

ECTS é a sigla correspondente a "*European Credit Transfer System*". Este sistema visa harmonizar, no espaço europeu, os critérios para a atribuição de unidades de crédito às unidades curriculares. O ECTS serve para medir, em unidades de crédito, a quantidade de trabalho que o aluno deve desenvolver em cada unidade curricular. São incluídas não só as horas de aulas colectivas mas também as horas que o aluno utiliza em sessões de ensino individualizado ou integrado em pequenos grupos,

em estágios, em projectos, em trabalhos de campo, em sessões autónomas de estudo e nos próprios momentos de avaliação. Um ano lectivo corresponde a um volume de trabalho equivalente a 60 unidades de crédito. Os créditos ECTS não medem ensino, medem aprendizagem. Indicam, em relação a cada unidade curricular, o volume de trabalho a efectuar pelo estudante para o completar com aproveitamento, incluindo todas as actividades de estudo ou procura e tratamento de informação.

Este sistema facilita o reconhecimento dos graus e das competências adquiridas na formação dentro do espaço europeu e possibilita aos diplomados uma grande mobilidade após a conclusão de cada ciclo de estudos, quer entre instituições nacionais, quer entre estabelecimentos nacionais e estrangeiros.

No processo de adequação dos cursos procedeu-se à redução das cargas lectivas presenciais, as quais são substituídas por outras modalidades formativas.

A mudança na duração dos cursos é a face mais visível da implementação do processo de Bolonha. Mas para além disso, Bolonha representa uma mudança de paradigma de ensino, preconizando a passagem de um modelo passivo, assente na transmissão de conhecimentos, para um modelo baseado no desenvolvimento das competências dos alunos. O aluno é o centro do processo de aprendizagem e implica, em muitos casos, a adopção de novas metodologias, colocando maior ênfase no trabalho do aluno e assentando numa aprendizagem mais activa, autónoma e prática, baseada no trabalho laboratorial e de campo, no estudo de casos, no desenvolvimento de projectos e na aprendizagem à distância.

Outro aspecto da declaração de Bolonha é o da comparabilidade dos critérios e métodos da avaliação e garantia da qualidade, como forma de facilitar o reconhecimento de diplomas, o que implica alguma homogeneidade dos sistemas de avaliação. Pouco antes da reunião de Bolonha, a recomendação do Conselho Europeu de 1998 (98/561/EC) sobre a qualidade do ensino superior pretendeu a introdução, à escala europeia, de métodos de garantia da qualidade e a promoção da cooperação europeia neste domínio, recomendando a todos os estados membros que estabeleçam sistemas de avaliação e de garantia de qualidade baseados em princípios comuns, já postos em prática na maioria dos processos de avaliação.

A declaração de Bolonha exprime a vontade dos governos signatários em edificar o "espaço europeu do ensino superior", o que levará:

- À adopção de um sistema de graus de fácil leitura e comparação dos diplomas obtidos nos diversos estados;
- Ao estabelecimento de um sistema de créditos reconhecidos em toda a Europa, para promoção da mobilidade dos estudantes;
- À promoção da mobilidade de estudantes e professores, com manutenção das regalias da origem durante estadias no estrangeiro;
- À cooperação para a garantia da qualidade do ensino superior, com estabelecimento de critérios de qualidade e de métodos de avaliação comparáveis à escala europeia.

A escola a que pertencemos foi pioneira na adequação dos seus cursos ao tratado de Bolonha e na sua implementação e encontra-se actualmente numa fase de avaliação de resultados.

3. O Politécnico

O ensino politécnico foi programado no início dos anos 70 e tornou-se uma grande necessidade dada a explosão da procura do ensino superior, a seguir à revolução de Abril de 1974, à qual a universidade não conseguia dar resposta. Foi uma época de grande desenvolvimento do ensino superior para o que contribuiu a assimetria territorial da oferta, a carência de recursos humanos e reivindicações regionais para a instalação de novos estabelecimentos de ensino superior. Assistiu-se a uma grande diversidade de oferta em termos de instituições de ensino superior. A diversidade é um elemento enriquecedor de qualquer sistema organizacional e é particularmente relevante no caso do ensino superior, face à grande complexidade dos desafios que se lhe colocam e à multiplicidade de solicitações que lhe são colocadas, a começar pela heterogeneidade crescente da população que a procura. A diversidade aumenta o leque de escolhas, ajusta-se à evolução rápida das exigências do mercado do trabalho, cria condições para experiências inovadoras e estimula a procura de padrões de excelência próprios de cada instituição. Aumenta a competitividade institucional, a qualidade proporciona oportunidades variadas de formação ao longo da vida, hoje um aspecto crucial da política do ensino superior.

Os politécnicos devem ter uma cultura institucional própria, diferente da das universidades. Isto inclui uma visão própria da tipologia e metodologia dos seus cursos, da composição e formação do seu pessoal docente, dos modelos organizativos e da governação, da natureza da sua investigação científica.

É comum afirmar-se que para muitos candidatos ao ensino superior, o politécnico é uma segunda escolha. No entanto, esta atitude tem vindo a mudar pela afirmação da qualidade das suas ofertas e uma maior consciência pública do seu valor social. Uma das suas vantagens posicionais, de que deve tirar partido, é a melhor cobertura geográfica em comparação com as universidades. Entre as universidades públicas, só 3 se localizam no interior, enquanto os institutos politécnicos cobrem todos os distritos do interior do país.

O mercado de emprego precisa de jovens com sólida preparação científica, de banda larga, com novas competências transversais e atitudes activas, com capacidade de conceptualizar projectos e de promover a integração de equipas. É o que se pede hoje à formação universitária. Mas precisa também de pessoas com "saber fazer", com sentido prático e empreendedor, com enraizamento no tecido económico-social e com grande capacidade de adaptação permanente à evolução tecnológica. É aqui que o politécnico tem um papel importante.

Outra questão importante da formação no politécnico é a grande articulação com a actividade profissional, incluindo a formação em ambiente de trabalho. Deste modo, a organização curricular dos cursos deve permitir uma ligação fácil entre o ensino e a prática, nos dois sentidos, seja através do ensino tradicional ou das novas formas de ensino à distância e de aprendizagem ao longo da vida promovendo práticas estreitamente ligadas à realidade.

4. O curso de Gestão

Este curso tem como objectivos a formação de quadros técnicos superiores de elevada qualificação com conhecimentos científicos, técnicos e práticos capazes de dar resposta às necessidades de empresas e outras organizações em domínios diversos tais como administrativo, contabilístico, financeiro e comercial. A concretização destes objectivos baseia-se numa formação marcadamente prática apoiada numa forte ligação ao meio empresarial.

As competências científicas e técnicas fornecidas durante o curso habilitam os licenciados em gestão ao desempenho de diversas funções em domínios como a organização e gestão, planeamento, estratégia e ensino. Estes diplomados estão aptos ao exercício de funções em gestão de pessoal, banca e seguros, contabilidade e fiscalidade, gestão financeira, gestão comercial, marketing, controlo de qualidade, auditoria fiscal e financeira, gestão de aprovisionamento, organização da produção, controlo de gestão, administração pública, estudos de viabilidade económico-financeira, ensino, gestão de pequenas e médias empresas e acesso a técnico oficial de contas.

O exercício de uma profissão na área da actividade empresarial implicará uma formação que deverá cobrir a generalidade das realidades abordadas pelo sector empresarial, as quais têm consubstanciação em todas as áreas organizacionais. Pretende-se, então, que o profissional de gestão preste apoio nos mais variados sectores internos e externos das organizações, existindo, desta forma, uma certa polivalência que é necessária neste domínio. Assim, entende-se que o perfil do diplomado em gestão obedeça a diversas aptidões gerais, nomeadamente

- deter aptidão numérica e habilidades quantitativas, de forma a desenvolver raciocínios logicamente consistentes;
- desenvolver prática de auto-gestão efectiva em termos de tempo, planeamento e comportamento, dinâmica pessoal, iniciativa individual e empreendedorismo;
- conhecer e aplicar tecnologias de informação, sistemas de apoio à decisão, sistemas operativos e ferramentas informáticas de execução e desenvolvimento da actividade empresarial;
- implementar técnicas de gestão com ênfase em gestão estratégica, qualidade e desenvolvimento organizacional;
- identificar novas formas de organização e de prestação de serviços;
- identificar e interpretar as directrizes dos vários tipos de planeamento aplicáveis à gestão das organizações;
- demonstrar visão sistémica e interdisciplinar da actividade empresarial e

- utilizar formulações matemáticas e estatísticas na análise dos fenómenos económicos e empresariais.

A gestão entendida como ciência coloca a questão da interdisciplinaridade. O edifício teórico da gestão engloba, em si, contribuições de um amplo leque de disciplinas científicas, cujo âmbito vai desde a área da ciência exacta à da ciência natural (Teixeira *et al.* (2007)).

Do plano de estudos de qualquer curso de gestão fazem parte disciplinas pertencentes a diversas áreas científicas tais como contabilidade e finanças, gestão comercial e administrativa, gestão de produção e métodos, matemática, estatística, economia, direito e informática.

No campo da estatística espera-se que um licenciado em gestão possua conhecimentos que vão desde a estatística descritiva, inferência estatística até à econometria. Interessa-nos pois saber quantas disciplinas da área de estatística fazem parte do plano de estudos do curso, os conteúdos programáticos nelas incluídos, qual a sua carga horária e as unidades de créditos correspondentes.

Na escola a que pertencemos este curso inclui duas unidades curriculares, estatística e estatística aplicada, ambas no 2º ano e que versam todos os tópicos de referência enumerados acima.

É nosso propósito analisar o que se passa nas restantes instituições.

Dos 15 institutos politécnicos existentes actualmente no país todos englobam diversos cursos na área da gestão e com designações bastante diversas. Os cursos desta área são leccionados nas escolas superiores de gestão, escolas superiores de tecnologia e gestão ou institutos superiores de contabilidade e administração (casos de Lisboa e Coimbra). Neste estudo vamos limitar-nos a analisar e comparar apenas os cursos com a designação de gestão ou gestão de empresas. Em 4 politécnicos constata-se a existência de vários cursos na área em análise mas com designações diferentes. No politécnico de Castelo Branco encontramos os cursos de gestão hoteleira, gestão turística, contabilidade e gestão financeira entre outros. Em Setúbal são leccionados, entre outros, os cursos de gestão de recursos humanos, gestão da distribuição e logística e gestão de sistemas de informação. No politécnico de Cávado e Ave encontram-se os cursos de gestão bancária e seguros e gestão de actividades turísticas entre outros. Finalmente no politécnico do Porto encontra-se a licenciatura em ciências empresariais. Os cursos destas instituições não são objecto da nossa análise.

Na tabela 1 pode observar-se a caracterização dos diversos cursos em termos das disciplinas de estatística que incluem no seu plano de estudos.

Em nosso entender a análise comparativa deverá ser feita em termos do número de disciplina da área de estatística constantes do plano de estudos mas também quanto ao número de unidades de crédito dessas disciplinas.

Assim, quanto ao número de disciplinas apenas um curso, o de Bragança, inclui 3 disciplinas. Em Beja, Portalegre e Santarém o curso de gestão contempla apenas uma disciplina desta área científica, o que nos parece ser manifestamente insuficiente. Os casos restantes incluem duas disciplinas. Na maioria dos casos as disciplinas situam-se no 2º ano do plano de estudos do curso.

Em termos do número de créditos observa-se uma variação entre 5 (Portalegre) e 18 (Bragança), situando-se a média nos 9,5.

Causa-nos preocupação verificar que são muitos os cursos que incluem número de créditos abaixo da média. São os casos de Beja, Portalegre e Santarém (que incluem apenas uma unidade curricular de estatística) e Lisboa e Viseu, apesar de incluírem duas disciplinas daquela área. Pode ser entendido como um sinal de não valorização destas competências.

Como critério de comparação utiliza-se o número de unidades de crédito ECTS, que como foi exposto atrás, é o sistema utilizado para esse fim. Seria ainda interessante comparar os cursos em termos das horas lectivas das disciplinas em análise, no entanto, em alguns casos há alguma dificuldade de obtenção dessa informação através dos sites consultados.

5. A importância da Estatística

A importância da estatística pode ser vista através da sua utilização a vários níveis: ao nível do estado, das empresas, das organizações sociais e profissionais, do cidadão comum e ao nível científico. Dada grande importância da estatística, praticamente todos os governos possuem organismos oficiais destinados à realização de estudos estatísticos, sendo em Portugal o Instituto Nacional de Estatística que tem essas funções.

Tabela 1 – Caracterização do curso de gestão em termos das disciplinas de estatística

INSTITUTO POLITÉCNICO	Designação do curso	Disciplinas da área de Estatística	Ano/Semestre	ECTS	Total
Beja	Gestão	Probabilidades e Estatística	2º/2º	6	6
Bragança	Gestão	Estatística	2º/1º	6	18
		Complementos de Estatística	2º/2º	6	
		Métodos Econométricos	3º/1º	6	
Coimbra	Gestão de Empresas	Probabilidades e Estatística	1º/2º	4	8
		Estatística Inferencial	2º/2º	4	
Guarda	Gestão	Estatística	2º/1º	7	12
		Estatística Aplicada	2º/2º	5	
Leiria	Gestão	Estatística Aplicada à Gestão I	1º/1º	5	10
		Estatística Aplicada à Gestão II	1º/2º	5	
Lisboa	Gestão	Estatística I	2º/1º	4	8
		Estatística II	2º/2º	4	
Portalegre	Gestão	Estatística	2º/1º	5	5
Santarém	Gestão de Empresas	Probabilidades e Estatística	1º/2º	5,5	5,5
Tomar	Gestão de Empresas	Estatística I	1º/2º	6	12
		Estatística II	2º/1º	6	
Viana do Castelo	Gestão	Probabilidades e Estatística	1º/2º	5	11
		Inferência Estatística e Investigação Operacional	2º/2º	6	
Viseu	Gestão de Empresas	Fundamentos de Estatística	2º/1º	4	9
		Estatística Aplicada	2º/2º	5	

Como a estatística é um ramo da matemática aplicada, os seus métodos são rigorosos e precisos. Apesar da objectividade que a matemática confere aos métodos estatísticos, deve ter-se em conta que os seus resultados incorporam alguma subjectividade. Tal subjectividade resulta principalmente da qualidade das medidas e das observações, o que é particularmente crítico no caso das ciências sociais e humanas.

Podemos afirmar que a estatística tem um papel muito importante no desenvolvimento científico em geral. Especialmente nas últimas décadas, tem crescido a sua importância, não apenas como uma disciplina independente, mas também como ferramenta ao serviço de outras ciências. Sendo a estatística uma ciência multidisciplinar que abrange praticamente todas as áreas do conhecimento humano, é conhecida a sua aplicabilidade nas ciências naturais, na medicina, na agronomia, na economia, na gestão, e mais recentemente na ciência forense. Também no âmbito das ciências humanas e sociais a estatística constitui um suporte de cientificidade de tal modo que lhes tem proporcionado consideráveis desenvolvimentos e aumento de credibilidade pública devido à sua utilização.

A estatística é uma ciência que estando ao serviço das demais permite a geração de conhecimento em outras áreas através de métodos quantitativos. Possibilita o desenvolvimento do raciocínio lógico dos estudantes. Dessa forma, eles estarão mais preparados para compreender os conteúdos de outras unidades curriculares. As informações estatísticas são concisas, específicas, eficazes e, quando analisadas com a ajuda dos instrumentos/técnicas formais de análise estatística, fornecem ajuda imprescindível para as tomadas racionais de decisão. Neste sentido, a Estatística fornece ferramentas importantes para que as empresas/instituições possam definir melhor suas metas, avaliar sua performance, identificar seus pontos fracos e actuar na melhoria contínua de seus processos.

Em termos da gestão das organizações, os programas de qualidade adoptados dependem em grande parte de modelos estatísticos. A implantação de programas de qualidade com vista à eliminação de desperdícios, redução de defeitos, diminuição de inspecções periódicas e aumento de satisfação do consumidor final vão deixando de lado as técnicas de controle antigas, tais como a inspecção da qualidade no produto acabado, substituíndo-as pela prevenção, baseando-se no controle do processo produtivo com a utilização de instrumentos estatísticos.

A diversidade de actuação é um dos grandes atractivos da estatística, que pode promover a melhoria da eficiência e também a solução de vários problemas práticos importantes em quase todas as áreas do saber.

De maneira mais expressiva, nota-se uma grande expansão da estatística como instrumento básico em todas as ciências que desejam explicar a realidade, bem como proporcionar maior criatividade e inovação dentro de cada ciência em particular.

O uso de ferramentas estatísticas nas ciências forenses tem actualmente grande importância nos meios judiciais. Os cientistas forenses ao avaliarem e interpretarem as evidências que incluem elementos de incerteza necessitam cada vez mais do apoio da ciência estatística nos seus mais diversos ramos.

Deste modo, teoria, métodos e aplicações de probabilidade e estatística estão na base da avaliação da evidência científica.

Nas últimas décadas o uso da estatística para fins legais teve um grande e rápido desenvolvimento e que continua nos dias de hoje. Os estatísticos são chamados a tribunal como peritos ou consultores. Eles avaliam e quantificam a incerteza. Muito embora já se verifique algum acolhimento relativamente à participação de estatísticos no ambiente jurídico, há ainda muito a fazer por uma aproximação mais efectiva entre a estatística e o direito.

Também a administração pública assim como o sector privado necessitam da actualização de conhecimentos dos seus quadros em estatística. A ausência de valorização dos métodos quantitativos tem implicações nos raciocínios e lógica na compreensão de diversos conceitos. Os conhecimentos estatísticos permitem encontrar metodologias e instrumentos de avaliação de processos e ponderar os resultados alcançados com as decisões tomadas. Dotam os indivíduos de capacidade de avaliar expectativas dos projectos, investimentos ou resultados a esperar das suas acções.

Bibliografia

Brito, C. *et. al* (2003). *Mat 10 - 2ª Parte*, *Lisboa Editora*.

Teixeira, A., Rosa, A. e António, N. (2007) "O doce amanhecer da ciência da GESTÃO: uma perspectiva filosófica". *Edições pedago*.

http://www.ipb.pt/; http://www.ipbeja.pt/; http://www.ipc.pt/; http://www.ipca.pt/; http://www.ipcb.pt/ http://www.ipg.pt/; http://www.ipl.pt/; http://www.ipleiria.pt/; http://www.ipp.pt/
http://www.ipportalegre.pt/; http://www.ips.pt/; http://www.ipsantarem.pt/; http://www.ipt.pt/
http://www.ipv.pt/; http://www.ipvc.pt/

Aplicação da teoria de valores extremos em Ciências Médicas: análise dos níveis elevados de colesterol total observado em Portugal

P. de Zea Bermudez, *pcbermudez@fc.ul.pt*[(1)] e Zilda Mendes, *zilda.mendes@anf.pt* [(2)]

(1) DEIO/FCUL e CEAUL; (2) CEFAR, INFOSAÚDE;

1. Introdução

Segundo o Portal da Saúde (http://www.min-saude.pt), as doenças cardiovasculares são de um modo geral definidas como "doenças que afectam o aparelho cardiovascular, designadamente o coração e os vasos sanguíneos". As doenças cardiovasculares têm um grande impacto na morbilidade e mortalidade da população Portuguesa, constituindo uma das principais causas de morte em Portugal. Efectivamente, de acordo com informações veiculadas pelo Portal da Saúde, as doenças cardiovasculares são responsáveis por cerca de 40% dos óbitos que ocorrem no nosso país. Infelizmente, o número exacto de indivíduos em risco de contrair alguma doença cardiovascular não é conhecido com precisão. Há diversos factores que contribuem para o risco de ocorrência destas doenças, tais como a idade e a história familiar de doença cardiovascular prematura (factores não modificáveis). Porém, há também factores de risco modificáveis, tais como sejam a obesidade, a falta de actividade física regular e o tabagismo que também contribuem para aumentar a probabilidade dos indivíduos virem a sofrer destas doenças. O controlo destes factores modificáveis pode melhorar consideravelmente a qualidade de vida da população e diminuir a incidência deste tipo de patologias.

Um dos factores de risco que é habitualmente observável nos indivíduos que desenvolvem alguma doença cardiovascular, tais como o enfarte de miocárdio, é a hipercolesterolemia, que consiste na presença no sangue de valores de colesterol superiores aos valores máximos recomendados. Actualmente, o valor de referência para o colesterol total é de 190 mg/dl (Task Force (2003) e Task Force (2007)).

No ano de 2005, a Associação Nacional de Farmácias (ANF) realizou uma campanha com o objectivo de identificar os indivíduos em risco de desenvolver uma doença cardiovascular em Portugal. O estudo pretendia fundamentalmente avaliar as características dos indivíduos que apresentavam valores anormais de certos parâmetros fisiológicos e bioquímicos. Como parte do protocolo da recolha de dados, os indivíduos que apresentavam valores demasiado elevados em alguns dos parâmetros foram referenciados a uma consulta médica, de acordo com o CheckSaúde (2005).

Nesta fase, o estudo que é aqui sucintamente apresentado pretende analisar os valores elevados de colesterol, nomeadamente verificar se existem diferenças regionais relativamente a esta variável.

2. **Descrição dos dados analisados**

A recolha da informação decorreu de 14 a 28 de Novembro de 2005 nas farmácias associadas da ANF que aceitaram participar na Campanha. Estas farmácias estão distribuídas ao longo dos 18 distritos do continente e dos arquipélagos da Madeira e dos Açores. Foram registadas as seguintes variáveis:

Sexo (F/M), Idade (anos), Distrito, Tabagismo (S/N), Índice de Massa Corporal (IMC, kg/m^2), Pressão arterial (mm/Hg), Glicemia em jejum e ocasional (mg/dl), Triglicerídeos (mg/dl), Colesterol Total (mg/dl) e Referência à consulta médica.

Nesta campanha foram recolhidas estas informações relativamente a 40065 utentes. O perfil dos indivíduos está representado no diagrama que é apresentado na Figura 1.

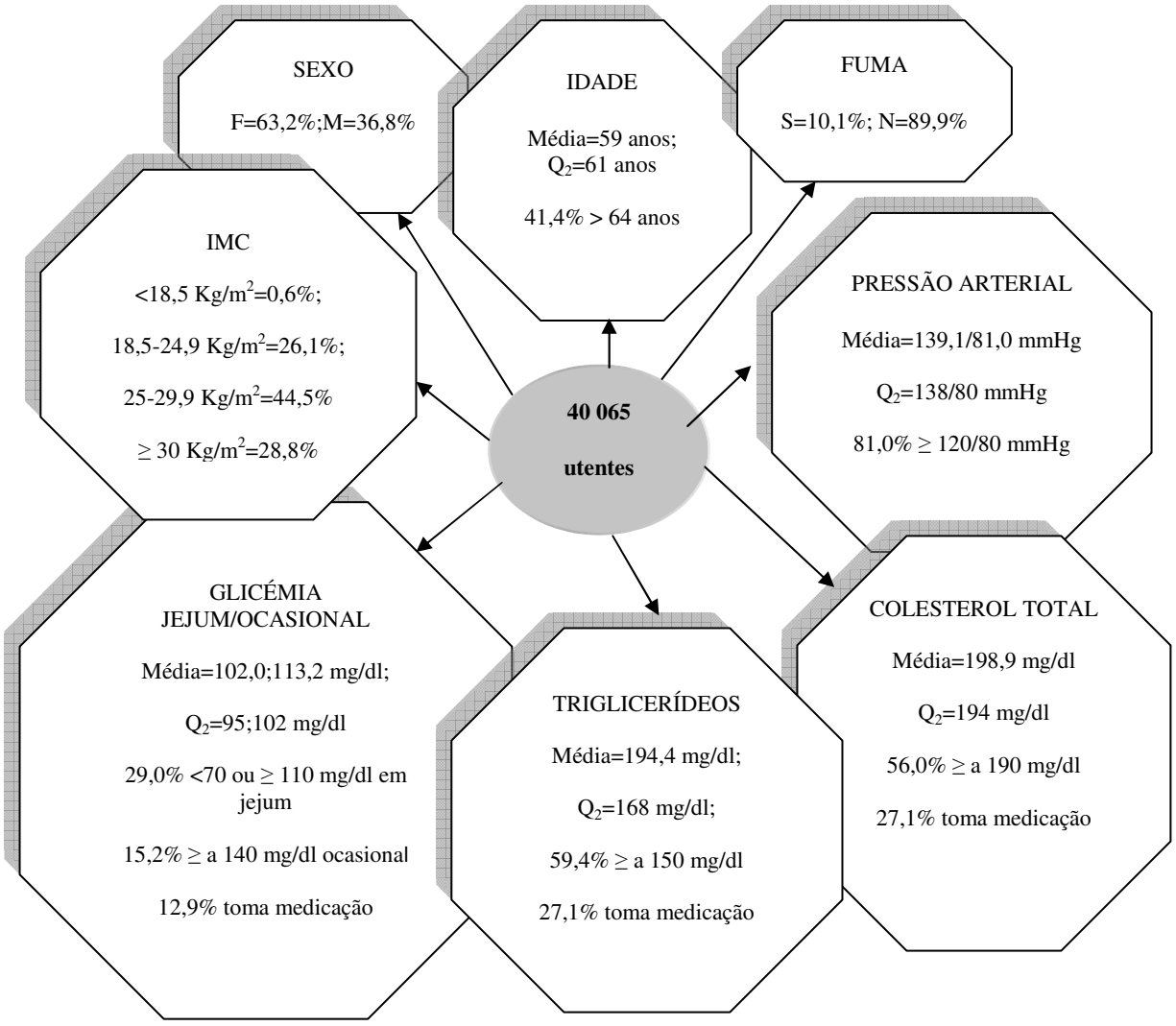


Figura 1 - Perfil dos 40065 indivíduos que constituem a amostra

É necessário ter em conta que se trata de uma amostra de indivíduos que se disponibilizaram a participar no estudo (voluntários), podendo não reflectir a realidade do país, nomeadamente no que se refere à idade e ao género. Também é preciso ter em conta que a distribuição das respostas por distrito não é de forma alguma equilibrada como se pode ver na Tabela 1. De notar também que, como é habitual nestes estudos, há uma considerável percentagem de valores omissos.

AVEIRO	BEJA	BRAGA	BRAGANÇA	C. BRANCO
3339	586	3057	1005	634
8,4%	1,5%	7,7%	2,5%	1,6%
COIMBRA	**ÉVORA**	**FARO**	**GUARDA**	**MADEIRA**
1493	749	841	427	1434
3,8%	1,9%	2,1%	1,1%	3,6%
AÇORES	**LEIRIA**	**LISBOA**	**PORTALEGRE**	**PORTO**
413	1718	9593	366	6519
1,0%	4,3%	24,3%	0,9%	16,5%
SANTARÉM	**SETÚBAL**	**V. CASTELO**	**VILA REAL**	**VISEU**
1722	3156	376	1027	1094
4,4%	8,0%	1,0%	2,6%	2,8%

Tabela 1 - Distribuição dos respondentes por distrito

Alguns gráficos de interesse relativamente à distribuição dos valores de colesterol segundo algumas das variáveis registadas são apresentados na Figura 2.

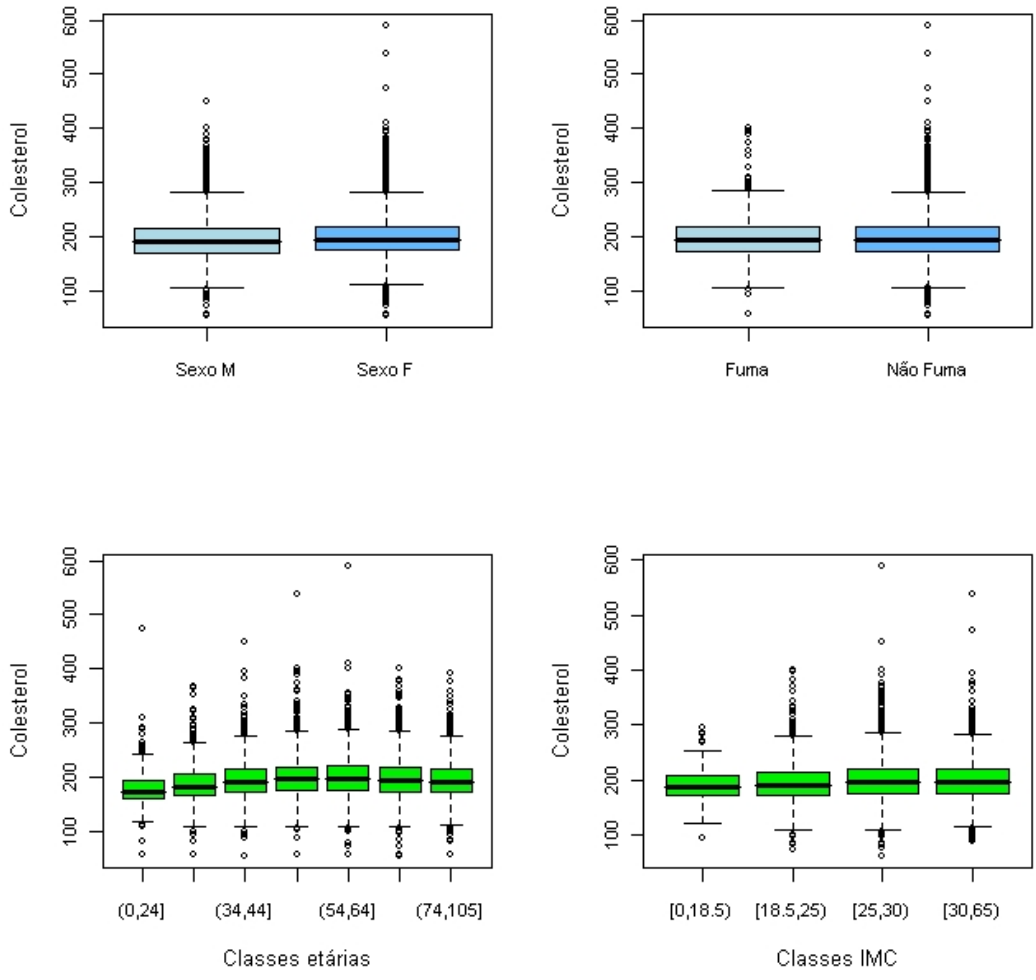


Figura 2 – Distribuição dos valores de colesterol total segundo o sexo, o tabagismo, a idade e os valores de IMC

Variável		Prob. Colesterol ≥ 190 mg/dl	Variável		Prob. Colesterol ≥ 190 mg/dl
Idade	**<25**	0,314	IMC	**<18,5**	0,472
	25-34	0,421		**[18,5; 24,9]**	0,522
	35-44	0,529		**[25,0; 29,9]**	0,573
	45-54	0,598		**≥30**	0,585
	55-64	0,600	**Fuma**	**Sim**	0,582
	65-74	0,566		**Não**	0,559
	≥75	0,536	**Sexo**	**Masculino**	0,528
				Feminino	0,579

Tabela 2 – Estimativa da probabilidade de que os valores de colesterol total não sejam inferiores ao valor de referência para algumas variáveis

Os valores que constam da Tabela 2 mostram que, para todas as variáveis (exceptuando os valores correspondentes aos indivíduos mais jovens), a estimativa da probabilidade de que os valores não sejam inferiores a 190 mg/dl é muito próxima de 0,50 ou até mesmo superior. Considerando que 190 mg/dl é o valor de referência do colesterol total, esta constatação pode ser preocupante. Obviamente, esta probabilidade pode estar de certo modo inflacionada pelo facto da amostra ser composta por indivíduos voluntários alguns dos quais, possivelmente, já têm problemas do foro cardiovascular. Apesar desta observação, tais probabilidades podem querer transmitir o facto problemático de que a hipercolesterolemia já constitui um sério problema de Saúde Pública na população Portuguesa.

3. Breve introdução metodológica

Dado que esta campanha foi desenvolvida com o intuito de identificar os indivíduos que estão em risco de desenvolver uma doença do foro cardiovascular, torna-se natural estudar os valores elevados de colesterol total. Nesta estrutura a utilização da teoria de valores extremos (TVE) torna-se fundamental. Também é de interesse avaliar se haverá diferenças a nível regional no que se refere aos valores elevados desta variável. As aplicações da TVE são comuns em diversas áreas, tais como a hidrologia e o ambiente, mas raras, ou mesmo inexistentes, nas ciências médicas. Neste trabalho, os valores elevados de colesterol foram modelados usando o método *Peaks Over Threshold*, usualmente conhecido pela sigla POT. Este método consiste em ajustar os excessos (ou excedências) acima de um limiar suficientemente elevado pela distribuição de Pareto Generalizada, GPD (ver Pickands (1975)). A função de distribuição da GDP (k,σ), $-\infty < k < \infty$ e $\sigma > 0$, é dada por:

$$F\left(x|k,\sigma\right)=\begin{cases}1-\left(1+k\,\dfrac{x}{\sigma}\right)^{-1/k}, k\neq 0\\ 1-\exp\left(-\dfrac{x}{\sigma}\right), k=0\end{cases},$$

sendo k e σ os parâmetros de forma e de escala, respectivamente. Tem-se que $x \geq 0$ se $k \geq 0$, enquanto que $0 \leq x \leq -\sigma / k$ se $k < 0$. A escolha de um *threshold (u)* adequado acima do qual é admissível modelar os dados com uma GPD pode ser um problema complicado. Habitualmente usam-se simultaneamente diversas técnicas, tais como a representação gráfica da função de excesso médio. Também se recorre frequentemente à representação gráfica de algum estimador (por exemplo, o estimador de máxima verosimilhança) do parâmetro de forma em função de u (ou do número de estatísticas ordinais superiores). Todos estes assuntos são extensivamente abordados em Embrechts et

al. (1997). Para uma visão da TVE com uma índole um pouco mais prática ver, por exemplo, Coles (2001).

4. Apresentação de alguns resultados

Os modelos ajustados aos dados de alguns dos distritos de Portugal são apresentados na Tabela 3.

Distrito	u	Est k	r	%	IC para k (95%)	Est σ	IC para σ (95%)
Aveiro	250	0,01	213	7,8	(-0,13; 0,14)	26,3	(21.3; 31,3)
Braga	270	0,42	71	2,9	(0,13; 0,70)	13,2	(8,5; 17,8)
Bragança	240	0,16	57	7,5	(-0,09; 0,41)	18,0	(11,5; 24,4)
Coimbra	270	0,18	53	4,6	(-0,11; 0,47)	28,1	(17,0; 39,2)
Faro	260	0,05	51	7,4	(-0,24; 0,34)	27,9	(16,8; 39,0)
Guarda	220	-0,21	56	19,1	(-0,48; 0,06)	29,2	(18,3; 40,1)
Leiria	250	-0,34	101	7,2	(-0,47 -0,21)	24,9	(19,3; 30,5)
Madeira	250	-0,27	76	7,6	(-0,41; -0,14)	26,0	(19,5; 32,8)

Tabela 3 – Modelos GPD ajustados a oito distritos de Portugal

Para os modelos ajustados aos oito distritos (Tabela 3), pode observar-se que:

- Os intervalos de confiança (IC) para o parâmetro de forma da GPD ajustada aos dados dos distritos de Aveiro, Coimbra e Faro francamente incluem o zero. Sendo assim, os modelos poderão, possivelmente, ser simplificados de forma a serem exponenciais. Os IC para o parâmetro *k* relativos aos distritos de Bragança e Guarda também incluem o zero. Porém este valor está muito próximo de um dos limites dos IC.
- Os modelos GPD ajustados aos dados dos distritos de Leiria e à Região Autónoma da Madeira reflectem uma distribuição subjacente de cauda leve, com limite superior de suporte finito.
- Os dados relativos ao distrito de Braga são modelados por uma distribuição GPD que reflecte uma distribuição com uma cauda moderadamente pesada.

Região	$q_{0.995}$	Modelo $q_{0.995}$	IC para $q_{0.995}$	$q_{0.999}$	Modelo $q_{0.999}$	IC para $q_{0.999}$
Aveiro	327,60	323,06	(313,85; 335,96)	380,00	366,43	(348,18; 401,46)
Braga	294,69	303,96	(295,14; 318,76)	381,82	366,45	(334,91; 460,97)
Bragança	288,94	301,08	(285,31; 335,11)	334,18	352,17	(318,13; 482,78)
Coimbra	348,45	346,60	(326,96; 384,63)	399,07	424,89	(377,36; 598,46)
Faro	343,42	340,80	(321,74; 383,78)	397,26	394,70	(356,91; 480,00)
Guarda	295,16	294,84	(282,56; 330,36)	305,14	313,71	(296,92; 394,72)
Leiria	289,99	293,90	(292,40; 299,97)	300,58	306,56	(304,08; 317,39)
Madeira	295,00	300,03	(295,52; 309,14)	297,00	316,07	(308,15; 332,96)

Tabela 4 – Quantis empíricos e quantis estimados usando os modelos ajustados

Alguns quantis empíricos e os correspondentes quantis estimados pelos modelos, assim como os IC a 95%, são apresentados na Tabela 4. Os valores obtidos, claramente mostram a concordância entre os quantis empíricos e os quantis estimados a partir dos modelos ajustados.

5. Comentários e trabalho futuro

- Constatou-se que a abordagem POT foi muito útil para modelar os maiores valores de colesterol. Em todas as regiões analisadas o limiar a partir do qual se considera que um valor de colesterol é elevado foi bastante superior ao valor de referência de 190 mg/dl e diferente em cada região. Apesar da população em análise apresentar já alguma frequência de co-morbilidades associadas, tal facto poderá indicar que existe uma grande percentagem desta população com risco acrescido, uma vez que os *thresholds* encontrados são sempre superiores ao valor de referência.

- Os valores dos parâmetros de forma das GPDs ajustadas aos dados de cada região são bastante diferentes podendo indicar a existência de uma componente espacial. Sendo assim, num futuro próximo, será analisada a possibilidade de modelar os valores máximos de colesterol usando uma abordagem espacial.

Referências

CheckSaúde, Guia Prático, Risco Cardiovascular: Parâmetros e Intervenção Farmacêutica (2005). Associação Nacional das Farmácias, Lisboa, Portugal.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag: London.

Embrechts P., Klüppelberg C., Mikosch T. (1997). *Modeling Extremal Events,* Springer-Verlag: Berlin.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, 3, 119-131.

Third Joint **Task Force** of European Society of Cardiology and other societies on Cardiovascular Disease Prevention in Clinical Pratice (2003). European guidelines on cardiovascular disease prevention in clinical pratice. *European Heart Journal*, 24,1601-1610.

The **Task Force** for the management of Arterial Hypertension of the European Society of Hypertension and of the European Society of Cardiology (2007). Guidelines for the Management of Arterial Hypertension. *Journal of Hypertension*, 25, 1105–1187.

Agradecimentos

As autoras deste estudo agradecem ao Departamento de Serviços Farmacêuticos pela cedência da base de dados para esta análise.

Epidemiologia espacial - aplicações em Saúde Pública

Dias JG[1], Matos E[2], Bento MJ[3], Queirós L[1], Correia AM[1], Mendonça D[2]

[1] *Departamento de Saúde Pública da Adm. Reg. de Saúde do Norte, I.P., jdias@arsnorte.min-saude.pt*

[2] *Instituto de Ciências Biomédicas de Abel Salazar da Universidade do Porto, dvmendon@icbas.up.pt*

[3] *Instituto Português de Oncologia do Porto, mjbento@ipoporto.min-saude.pt*

1. Introdução

O estudo da distribuição geográfica/espacial da incidência de doenças ou de outros acontecimentos relacionados com a saúde é designada por Epidemiologia Espacial[1,2,3]. Actualmente, é uma área com crescente relevância e grande utilização por parte dos investigadores em estudos epidemiológicos e de saúde pública[2,3], constituindo uma ferramenta importante para o estudo do comportamento de doenças e para a construção de indicadores de saúde[4]. A identificação de padrões de ocorrência da doença é uma componente importante no contexto da saúde pública[5]. Estes padrões podem sugerir factores relacionados com a ocorrência de doenças e levar a uma melhor compreensão e controlo das mesmas[1,2,6]. A localização espacial dos eventos é um componente do padrão observado que fornece informações importantes sobre a distribuição da doença numa região[10].

O desenvolvimento desta área de investigação deve-se em grande parte ao desenvolvimento da computação, o qual possibilitou um maior e melhor acesso aos sistemas de informação geográfica, mas também aos avanços metodológicos na área da estatística, destacando-se os métodos Bayesianos[1,7,8].

A Epidemiologia Espacial pode ser dividida em três grandes áreas[1]:

- Mapeamento de doenças: tem como objectivo avaliar a variação espacial das taxas de incidência ou de mortalidade, de forma a identificar diferentes níveis de risco; prever epidemias; e levantar hipóteses de forma a orientar acções em saúde.

- Estudos ecológicos: têm como objectivo estudar a relação entre potenciais factores de risco e a incidência de doenças ou a mortalidade. Estes estudos baseiam-se, essencialmente, na construção de modelos com os quais se pretende explicar a variação das taxas através de indicadores sócio-económicos, ambientais, comportamentais, além da sua localização espacial.

- Detecção de *cluster*: o objectivo principal deste tipo de análise é estabelecer a significância de um risco mais elevado em determinadas áreas.

Os exemplos apresentados neste trabalho inserem-se no contexto do mapeamento de doenças e têm como objectivo avaliar a variação espacial de riscos relativos e de taxas de incidência de doenças oncológicas[2].

O mapeamento de doenças pode ser feito usando dados pontuais, quando temos a localização exacta do evento, ou usando dados agregados por área, quando temos o número de eventos e a população em risco para cada uma das áreas analisadas[1,2,9]. Neste trabalho foram utilizados dados agregados por área. Este tipo de dados são os mais comummente utilizados, no entanto, em populações de pequena dimensão e/ou valores pequenos de ocorrência de doença, podemos ter valores extremos na incidência devido às variações populacionais[1,10,11]. Estas variações são designadas por flutuações aleatórias e podem influenciar a interpretação dos mapas produzidos[1,10,11]. Para evitar viés de interpretação, é necessário controlar as flutuações aleatórias usando métodos estatísticos que permitam o mapeamento do verdadeiro risco da doença[1,10-12].

Assim, um dos desafios da Epidemiologia Espacial é o desenvolvimento de métodos para a estimação das taxas de forma a diminuir a variabilidade aleatória nas regiões com populações em risco de pequena dimensão, assim como a possibilidade de incorporar covariáveis nos modelos[1,13,14]. As áreas com populações pequenas e/ou incidências baixas podem apresentar grande variabilidade nas taxas, podendo-se observar valores muito altos (sobrestimação do risco) ou muito baixos (subestimação do risco) nas taxas, sem que, estas áreas se caracterizem de facto como áreas de alto ou baixo risco[10,11,15].

Os modelos Bayesianos Hierárquicos são actualmente o método de modelação mais utilizado para a correcção ou para suavização das estimativas. Esta metodologia tem um papel cada vez mais relevante em estudos epidemiológico[4,16].

Neste trabalho apresentam-se dois exemplos de aplicação de modelos espaciais na área da saúde pública: incidência de cancro da tiróide e de cancro do estômago na região Norte de Portugal.

2. Métodos

Existem diversas abordagens estatísticas para o mapeamento de dados agregados por área, neste trabalho usaram-se modelos Bayesianos Hierárquicos[1,10,17]. O modelo básico para descrever o número de eventos observados em cada área i é o modelo de Poisson. Os modelos de Poisson têm sido usados frequentemente na análise de dados epidemiológicos[1,6,11,17,18]. O uso deste tipo de modelos é apropriado para a análise de doenças raras, dado que estes estudos implicam taxas de incidência baixas e/ou áreas de pequena dimensão.

Assim, considerou-se que os valores observados y_i em cada área i com $i = 1, \dots, n$ seguem uma distribuição de Poisson com média $\mu_i = e_i\rho_i$:

$$y_i \sim Poisson(\mu_i) = Poisson(e_i\rho_i)$$

onde e_i representa o número de casos esperados em cada área (calculados com base na população em risco e numa taxa global de referência para a doença na região) e ρ_i representa o risco relativo em cada área i.

Para "remover" ou "controlar" o efeito de possíveis variáveis confundidoras, como por exemplo o sexo e a idade foi necessário recorrer à padronização[10,19]. Sejam i o índice da área, j o índice do sexo, k o índice do grupo etário, então y_{ijk} representa o número de eventos da classe j e k na área i, e N_{ijk} o

número de pessoas em risco por idade e sexo na área i. A taxa global para a região em estudo por idade e sexo é dada por $r_{jk} = \frac{\sum_i y_{ijk}}{\sum_i N_{ijk}}$. Logo, $e_{ijk} = r_{jk} \times N_{ijk}$ é o número de eventos esperados na área i em cada classe idade e sexo. Assim, o número esperado de eventos na área i, assumindo que o risco é constante em cada classe idade e sexo, é dado por $e_i = \sum_j \sum_k e_{ijk}$.

Com este tipo de padronização, a modelação focaliza-se no logaritmo do risco relativo da doença $\theta_i = log\ \rho_i$. Ao considerar-se θ_i como efeitos fixos, as suas estimativas de máxima verosimilhança são $\widehat{\theta_i} = log\left(\frac{y_i}{e_i}\right)$, isto é, as estimativas do risco relativo são as razões padronizadas de morbilidade ou mortalidade (*Standardized Morbidity* ou *Mortality Ratio* - SMR). Mas, como referido anteriormente, a SMR em áreas pequenas pode ser enviesada, uma vez que valores extremos da SMR são geralmente baseados num número pequeno de casos.

Optou-se, então por um alisamento das SMR incorporando no modelo efeitos aleatórios (θ_i). Assim, permite-se a sobredispersão no modelo de Poisson causada por factores de confundimento não observados[20,21,22]. O método mais usual para estimar o vector de "efeitos aleatórios" $\theta = (\theta_1, \dots, \theta_n)$ é utilizar uma abordagem bayesiana. Na prática, a formulação do modelo é dada por[10,23]:

$$y_i \sim Poisson(\mu_i) = Poisson(e_i\rho_i)$$
$$log\mu_i = loge_i + log\ \rho_i = loge_i + \theta_i$$

No caso do mapeamento de doenças uma escolha frequente é considerar-se $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$.

Logo, como referido anteriormente, ao considerarmos os efeitos aleatórios pode-se, parcialmente, ter em consideração covariáveis não estudadas/medidas, que podem induzir a dependência espacial entre os y_i, no entanto, não permite a dependência espacial explícita entre os y_i. Esta dependência espacial explícita pode, por exemplo, ser decorrente da menor variabilidade das taxas nas áreas urbanas com elevada densidade populacional, em oposição à baixa densidade populacional nas zonas rurais, ou então devido à etiologia da doença. Esta dependência espacial explícita pode ser incluída no modelo acrescentando-se um termo de efeitos aleatórios espacialmente estruturado[20,21].

Assim, o modelo extendido é:

$$log\mu_i = loge_i + \alpha + \varphi_i + \nu_i$$

onde θ_i foi substituído por $(\alpha + \varphi_i + \nu_i)$, α corresponde à média do logaritmo do risco relativo de todas as áreas, φ_i representa a componente espacialmente não estruturada com média zero referente ao risco relativo da área i comparada com o global da região e ν_i corresponde ao efeito aleatório espacialmente estruturado. Uma escolha usual para a distribuição *a priori* correspondente à componente espacialmente estruturada ν_i é considerar-se um modelo condicional autoregressivo Gaussiano (CAR) intrínseco[24]:

$$\nu_i | \nu_{j\neq i} \sim N\left(\frac{\sum_{j\neq i} w_{ij}\nu_i}{\sum_{j\neq i} w_{ij}}, \frac{\sigma_\nu^2}{\sum_{j\neq i} w_{ij}}\right)$$

onde w_{ij} são os elementos da matriz de vizinhanças (considerou-se, como usualmente que $w_{ij} = 1$, quando a área i partilhava uma fronteira com a área j e $w_{ij} = 0$ caso contrário) e σ_ν é o hiperparâmetro que controla a força da dependência espacial.

Assim, o modelo hierárquico considerado neste trabalho é dado por:

$$y_i \sim Poisson(\mu_i) = Poisson(e_i\rho_i) \qquad i = 1, \dots, 68$$
$$log\mu_i = loge_i + log\ \rho_i = loge_i + \alpha + \varphi_i + \nu_i$$

onde $\alpha \sim U(-\infty, +\infty)$, $\varphi_i \sim N(0, \sigma_\varphi^2)$ and $\nu_i \sim CAR(\sigma_\nu^2)$.

Atribuíram-se distribuições *a priori* Gama independentes Gama$(0{,}5; 0{,}0005)$ para as precisões correspondentes a cada um dos hiperparâmetros ($\tau_\varphi = \frac{1}{\sigma_\varphi^2}, \tau_\nu = \frac{1}{\sigma_\nu^2}$).

Foram usados os softwares R e WinBugs para ajustar o modelo, tendo-se utilizado duas cadeias MCMC paralelas, cada uma com 25000 amostras incluindo 20000 para o período de *burn-in*.

3. Casos de estudo

3.1 Cancro da tiróide na região Norte de Portugal: 2004-2005

A incidência de cancro da tiróide tem aumentado continuamente ao longo das últimas décadas nos países da Europa Ocidental. Portugal apresenta uma das taxas mais altas de cancro da tiróide. Este tipo de tumor é muito mais frequente nas mulheres do que nos homens[25]. O conhecimento de como a incidência deste tipo de cancro varia geograficamente pode dar indicações para uma melhor compreensão dos factores determinantes desta doença. O objectivo deste caso de estudo foi analisar a distribuição geográfica da incidência de cancro da tiróide na região Norte de Portugal ao nível do concelho de residência dos doentes.

O Registo Oncológico Regional do Norte (RORENO) de base populacional foi responsável pela recolha e validação dos dados sobre a incidência de cancro da tiróide. Em 2004 e 2005, registaram-se 950 novos casos de cancro da tiróide na região Norte, sendo 165 (17,4%) em homens e 785 (82,6%) em mulheres. A taxa de incidência média anual bruta foi 14,7/10^5 habitantes (sendo 5,3/10^5 homens e 23,5/10^5 mulheres).

A Figura 1 apresenta a distribuição da taxa média anual de incidência de cancro da tiróide na região, por concelho de residência no período em estudo. Verifica-se que os concelhos de Vinhais e Mesão Frio são os que apresentam as taxas de incidência mais elevadas. Ambos os concelhos têm uma população de dimensão reduzida e dispersa no caso de Vinhais.

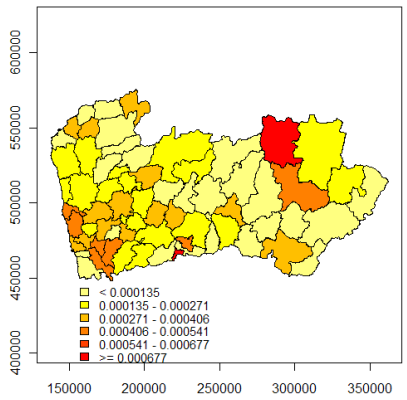


Figura 1 – Distribuição espacial da taxa média anual de incidência de cancro da tiróide na região Norte, por concelho de residência (2004-2005)

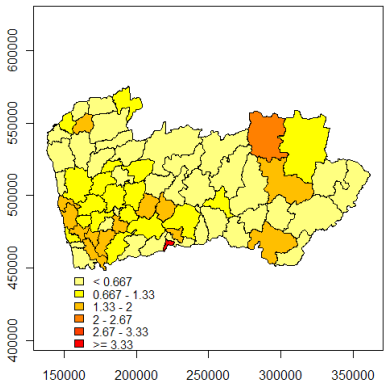


Figura 2 – Distribuição espacial da razão padronizada de morbilidade de cancro da tiróide na região Norte, por concelho de residência (2004-2005)

A Figura 2 apresenta o mapa da SMR por sexo e grupo etário, verificando-se que a sua distribuição espacial por concelho é diferente da distribuição da taxa média anual observada na Figura 1. Estas diferenças reflectem a variação na distribuição da população por idade e sexo nos concelhos da região Norte. No entanto, o concelho de Mesão Frio é o que mantém um risco mais elevado.

Na Figura 3 apresenta-se o mapa com os riscos relativos modelados. Como se pode verificar na figura os riscos relativos modelados aproximaram-se de 1 (média da região), reflectindo o alisamento efectuado pelo modelo e uma maior homogeneidade na distribuição do risco de doença. De realçar que os concelhos de Gondomar e Paços de Ferreira apresentam um risco igual ou superior a 1,47.

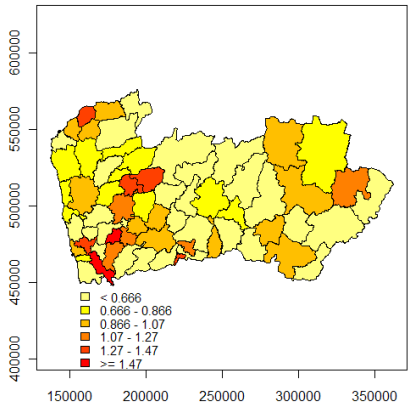


Figura 3 – Distribuição espacial dos riscos relativos estimados de cancro da tiróide na região Norte, por concelho de residência (2004-2005)

Este estudo revelou padrões espaciais nos dados e destacou as áreas de disparidade nas taxas de cancro da tiróide. É necessária pois a continuidade desta investigação para analisar este padrão de incidência relacionando-o com possíveis factores de risco como por exemplo, factores ambientais (concentração de rádon, radiação para o tratamento de doenças benignas) ou factores genéticos entre outros.

3.2 Cancro do estômago na região Norte de Portugal: 1995-2005

De acordo com dados mundiais recentes, o cancro do estômago é quarto cancro mais comum e o segundo como causa de morte por cancro[26]. A incidência de cancro do estômago apresenta variações geográficas e diferenças quer na incidência quer na mortalidade, tendo sido observado em vários países, que o norte apresenta um maior risco de mortalidade do que o sul. Este gradiente é particularmente acentuado no hemisfério Norte. O cancro do estômago é também um importante problema de saúde pública em Portugal.

Neste estudo os dados foram recolhidos a partir de registo oncológico regional -RORENO entre 1995 e 2005. No período em estudo registaram-se 10826 novos casos de cancro do estômago (4457 nas mulheres e 6369 nos homens) numa população de aproximadamente 3,24 milhões de habitantes (Censos 2001). Nos 68 concelhos da região as taxas brutas médias anuais foram de $41/10^5$ habitantes nos homens e de $27/10^5$ habitantes nas mulheres. As taxas específicas por grupo etário aumentaram progressivamente com a idade, variando de $4/10^5$ a $127/10^5$ nas mulheres e de $5/10^5$ a $238/10^5$ nos homens.

A Figura 4 apresenta o mapa da SMR por sexo e grupo etário, verificando-se que a distribuição destes valores por concelho não é homogénea na região. Os valores mais elevados de SMR estão concentrados em alguns concelhos dos distritos do Porto, Braga e Viana do Castelo.

Na Figura 5 apresenta-se o mapa com os riscos relativos modelados. Neste caso, apesar da forma da distribuição ser mais suavizada, confirmam-se as mesmas conclusões.

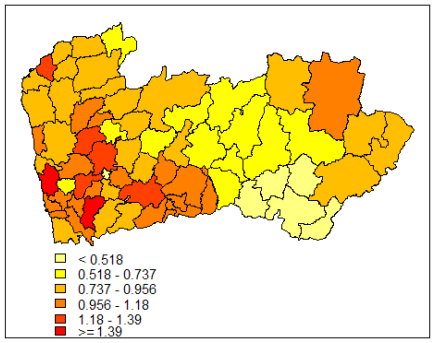


Figura 4 – Distribuição espacial da razão padronizada de morbilidade de cancro do estômago na região Norte, por concelho de residência (1995-2005)

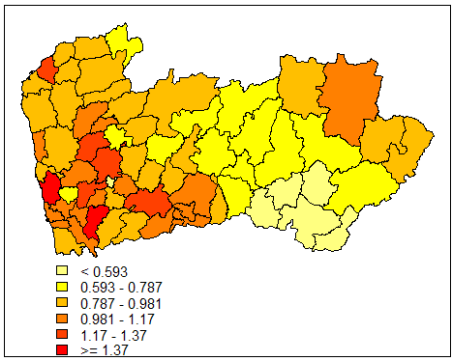


Figura 5 – Distribuição espacial dos riscos relativos estimados de cancro do estômago na região Norte, por concelho de residência (1995-2005)

A maior incidência de cancro do estômago observada em alguns concelhos da região justifica a necessidade de mais estudos para compreender os principais factores que contribuem para estes valores, nomeadamente, os hábitos alimentares/tradição, que podem justificar a manutenção de altos valores de incidência de cancro do estômago nestes concelhos.

4. Discussão

A análise bayesiana em conjunto com os sistemas de informação geográfica permitiram evidenciar áreas de disparidade no risco de cancro da tiróide e do estômago nos concelhos da região norte de Portugal.

Este estudo utiliza técnicas de modelação espacial e representa um passo inicial importante para compreender a dinâmica destas doenças. Alguns dos resultados encontrados neste trabalho justificam a necessidade de realizar estudos de investigação dirigidos à identificação dos factores que explicam a existência de diferentes padrões geográficos nas incidências de cancro na região Norte. Para tal deverão ser desenvolvidos modelos que permitam incorporar covariáveis como, por exemplo, factores socioeconómicos, comportamentais, de saúde ou ambientais que poderão ajudar a explicar as diferenças no padrão encontrado.

Neste estudo apenas foram considerados modelos envolvendo uma componente espacial, mas a evolução ao longo do tempo pode também ser de grande interesse. O próximo passo poderia ser descrever também a distribuição temporal, utilizando modelos espaço-temporais, para investigar estes padrões.

Os avanços estatísticos em epidemiologia espacial mais recentes permitem também incluir nos modelos, para além da dimensão do tempo, a combinação de informações a nível individual e a nível da área[27,28,29]. As limitações de usarmos apenas dados agregados são as que decorrem dos estudos ecológicos[27-29]. Embora este tipo de estudos seja útil para detectar associações entre níveis de exposição e ocorrência da doença, o uso de dados agregados tem limitações associadas, nomeadamente a falácia ecológica[27-29]. A falácia ecológica resulta de inferências causais individualizadas, tendo como base observações de um grupo de indivíduos agregadas por área. Portanto, esta recente metodologia visa incorporar os efeitos a nível individual, quer na exposição quer na doença, utilizando dados a nível de área, complementados com dados a nível individual, mesmo que estes sejam obtidos a partir de amostras pequenas[27-29]. A combinação de ambas as fontes de informação, área e individual,

permitirá corrigir o viés ecológico no efeito estimado[29], e assim dar um contributo para a implementação de estratégias de prevenção e controlo da doença oncológica.

Referências

1 Lawson A. *Statistical methods in spatial epidemiology*. Sussex: John Wiley & Sons, 2001.

2 Elliot P, Wakefield J, Best N, Briggs D. *Spatial Epidemiology: Methods and Applications*. London: Oxford University Press, 2000.

3 Elliot P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. *Environ Health Perspect*. 2004;112(9):998-1006.

4 Carvalho MS, Souza-Santos R. Análise de dados espaciais em saúde pública: métodos, problemas, perspectivas. *Cad. Saúde Pública*. 2005;21(2):361-378.

5 Crighton EJ, Elliott SJ, Moineddin R, Kanaroglou P, Upshur RE. An exploratory spatial analysis of pneumonia and influenza hospitalizations in Ontario by age and gender. *Epidemiol Infect*. 2007;135(2):253-61.

6 Bailey T. An Introduction to Spatial and Spatio-temporal Modelling of Small Area Disease Rates. Spring Course organizado pelos Instituto de Ciências Biomédicas de Abel Salazar, Faculdade de Medicina e Instituto de Saúde Pública da Universidade do Porto. 2008.

7 Congdon P. *Bayesian Statistical Modelling*. Sussex: Jonh Wiley & Sons, 2006

8 Paulino C, Turkman A, Murteira B. *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 2003.

9 Bailey T, Gatrell A. *Interactive Spatial Data Analysis*. England: Longman. 1995.

10 Bailey T. Spatial statistical methods in health. *Cad Saude Publica*. 2001;17(5):1083-1098.

11 Silva G. An Introduction to Spatiotemporal Model Analysis. Porto. Curso do Congresso Nacional de Epidemiologia. Porto. 2008.

12 Elliott P, Wartenberg D.Spatial epidemiology: current approaches and future challenges.
Environ Health Perspect. 2004;112(9):998-1006.

13 MacNab Y, Dean C. Spatio-temporal modeling of rates for the construction of disease maps. *Statist Med*. 2002;21:347-358.

14 Elliott P, Savitz D. Design issues in small area studies of environment and health. *Environ Health Perspect*. 2008;116(8):1098-1104.

15 Assunção RM. Estatística Espacial com Aplicações em Epidemiologia, Economia e Sociologia. São Paulo. Associação Brasileira de Estatística. 2001.

16 Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics, *Ann Inst Statist Math* 1991;43:1-59.

17 Carvalho M, Natário I. Análise de dados espaciais. Vila Real. *Sociedade Portuguesa de Estatística*. 2008.

18 Lowe R, Bailey T, Stephenson D, Graham J, Coelho C, Carvalho M, Barcellos C. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Computers and Geosciences 2010. Article in press.

19 Lawson A. Disease map reconstruction. Stat Med. 2001;20(14):2183-2204.

20 Clayton D, Bernardinelli L. *Bayesian methods for mapping disease risk*. Oxford. Oxford University Press. 1996.

21 Mollie A. Bayesian mapping of disease. London: Chapman & Hall.1995.

22 Lawless J. Negative binomial and mixed Poisson regression. Canadian Journal of Statistics. 1987;15(3):209–225.

23 Ricardson S, Thomson A, Best N, Elliot P. Interpreting posterior relative risk estimates in disease-mapping studies. Environmental Health Perspectives. 2004; 112:1016-1025.

24 Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems. *Statistical Science*. 1995;10(1):3–41.

25 Wartofsky L. Increasing world incidence of thyroid cancer: Increased detection or higher radiation exposure? Hormones 2010;9(2):103-108.

26 Brenner H, Rothenbacher D, Arndt V. Epidemiology of stomach cancer.*Methods Mol Biol*. 2009;472:467-77.

27 Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. *Stat Med*. 2006;25(12):2136-2159.

28 Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J R Statist Soc* A. 2008;171:159–178.

29 Jackson C, Richardson S, Best N. Studying place effects on health by synthesising individual and area-level outcomes. *Soc Sci Med*. 2008;67(12):1995-2006.

Pós - Doc

Tese de Doutoramento: Modelos Aditivos Generalizados em Análise de Sobrevivência (Boletim SPE Primavera de 2007, p.53)

Ana Luisa Papoila, *ana.papoila@fcm.unl.pt*

Departamento de Bioestatística e Informática
Faculdade de Ciências Médicas da Universidade Nova de Lisboa

Caros colegas e amigos,

Realizei as minhas provas públicas de Doutoramento na Reitoria da Universidade de Lisboa, no dia 19 de Julho de 2006. Foi um dia muito importante não só para mim mas também para todos os que me rodearam durante o período em que tive que me dedicar a tão importante empreendimento. Confesso que a elaboração deste projecto constituiu uma tarefa árdua mas que, simultaneamente, me muniu de conhecimentos teóricos e práticos que nunca pensei virem a ser de uma utilidade inquestionável. É comum associar à tese de doutoramento a ideia de que é algo muito teórico (por vezes até "aborrecido"!) e que temos que realizar independentemente de ser ou não interessante e/ou de ter ou não uma utilidade prática. Felizmente, no que me diz respeito, devo dizer que gostei do tema, estudei e aprendi imenso e os novos conhecimentos que adquiri constituíram um precioso complemento à minha anterior formação académica.

A tese incidiu sobre a modelação estatística de dados recorrendo aos Modelos Aditivos Generalizados (GAMs) e à Análise de Sobrevivência. Foi uma aposta no desenvolvimento de modelos com maior flexibilidade. De facto, a importância conferida aos dados aquando da sua modelação tem vindo a crescer; o papel que desempenham deixou de estar limitado ao ajustamento de determinado modelo teórico e passou a ser mais amplo com vista à obtenção de modelos que melhor reflectem a dependência das variáveis em estudo. Neste contexto, têm vindo a ser desenvolvidos, em várias áreas de investigação, modelos de regressão semi-paramétricos e não paramétricos possuidores de uma maior flexibilidade que advém não só do método de estimação da forma funcional que relaciona cada covariável com a variável resposta, mas também da estimação da função de ligação. Tais modelos constituem uma generalização dos Modelos Lineares Generalizados (GLMs) e dos GAMs. Neste projecto, foi explorada a ligação existente entre os GAMs e os modelos de Análise de Sobrevivência na medida em que existe uma correspondência entre estes modelos e os modelos de regressão de resposta binária. Assim sendo, também em Análise de Sobrevivência podemos encontrar modelos com o mesmo tipo de flexibilidade.

Considero estas áreas da Estatística apaixonantes e indispensáveis para quem, como eu, se dedica a analisar dados médicos. No entanto, os conhecimentos que há que ter para poder responder a todos os pedidos de colaboração provenientes das áreas biomédicas, estão bastante além do que podemos imaginar. Cada "caso é um caso" e, na realidade, cada conjunto de dados proveniente de estudos

médicos deveria ser analisado como se de um "doente" se tratasse; com isto quero dizer que, paralelamente ao mundo clínico em que o médico apenas diagnostica as doenças das quais conhece a sintomatologia, também nós, estatísticos, apenas conseguimos analisar correctamente um conjunto de dados se tivermos um vasto conhecimento dos métodos que nos permita fazê-lo. Quantos de nós já analisaram dados longitudinais considerando a média ou a mediana das várias medidas repetidas ou utilizaram modelos que não têm em consideração a estrutura de correlação existente entre as várias observações obtidas ao longo do tempo? E isto é só um exemplo! Como ultrapassar este problema? através de uma indispensável actualização contínua não só de conceitos mas também de novas metodologias estatísticas; só assim se consegue manter a qualidade do tratamento estatístico dos dados. Por este motivo, tenho tentado frequentar acções de formação em várias áreas da Estatística e não só. Vejamos, por exemplo, a Epidemiologia; definida como "o estudo da distribuição e factores determinantes das doenças e lesões nas populações humanas", permite-nos compreender se, e como, as diferenças entre indivíduos permitem explicar os vários padrões da distribuição da doença existentes na população. Ao realizarmos este tipo de estudos é importante tomar precauções desde o início, de forma a garantir a qualidade final dos dados recolhidos; na realidade, não queremos desperdiçar o tempo despendido e os custos, por vezes altos, devido a um deficiente delineamento do projecto. Não nos podemos esquecer que, por muito sofisticada que seja a metodologia estatística que utilizarmos para analisar um conjunto de dados obtidos através de um planeamento incorrecto e/ou sem uma monitorização de todo o processo de recolha, os resultados que obtivermos não nos darão qualquer garantia de qualidade. Há pois que reflectir, amadurecer ideias e escolher o desenho epidemiológico mais apropriado e que permitirá atingir os objectivos previstos. Existem ainda muitos outros aspectos de índole epidemiológica a considerar tais como a escolha das medidas de ocorrência da doença, a identificação de variáveis de confundimento, os tipos de viés com que teremos que lidar e qual a forma de os evitar, como estabelecer uma relação causal e não apenas uma associação estatística entre um factor de risco e uma doença ou condição, enfim! todas estas reflexões devem ser tidas em consideração aquando da elaboração do protocolo, primeiro passo a ser dado quando se pretende alcançar resultados com fiabilidade e credibilidade aceitáveis.

Concluindo, se já antes do Doutoramento existia alguma preocupação da minha parte em actualizar conhecimentos, presentemente, a formação continua a fazer parte da minha agenda e é levada muito a sério. De facto, sendo grande a diversidade de metodologias estatísticas a que tenho que recorrer para poder analisar os dados que me são facultados, só assim é possível ajudar com alguma qualidade quem me procura.

Aparte de colaborar com os médicos que acorrem ao Departamento de Bioestatística e Informática da Faculdade de Ciências Médicas (FCM) da Universidade Nova de Lisboa (UNL), ainda exerço as funções de docência inerentes ao cargo que ocupo na referida instituição. Contrariamente ao que desejaria, esta actividade representa uma carga adicional bastante pesada na medida em que, além das aulas da licenciatura, ainda tenho que ministrar aulas de cursos de actualização, mestrados e cursos de doutoramento, organizados pela FCM. Todos conhecemos as dificuldades financeiras que as faculdades do nosso país atravessam no momento pelo que, também nós sofremos com esta situação.

Recentemente, foi-me proposto o desafio de colaborar na criação de um Centro de Investigação (CI) no Centro Hospitalar de Lisboa Central (CHLC), no âmbito do protocolo estabelecido entre a FCM da UNL e o referido centro hospitalar. Como é óbvio, vi nesta proposta uma oportunidade única de ajudar os profissionais de saúde a delinear os seus estudos de uma forma correcta, com vista à obtenção de dados mais fiáveis e, se possível, com características que exijam análises mais motivadoras do ponto de vista da modelação estatística. A minha intervenção na organização do CI tem sido muito interessante e, sem dúvida, representa uma nova experiência profissional. A missão da equipa seleccionada para constituir o CI é a de promover uma investigação clínica de qualidade. Entre outros objectivos a atingir, salienta-se a prestação dos seguintes serviços: ajuda na pesquisa de financiamento,

apoio epidemiológico, estatístico e na gestão de dados e ainda assessorar o Conselho de Administração do CHLC através de pareceres a propostas de projectos de investigação internos e externos ao CHLC. Para isso, foram criadas três estruturas funcionais: o Gabinete de Apoio aos Projectos, o Gabinete de Análise Epidemiológica e Estatística e o Gabinete de Articulação. Cabe a este último a optimização de sinergias intra- e interinstitucionais.

Agora que o CI está preparado para receber os pedidos de colaboração, dado que se encontra terminada a fase de organização, tenho plena consciência de que não será uma tarefa simples. De facto, sendo o CHLC actualmente constituído por quatro pólos hospitalares: Hospital de S. José, Hospital de Santa Marta, Hospital de Santo António dos Capuchos e Hospital de Dona Estefânia, será grande o número de profissionais de saúde que irá solicitar os nossos serviços; o sucesso deste projecto passa inevitavelmente pela constituição de uma forte equipa de epidemiologistas e de estatísticos. Por outro lado, a análise que os dados requerem poderá vir a ser um pouco repetitiva e até sem interesse nalgumas situações. No entanto, as limitações inerentes a um trabalho deste tipo serão, de alguma forma, minimizadas, dada a riqueza de dados que seguramente irão surgir de tão grande manancial de informação. Além disso, este projecto representa a oportunidade de integrar o meio académico com a actividade clínica de um grande hospital que desponta para a investigação e que tem como objectivo fazê-lo de uma forma estruturada. Um projecto destes, a ter êxito como todos esperamos, será seguramente muito gratificante.

A partir das linhas que acabo de escrever, facilmente se conclui que pouco tempo resta para a investigação. Felizmente que alguns dos estudos médicos que tenho apoiado são, do ponto de vista da investigação clínica, de grande valor científico. Em relação à investigação na área da Estatística, esta é realizada com grande esforço e, frequentemente, com prejuízo das horas de lazer. Pena é que o dia só tenha 24 horas!

Termino reforçando a ideia de que, após as provas de doutoramento, surgem mais responsabilidades, aumenta não só o número mas também a diversidade de solicitações e a criação de prioridades passa a ser vital para a nossa "sobrevivência" profissional.

Saudações académicas
Ana Luisa Papoila

Tese de doutoramento: Uma Abordagem Bayesiana à Determinação de Modelos

(Boletim SPE Primavera de 2007, p. 54)

Júlia Teles, *jteles@fmh.utl.pt*

Secção de Métodos Matemáticos e Centro Interdisciplinar para o Estudo da Performance Humana, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa

A determinação de modelos, que engloba a selecção e a validação, é um problema fulcral na Estatística. A abordagem bayesiana deste problema, com especial ênfase para a selecção de variáveis em modelos de regressão, foi o objecto de estudo na minha dissertação de doutoramento. Neste texto faço um enquadramento do problema da selecção de modelos, descrevendo, de forma sucinta, algum do trabalho realizado durante e depois do doutoramento e aponto alguns possíveis caminhos para investigação futura. Embora não seja um assunto tratado neste texto, quero também relembrar a importância da validação de modelos, muitas vezes relegada para um segundo plano nas aplicações práticas, para o sucesso do procedimento de inferência, seja ele paramétrico ou preditivo.

Na literatura estatística encontra-se uma grande variedade de critérios para comparação de modelos. Os mais conhecidos e utilizados são os que se baseiam na minimização de uma quantidade, que pode ser expressa como menos duas vezes o máximo da log-verosimilhança mais uma função de penalização, que depende do número de parâmetros do modelo e/ou da dimensão da amostra. Entre eles destaca-se o critério de informação de Akaike (AIC) (Akaike, 1973) e o critério de informação bayesiano (BIC) (Schwarz, 1978). Apesar de algumas aproximações assintóticas que procuram justificar a utilização destes critérios na abordagem bayesiana (Kass e Raftery, 1995), estes não são intrinsecamente bayesianos, na medida em que não permitem incorporar a informação *a priori* eventualmente existente.

O factor Bayes (Jeffreys, 1961) foi considerado, durante muito tempo, como o único modo adequado para a comparação bayesiana de modelos. Um dos problemas inerentes à utilização do factor Bayes, na comparação de modelos, é o facto de ele poder não se encontrar definido quando se considera distribuições *a priori* impróprias, o que acontece frequentemente quando se utilizam distribuições não informativas. Para contornar este problema, foram propostas diversas modificações ao factor Bayes, tais como o factor Bayes *a posteriori* (Aitkin, 1991), parcial (O'Hagan, 1991), intrínseco (Berger e Pericchi, 1996), fraccionário (O'Hagan, 1995) e o pseudo-Bayes (Gelfand *et al.*, 1992).

Alternativamente a estas modificações do factor Bayes, Carlin e Louis (2000) propuseram a utilização de uma versão bayesiana do BIC. A aplicação deste critério leva à escolha do modelo que minimiza menos duas vezes o valor esperado *a posteriori* da log-verosimilhança mais uma função de

penalização, que depende da dimensão da amostra e do número de parâmetros do modelo. A estrutura tipicamente hierárquica da modelação bayesiana dificulta, contudo, a determinação do número de parâmetros do modelo e a utilização da versão bayesiana do BIC, sendo o critério de informação da *deviance* (DIC) (Spiegelhalter *et al.*, 2002) uma alternativa adequada. Este critério, apresentado como uma versão bayesiana do AIC, tem duas componentes: uma medida da qualidade do ajustamento – o valor esperado *a posteriori* da *deviance* bayesiana – mais uma medida da complexidade do modelo – a diferença entre o valor esperado *a posteriori* da *deviance* bayesiana menos a *deviance* bayesiana calculada no valor esperado *a posteriori* do vector de parâmetros. Note-se que, a *deviance* bayesiana, tal como é definida em Spiegelhalter *et al*. (2002), é menos duas vezes a log-verosimilhança mais duas vezes o logaritmo de uma função que depende apenas dos dados. Na comparação de modelos é usual considerar que esta função é igual a 1. Deste modo, a medida da qualidade de ajustamento do DIC é a mesma da versão bayesiana do BIC, havendo apenas diferença na medida da complexidade do modelo. Outros critérios têm vindo a ser propostos, *e.g.*, o critério de informação bayesiano preditivo (BPIC) (Ando, 2007) e critérios baseados em *scores* (Draper e Krnjajic, 2007). Diversos estudos de simulação para comparação destes e doutros critérios têm sido realizados, *e.g.*, Teles (2005), Teles e Amaral Turkman (2007) e Ando (2007).

Quando existem muitos modelos para comparar não é possível examinar exaustivamente todos eles e, deste modo, a arbitrariedade na escolha do modelo é devida, não só ao critério de selecção de modelos mas, também, ao método de pesquisa utilizado. No caso particular da selecção de variáveis em modelos de regressão, o número de modelos a comparar é extremamente elevado, mesmo com um número moderado de possíveis variáveis. Assim, a pesquisa exaustiva é uma solução que não pode ser considerada na maior parte das situações, sendo necessário recorrer a métodos alternativos.

A abordagem bayesiana à selecção de variáveis em modelos de regressão engloba diversos tipos de métodos. Alguns deles, inspirados nos usados na abordagem clássica, combinam métodos de pesquisa com critérios de avaliação de modelos, usualmente um critério de verosimilhança penalizada ou uma função de discrepância (Carlin e Louis, 2000). O método informal de pesquisa manual (Paulino *et al*., 2003), e o método de selecção progressiva ou de eliminação regressiva via DIC (Teles e Amaral Turkman, 2004; Teles, 2005), são exemplos deste tipo de métodos de selecção de variáveis.

Contudo, a solução intrinsecamente bayesiana para o problema de selecção de variáveis é, na maior parte das vezes, encarada como um problema de estimação, nomeadamente de estimação das probabilidades *a posteriori* dos diversos modelos e das probabilidades *a posteriori* de inclusão de cada uma das variáveis no modelo. A estimação destas probabilidades permite identificar o modelo de probabilidade *a posteriori* máxima (Ntzoufras, 2009) e, caso exista, do modelo de probabilidade *a posteriori* mediana (Barbieri e Berger, 2003), *i.e.*, o modelo que inclui as variáveis que têm probabilidade *a posteriori* de inclusão no modelo maior ou igual a 0.5.

O uso do amostrador Gibbs e a inclusão de um vector de variáveis indicatrizes no algoritmo de amostragem possibilita a estimação das probabilidades mencionadas no parágrafo anterior. Esta estratégia deu origem a diversos métodos bayesianos de selecção de variáveis nos quais a pesquisa é feita, simultaneamente, no espaço dos modelos e no espaço dos parâmetros (Carlin e Louis, 2000), nomeadamente, o método de selecção de variáveis por pesquisa estocástica (SSVS) (George e McCulloch, 1993), o método das distribuições *a priori* não condicionais (KM) (Kuo e Mallick, 1998), o método Gibbs de selecção de variáveis (GVS) (Dellaportas *et al*., 1997) e o método MCMC dos saltos reversíveis (RJMCMC) (Green, 1995). Dellaportas *et al*. (2000) descreveram e compararam alguns destes métodos de selecção de variáveis baseados no amostrador Gibbs no contexto dos modelos lineares generalizados. O'Hara e Sillanpäa (2009) recorreram a exemplos reais e a dados simulados para estudar a performance destes métodos. Teles e Amaral Turkman (2010) efectuaram um estudo de simulação para averiguar a importância da dimensão da amostra e do número de preditores na performance do método GVS.

Recentemente têm surgido novas abordagens à selecção bayesiana de variáveis em modelos de regressão. Uma delas passa pela utilização do LASSO (*least absolute shrinkage and selection operator*) de Tibshirani (1996), que é um método de estimação dos coeficientes da regressão linear por mínimos quadrados penalizados com imposição da norma L_1. As estimativas LASSO podem ser interpretadas como estimativas das modas *a posteriori* dos parâmetros de regressão quando as suas distribuições *a priori* são independentes e identicamente distribuídas com distribuição de Laplace. Park e Casella (2008) referem que as estimativas intervalares, obtidas através da abordagem bayesiana do LASSO, podem ser usadas na selecção de variáveis em modelos de regressão. Hans (2009) introduz um novo amostrador Gibbs para a abordagem bayesiana à regressão LASSO e refere que a incerteza, relativa à especificação do modelo de regressão, pode ser incorporada na análise atribuindo uma distribuição *a priori* ao espaço dos modelos. Lykou e Ntzoufras (2010) propõem um método de selecção de variáveis que resulta da combinação do método LASSO e KM. A par com o LASSO, também o crescente desenvolvimento da metodologia ABC (*Approximate Bayesian Computation*) (*e.g.*, Beaumont *et al*., 2002 e Ratmann *et al*., 2009) perspectiva o aparecimento de outras soluções para o problema da selecção de variáveis em modelos de regressão. Esta é, portanto, uma área onde ainda existe muito trabalho por realizar e na qual se espera, a curto prazo, o surgimento de novas abordagens.

Agradecimentos

Ao Fundo Social Europeu e ao Estado Português que financiaram o doutoramento, no âmbito do 3º Quadro Comunitário de apoio, através do programa PRODEP 2001 para a formação Avançada de Docentes (Medida 5/Acção 5.3/Pedido de Financiamento nº 5.3/LVT/296.003/01 do Concurso 2/5.3/PRODEP/2001). Um agradecimento muito especial aos meus orientadores de doutoramento, a Professora Doutora Maria Antónia Amaral Turkman e o Professor Doutor Luís Canto de Loura, pela orientação, incentivo, amizade e paciência inexcedíveis que tornaram possível a realização da minha dissertação.

Referências

Aitkin, M. (1991). Posterior Bayes factor (with discussion). *J. Royal Statist. Soc.*, B, 53, 111-142.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Em B.N. Petrov e F. Czaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267-281. Academia Kiado, Budapest.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94, 443-458.

Barbieri, M.M. e Berger, J.O. (2003). Optimal predictive model selection. *Ann. Statist.*, 32, 870-897.

Beaumont, M.A., Zhang, W. e Balding, D.J. (2002). Approximate Bayesian Computation in population genetics. *Genetics*, 162, 2025-2035.

Berger, J.O. e Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, 91, 109-122.

Carlin, B.P. e Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall, New York.

Dellaportas, P., Forster, J.J. e Ntzoufras, I. (1997). *On Bayesian model variable selection using MCMC*. Technical Report, Department of Statistics, Athens University of Economics and Business, Greece.

Dellaportas, P., Forster, J.J. e Ntzoufras, I. (2000). Bayesian variable selection using the Gibbs sampler. Em D. Dey, S. Ghosh e B. Mallick (Eds.), *Generalized Linear Models: A Bayesian Perspective*, pp. 271-286. Marcel Dekker, New York.

Draper, D. e Krnjajic, M. (2007). *Bayesian model specification*. Technical Report ams2007-09, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

Gelfand, A.E., Dey, D.K. e Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods (with discussion). Em J.M. Bernardo, J.O. Berger, A.P. Dawid e A.F.M. Smith (Eds.), *Bayesian Statistics*, 4, pp. 147-167. Oxford University Press, Oxford.

George, E.I. e McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, 88, 881-889.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711-732.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96, 835-845.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, Oxford.

Kass, R.E. e Raftery, A.E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90, 773-795.

Kuo, L. e Mallick, B. (1998). Variable selection for regression models. *Sankhya,* B, 60, 65-81.

Lykou, A. e Ntzoufras, I. (2010). On Bayesian Variable Selection Using Lasso and Related Methods. *Book of Abstracts of The Ninth Valencia International Meeting on Bayesian Statistics and The 2010 ISBA World Meeting*, pp. 188-189. Benidorm, Spain.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Wiley Series in Computational Statistics. Hoboken, New Jersey.

O'Hagan, A. (1991). (In discussion following a paper by Aitkin) Posterior Bayes factors. *J. Royal Statist. Soc.*, B, 53, 136.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *J. Royal Statist. Soc.*, B, 57, 99-138.

O'Hara e Sillanpäa (2009). A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis*, 4, 85-118.

Park, T. e Casella, G. (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.*, 103, 681-686.

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

Ratmann, O., Andrieu, C., Wiuf, C. e Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 106, 10576-10581.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6, 461-464.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. e van der Linden, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Soc.*, B, 64, 583-639.

Teles, J. (2005). *Uma Abordagem Bayesiana à Determinação de Modelos*. Tese de doutoramento, Faculdade de Motricidade Humana da Universidade Técnica de Lisboa.

Teles, J. e Amaral Turkman, M.A. (2004). Selecção bayesiana de covariáveis em modelos de regressão linear. Em C.A. Braumann, P. Infante, M.M. Oliveira, R. Alpízar-Jara e F. Rosado (Eds.), *Estatística Jubilar – Actas do XII Congresso Anual da SPE*, pp. 801-808. Edições SPE, Lisboa.

Teles, J. e Amaral Turkman, M.A. (2007). Bayesian model selection criteria: a comparative study through simulation. Em J. del Castillo, A. Espinal, e P. Puig (Eds.), *Proceedings of the 22nd International Workshop on Statistical Modelling*, pp. 560-563. Institut d'Estadistica de Catalunya, Barcelona.

Teles, J. e Amaral Turkman, M.A. (2010). Some remarks about Gibbs variable selection performance. [Submetido à *Studies in Theoretical and Applied Statistics. Springer Series*.]

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Royal Statist. Soc.*, B, 58, 267-288.

Tese de Doutoramento: Medidas de Desempenho para Testes de Discordância em Populações Normais (Boletim SPE Primavera de 2007, p. 54)

José Palma, *jose.palma@estsetubal.ips.pt*

Departamento de Matemática, Escola Superior de Tecnologia de Setúbal,
Instituto Politécnico de Setúbal

1. INTRODUÇÃO

A detecção de outliers complica-se no momento da escolha da estatística de teste a adoptar face à grande variedade existente. Para populações com distribuição normal, Barnett e Lewis (1994) apresentam um conjunto de mais de quarenta testes. No entanto, a opção por uma ou outra estatística é uma questão fundamental pois uma observação poderá ser considerada outlier por um teste e não o ser por outro.
Face à multiplicidade de métodos de construção de testes de discordância e de estatísticas de teste daí resultantes torna-se necessário avaliar a sua eficácia. O cálculo das medidas de desempenho requer o conhecimento da distribuição da estatística de teste na hipótese alternativa que consideramos para explicar os outliers. Isto coloca problemas de difícil solução, especialmente no quadro da distribuição normal.
Este trabalho teve por objectivo o cálculo das medidas de desempenho para as estatísticas de teste obtidas a partir da aplicação do método Generativo de Alternativa Natural, no quadro de populações normais. Até esta data, os estudos de desempenho no âmbito da distribuição normal eram praticamente inexistentes.

2. ESTATÍSTICAS DE TESTE

Consideremos uma hipótese nula, H_0, de não existência de outliers, segundo a qual a amostra é extraída de uma determinada população F. Na formulação de um modelo de discordância, pelo qual se irá introduzir a hipótese de existência de outliers, têm sido consideradas diversas hipóteses alternativas $\overline{H}$. Pela hipótese $\overline{H}$ de alternativa natural, admitimos a presença de um valor discordante na amostra e que, em princípio, pode ser qualquer uma das n observações. Supondo que x_j é a observação discordante, então na sub-hipótese $\overline{H}_j$ admite-se que x_j tem densidade de probabilidade $f(x;\delta')$ para algum $j \in [1,2,\ldots,n]$ e os restantes x_i (com $i \neq j$) têm densidade de probabilidade $f(x;\delta)$. Com base nesta hipótese podemos determinar as estatísticas de teste. Vejamos alguns exemplos.
No quadro da distribuição normal na hipótese de μ e σ conhecidos e supondo μ', o parâmetro subjacente à distribuição da observação outlier, desconhecido e que não temos informação suplementar, a estatística de teste é dada por

$$S_3 = \max_j \left| \frac{x_j - \mu}{\sigma} \right|.$$

Como candidatos a outlier o método GAN propõe neste caso as observações $x_{(1)}$ e $x_{(n)}$.
As funções de distribuição para a estatística de teste, respectivamente na hipótese nula e alternativa

$$F^0_{S_3}(s) = P(S_3 \le s | H_0) = [\phi(s) - \phi(-s)]^n, \text{ com } s \ge 0$$

$$F^1_{S_3}(s) = [\phi(s) - \phi(-s)]^{n-1}\left[\phi\left(s - \frac{\delta}{\sigma}\right) - \phi\left(-s - \frac{\delta}{\sigma}\right)\right], \text{ com } s \ge 0.$$

Na hipótese de μ conhecido, μ' desconhecido com informação suplementar $\mu < \mu'$ a estatística de teste é dada por

$$S_2 = \max_j \frac{x_j - \mu}{\sigma},$$

com região de rejeição da forma $S_2 > c$ e candidato a outlier $x_{(n)}$, desde que $c \ge 1$. As funções de distribuição para a estatística de teste, respectivamente na hipótese nula e alternativa

$$F^0_{S_2}(s) = [\phi(s)]^n \text{ e } F^1_{S_2}(s) = [\phi(s)]^{n-1}\left[\phi\left(s - \frac{\delta}{\sigma}\right)\right].$$

3. MEDIDAS DE DESEMPENHO

Assumindo a alternativa de deslizamento ou a alternativa natural, uma das observações da amostra sob $\overline{H}$ será o contaminante, suponha-se que é a observação x_n; e supondo que S é uma estatística de teste, existe a correspondente medida S_n para o contaminante. David (1981) e Braumann (1994) sugerem as seguintes cinco probabilidades como medidas do desempenho das estatísticas de teste: sendo $\Re$ a região de rejeição, $P_1 = P(S \in \Re \mid \overline{H})$ probabilidade de se concluir pela existência de contaminação admitindo que ela existe, noutras palavras a função potência; $P_2 = P(S_n \in \Re \mid \overline{H})$ probabilidade de x_n ser suficientemente discordante, admitindo que há contaminação; $P_3 = P(S \in \Re \text{ e } S = S_n \mid \overline{H})$ probabilidade do contaminante ser o outlier e ser identificado como discordante, admitindo que há contaminação; $P_4 = P(S \in \Re \text{ e } S = S_n \text{ e } S_1, \ldots, S_{n-1} \notin \Re \mid \overline{H})$ probabilidade de se concluir pela existência de contaminação e a única observação que satisfaz o critério de discordância ser a observação contaminante x_n, admitindo que há contaminação; $P_5 = P(S_n \in \Re \mid S = S_n, \overline{H})$ probabilidade de a observação contaminante satisfazer o critério de discordância caso tenha sido considerada responsável e que haja contaminação. Esta medida é equivalente a P_3/P_6, com $P_6 = P(S = S_n \mid \overline{H})$, probabilidade de se considerar responsável a observação contaminante caso haja contaminação. Braumann (1994) considera ainda a medida $P_7 = (P_1 - P_3)$.
Vejamos algumas medidas de desempenho para os exemplos citados. No caso da estatística S_2 :

$$P_1 = 1 - [\phi(c)]^{n-1}\left[\phi\left(c - \frac{\delta}{\sigma}\right)\right], \; P_2 = 1 - \phi\left(c - \frac{\delta}{\sigma}\right)$$

$$P_3 = \int_c^{+\infty} (\phi(y))^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(y - \frac{\delta}{\sigma}\right)^2}{2}\right) dy, \; P_4 = [\phi(c)]^{n-1}\left[1 - \phi\left(c - \frac{\delta}{\sigma}\right)\right],$$

$$P_6 = \int_{-\infty}^{+\infty} (\phi(y))^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left(y - \frac{\delta}{\sigma}\right)^2}{2}\right) dy.$$

As restantes medidas de desempenho podem obter-se a partir das anteriores.
Para a estatística S_3 :

$$P_1 = 1 - [\phi(c) - \phi(-c)]^{n-1}\left[\phi\left(c - \frac{\delta}{\sigma}\right) - \phi\left(-c - \frac{\delta}{\sigma}\right)\right] \ , \ P_2 = 1 - \left[\phi\left(c - \frac{\delta}{\sigma}\right) - \phi\left(-c - \frac{\delta}{\sigma}\right)\right],$$

$$P_3 = \int_c^{+\infty} [\phi(z) - \phi(-z)]^{n-1}\left[f_N\left(z - \frac{\delta}{\sigma}\right) + f_N\left(-z - \frac{\delta}{\sigma}\right)\right] dz \,,$$

$$P_4 = [\phi(c) - \phi(-c)]^{n-1}\left[1 - \left[\phi\left(c - \frac{\delta}{\sigma}\right) - \phi\left(-c - \frac{\delta}{\sigma}\right)\right]\right],$$

$$P_6 = \int_0^{+\infty} [\phi(z) - \phi(-z)]^{n-1}\left[f_N\left(z - \frac{\delta}{\sigma}\right) + f_N\left(-z - \frac{\delta}{\sigma}\right)\right] dz \,.$$

4. CONCLUSÕES

Foram calculados as diversas medidas de desempenho, para todas as estatísticas obtidas por Rosado [1984], foram ainda propostas algumas outras estatísticas, nomeadamente na situação em que se considera algum tipo de informação suplementar acerca do comportamento dos parâmetros das distribuições envolvidas. Foram determinadas as diferentes distribuições das estatísticas de teste na hipótese nula de ausência de outliers e na hipótese alternativa de contaminação. Foram ainda apresentados valores críticos para todas as estatísticas.

Em termos de resultados finais concluímos que em geral à medida que aumenta o tamanho da amostra o desempenho dos diferentes testes diminui. Desta conclusão resulta que se a amostra é grande dificilmente se detectará um outlier. Em amostras grandes só será viável a detecção quando os valores do parâmetro subjacente à observação outlier forem bastante desviantes. Esta conclusão só é contrariada em testes relativamente mais complexos, em que aparentemente o aumento da dimensão da amostra não implica necessariamente perda de eficácia.

A informação disponível é um elemento fundamental em termos de opção pelos diferentes testes e como tal na eficácia com que se detecta outliers. Assim, verifica-se a degradação das medidas de desempenho à medida que se considera testes obtidos na hipótese de menos informação disponível.

Vimos ainda a possibilidade do elemento mais próximo da média ser um outlier, o que constitui elemento inovador em relação aos métodos de detecção tradicionais que apenas detectam valores extremos. No entanto o desempenho das diferentes medidas, concebidas para detectar este tipo de contaminante, revela-se relativamente fraco.

Referências

Barnett, V., Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley & Sons. New York.

Braumann, M. M. (1994). *Sobre Testes de Detecção de Outliers em Populações Exponenciais*. Dissertação de Doutoramento. Universidade de Évora.

David, H. A. (1981). *Order Statistics.* John Wiley & Sons. New York.

Rosado, F. (1984). *Existência e Detecção de Outliers - Uma Abordagem Metodológica.* Tese de doutoramento. Faculdade de Ciências, Lisboa.

Tese de Doutoramento: Desigualdades Exponenciais e Velocidades de Convergência em Problemas de Estimação para Amostras Associadas

(Boletim SPE Primavera de 2007, p. 53)

Carla Henriques, *carlahenriq@estv.ipv.pt*

Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Viseu

No decurso do meu trabalho para o doutoramento, houve um tema que me seduziu particularmente e por isso decidi dedicar-lhe aqui umas poucas linhas. Refiro-me *aos grandes desvios*, em particular no âmbito da *associação*, estrutura de dependência positiva bastante estudada nas últimas décadas. A sedução pelo tema tem a ver, naturalmente, com o facto de ser promissor e muito interessante. Depois do doutoramento, infelizmente, dediquei-lhe pouco tempo porque outros projectos se interpuseram. Mas, sem dúvida, ficou o gosto pelo assunto e a grande vontade de continuar o trabalho iniciado que ainda tem muito para explorar.

O resultado estabelecido na minha tese, e publicado em Henriques e Oliveira (2008), fornece condições suficientes para que a sucessão de médias $\bar{X}_n = 1/n \sum_{i=1}^{n} X_i$, de uma sucessão estacionária e associada de variáveis aleatórias reais, satisfaça *o princípio de grandes desvios*. Este resultado constitui mais uma das muitas extensões do Teorema de Cramér, como é conhecido na literatura o resultado clássico que descreve o comportamento assimptótico de caudas da média empírica de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.). O Teorema de Cramér tem as suas raízes num trabalho de Cramér (1938), com contribuições de Chernoff (1952) e Bahadur (1971), e tem sido estendido em várias direcções, particularmente a situações em que as variáveis de base não são independentes. Vamos então fazer uma breve incursão na Teoria dos Grandes Desvios.

A Teoria de Grandes Desvios debruça-se sobre a descrição da velocidade de decrescimento exponencial de certas probabilidades, ditas probabilidades de grandes desvios, probabilidades essas associadas a acontecimentos "extremos" ou de "cauda".

Começo por apresentar um exemplo clássico, mas, sem dúvida, bastante ilustrativo. No lançamento sucessivo de uma moeda equilibrada, sabemos, pelas Leis dos Grandes Números, que há medida que aumenta o número de lançamentos, aumenta também a probabilidade de a proporção de faces se situar numa vizinhança restrita de 1/2. No entanto, apesar de muito pouco provável, pode acontecer que ao fim de um elevado número de lançamentos a proporção de caras seja superior a, por exemplo, 9/10, ou mesmo 99/100. Ao descrever a velocidade de decrescimento das probabilidades de desvios da proporção de faces ao seu valor esperado 1/2, poderemos aproximar a probabilidade de cada um daqueles acontecimentos extremos, obtendo assim uma ideia de quão pouco prováveis são. Na verdade, a probabilidade de um tal acontecimento extremo decresce exponencialmente à medida que *n* aumenta. Mais precisamente, sendo $\bar{X}_n$ a proporção de faces em *n* lançamentos, tem-se

$$\lim_{n\to+\infty} \frac{1}{n} \log P(\bar{X}_n \geq a) = -I(a),$$

onde $I(a) = \sup_{u\geq 0}\{au - \log(1/2 + 1/2e^u)\}$. Para *n* suficientemente grande, tem-se então $P(\bar{X}_n \geq a) \approx \exp(-nI(a))$, onde $I(a)$ descreve a velocidade de decrescimento exponencial de $P(\bar{X}_n \geq a)$. O resultado anterior é típico da teoria dos grandes desvios.

Em vez de *n* lançamentos de uma moeda, situemo-nos no contexto dos seguros e pensemos na utilidade, mais do que isso, na importância, de poder aproximar a probabilidade de certos acontecimentos extremos. Podemos, por exemplo, conjecturar que num determinado período sucedem muitos acidentes automóveis com carros segurados pela companhia, obrigando a companhia ao

pagamento de avultadas quantias, de tal forma que esta se encontraria numa situação, pouco provável mas possível, de prejuízo relativamente aos seguros do ramo automóvel. Acontecimentos desta natureza têm um impacto de tal forma grande, que se torna relevante poder quantificar, pelo menos aproximadamente, a probabilidade da sua ocorrência. Os exemplos que acabámos de apresentar podem facilmente ser transpostos pelo leitor para outros campos. De facto, a teoria dos grandes desvios tem hoje aplicações numa grande variedade de áreas, das quais são exemplos, a mecânica estatística, sistemas de filas de espera, redes de comunicação, teoria da informação (cf. Ellis (1995) e referências aí contidas). Particularmente na Inferência Estatística, constituem, por exemplo, uma ferramenta útil para a caracterização de velocidades de convergência de estimadores, tanto no sentido da convergência em probabilidade, como no sentido da convergência quase certa. Têm também um papel crucial na comparação assimptótica de testes de hipóteses. Com efeito, o cálculo das eficiências relativas assimptóticas, no sentido de Bahadur, Hodges-Lehmann ou Chernoff têm por base resultados de grandes desvios (cf., por exemplo, van der Vaart (1998) ou Nikitin (1995)).

A teoria dos grandes desvios teve origem em problemas relacionados com somas de variáveis aleatórias reais, independentes e identicamente distribuídas, num esforço para ir um pouco além das Leis dos Grandes Números e do Teorema Limite Central. Sejamos mais precisos. As Leis dos Grandes Números estabelecem a convergência para zero da média empírica centrada no seu valor esperado. Isto é, supondo, para simplificar, que $X_1, \cdots, X_n$ são variáveis aleatórias i.i.d., com $\mathrm{E}(X_i)=\mu$ e $\mathrm{Var}(X_i)=\sigma^2$, então, quando n cresce, a distribuição de $\bar{X}_n$vai-se tornando mais concentrada à volta de μ e, consequentemente, as caudas dessa distribuição vão diminuindo. A questão natural que se colocou aos investigadores foi a de descrever a velocidade de decrescimento destas caudas. Isto é, dado $t \in \mathbb{R}$, interessava caracterizar a velocidade de decrescimento das probabilidades $P(|\bar{X}_n - \mu| \geq t)$. A teoria dos grandes desvios debruça-se sobre este tipo de questões. Refira-se ainda que o Teorema Limite Central fornece uma aproximação para a probabilidade de um desvio da forma $t/\sqrt{n}$, da média empírica relativamente ao seu valor esperado:

$$P\left(\bar{X}_n - \mu \geq t\sigma/\sqrt{n}\right) \cong P(N(0,1) \geq t),$$

para n suficientemente grande. Então, o Teorema Limite Central fornece aproximações para "pequenos desvios", como deverão ser considerados por comparação com os que são estudados na teoria dos grandes desvios, $P(\bar{X}_n - \mu \geq t)$. Entre o Teorema Limite Central e os resultados de grandes desvios, encontram-se os resultados de desvios moderados que estudam o comportamento assimptótico de probabilidades do tipo $P(\bar{X}_n - \mu \geq t/a_n)$, $a_n \to +\infty$ com velocidade inferior a $\sqrt{n}$.

Um resultado clássico da teoria dos grandes desvios, atribuído na literatura a Cramér e a Chernoff, estabelece que, assumindo para simplificar que as variáveis têm médias nulas,

$$\lim_{n\to+\infty} \frac{1}{n} \log P(\bar{X}_n \geq a) = \inf_{u\geq 0}\{\log \mathrm{E}(e^{uX_1}) - au\}$$

(ver, por exemplo, Proposição 14.23 de van der Vaart (1998)). Assim, em relação à Lei Forte dos Grandes Números e ao Teorema Limite Central, este resultado permite ir um pouco mais longe na medida em que fornece aproximações para probabilidades de grandes desvios.

A teoria dos grandes desvios está hoje largamente desenvolvida, tendo ultrapassado o âmbito das somas de variáveis aleatórias reais. Numa perspectiva mais abrangente, consideremos uma sucessão $\{X_n, n \geq 1\}$ de elementos aleatórios em $(S, \mathcal{B}_S)$, onde S é um espaço métrico e $\mathcal{B}_S$ a sua σ-álgebra de Borel. Para cada $n\geq 1$, admitimos que X_n está definida no espaço de probabilidade $(\Omega_n, \mathcal{F}_n, P_n)$. Representamos por $d(\cdot,\cdot)$ a distância em S e por Q_n a distribuição de X_n em $(S, \mathcal{B}_S)$. À semelhança do que acontece com a média empírica de n variáveis aleatórias i.i.d., em muitas situações $\{X_n, n \geq 1\}$ converge fracamente para um ponto $x_0 \in S$. Neste caso, dado $A \in \mathcal{B}_S$ tal que o fecho de A não contém o ponto x_0, tem-se $Q_n(A) = P_n(X_n \in A) \overset{n}{\to} 0$. Uma questão que imediatamente se levanta é a de investigar a velocidade a que estas probabilidades convergem para zero. Em muitos casos é possível demonstrar que esta velocidade é exponencial, isto é,

$$Q_n(A) \sim \exp(-a_n I(a)),$$

onde $I(a) = \inf_{x \in A} I(x)$, I é uma função de S em $[0,+\infty]$ e $\{a_n, n \geq 1\}$ é uma sucessão tal que $a_n \to +\infty$. Interessa pois investigar em que situações se verifica aquela relação, bem como identificar, se possível, a sucessão $\{a_n, n \geq 1\}$ e a função I. Estas questões estão no âmbito da teoria dos grandes desvios.

É de referir que grande parte dos resultados de grandes desvios não investigam directamente o comportamento assimptótico das probabilidades $Q_n(A)$, mas sim o de $\log Q_n(A)$.

A literatura sobre este assunto é hoje muito vasta. Para além de um grande número de artigos, existem também vários livros que abordam a teoria dos grandes desvios sob diversos aspectos. O artigo de Cramér (1938) é um dos primeiros trabalhos de relevo no âmbito dos grandes desvios. Citando Ellis (1995),

> "*The seed from which this forest grew is the 1938 paper of Harald Cramér which treats the large deviation theory of the sample means of independent, identically distributed (i.i.d.) random variables*."

Depois de Cramér surgem inúmeros trabalhos fundamentais, com contribuições relevantes para o desenvolvimento da teoria dos grandes desvios. Apresentando uma lista muitíssimo incompleta, atrevemo-nos a destacar os trabalhos de Chernoff (1952), Bahadur e Ranga Rao (1960), Bahadur (1971), Bahadur e Zabell (1979), Donsker e Varadhan (1976), Borovkov e Mogulskii (1978, 1980) (grandes desvios da média empírica, incluindo, nos últimos, versões em espaços gerais), Sanov (1957) (grandes desvios da função de distribuição empírica), Varadhan (1966), Donsker e Varadhan (1975a, 1975b, 1976}, Freidlin e Wentzell (1970, 1972, 1984} (formulação do princípio dos grandes desvios num espaço geral; desenvolvimentos no âmbito dos processos estocásticos). Muito do material publicado nestes e noutros artigos pode ser encontrado em livros, de que são exemplo, Saulis e Statulevicius (1991), Dupuis e Ellis (1997) e Dembo e Zeitouni (1998).

O princípio de grandes desvios (PGD) da média empírica tem sido estudado considerando diversas condições dependência. Na minha tese foi estabelecido um PGD para a média empírica, considerando que as variáveis de base satisfazem uma estrutura de dependência dita associação positiva ou, simplesmente, associação. A noção de associação foi introduzida na estatística por Esary, Proschan e Walkup (1967), que encontraram motivação em aplicações no âmbito da teoria da fiabilidade. Independentemente, esta noção surge também na área da mecânica estatística com o trabalho de Fortuin, Kasteleyn e Ginibre (1971), o qual deu nome a este conceito na literatura daquela área (diz-se que as variáveis satisfazem as desigualdades FKG). O resultado estabelecido na minha tese segue a mesma linha de demonstração do Teorema 6.4.4 de Dembo e Zeitouni (1998), onde se estabelece o PGD sob uma certa condição de mistura. Como disse anteriormente, muito ficou por explorar, o que não deixa de ser ao mesmo tempo inquietante e desafiador: A inquietude própria da dificuldade de gerir muitos projectos ao mesmo tempo, ficando forçosamente alguns para trás, e o desafio de explorar um campo que promete.

Antes de finalizar, e já que estamos a falar de grandes desvios, não posso deixar de salientar o reconhecimento do trabalho de Varadhan através do prémio Abel 2007,

> "*for his fundamental contributions to probability theory and in particular for creating a unified theory of large deviations.*"
>
> (*http://www.abelprisen.no/en/prisvinnere/2007/index.html*)

Referências

Bahadur, R.R. (1971). Some limit theorems in statistics. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 4., SIAM.

Bahadur, R.R. e Ranga Rao, R. (1960). On deviations of the sample mean. *Ann. Math. Statist.* 31, 1015-1027.

Bahadur, R.R. e Zabell, S.L. (1979). Large deviations of the sample mean in general vector spaces. *Ann. Probab.* 7, 587-621.

Borovkov, A.A. e Mogulskii, A.A. (1978). Probabilities of large deviations in topological spaces, I (in Russian). *Siberian Mathem. J.* 19, 988-1004.

Borovkov, A.A. e Mogulskii, A.A. (1980). Probabilities of large deviations in topological spaces, II (in Russian). *Siberian Mathem. J.* 21, 12-26.
Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on sums of observations. *Ann. Math. Statist.* 23, 493-507.
Cramér, H. (1938). Sur un nouveau théorème limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles* 736, 5--23. Colloque consacré àla théorie dês probabilités, Vol. 3, Hermann, Paris.
Dembo, A. e Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer-Verlag, New York.
Donsker, M.D. e Varadhan, S.R.S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.* 28, 1-47.
Donsker, M.D. e Varadhan, S.R.S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, II. *Comm. Pure Appl. Math.* 28, 279-301.
Donsker, M.D. e Varadhan, S.R.S. (1976). Asymptotic evaluation of certain Markov processexpectations for large time, III. *Comm. Pure Appl.Math.* 29, 389-461.
Dupuis, P. e Ellis, R.S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley Series in Probability and Statistics. John Wiley and Sons, New York.
Ellis, R.S. (1995). An overview of the theory of large deviations and applications to statistical mechanics. Harald Cramér Symposium (Stockholm, 1993). *Scand. Actuar. J* 1, 97-142.
Esary, J.D., Proschan, F. e Walkup, D. W. (1967). Association of random variables, with applications. *Ann. Math. Statist.* 38, 1466-1474.
Fortuin, C.M., Kasteleyn, P.W. e Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Commun. Math. Phys.* 22, 89-103.
Freidlin, M.I. e Wentzell, A.D. (1970). On small random perturbations of dynamical systems. *Russian Math. Surveys* 25, 1-5.
Freidlin, M.I. e Wentzell, A.D. (1972). Some problems concerning stability under small random perturbations. *Th. Probab. Appl.* 17, 269-283.
Freidlin, M.I. e Wentzell, A.D. (1984). *Random Perturbations of Dynamical Systems*. Translated by J. Szücs. Springer-Verlag, New York.
Gärtner, J. (1977). On large deviations from the the invariant measure. *Th. Probab. Appl.* 22, 24-39.
Henriques, C. e Oliveira, P.E. Large deviations for the empirical mean of associated random variables. *Statistics and Probability Letters* 78, 594-598.
Nikitin, Y. (1995). *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press, Cambridge.
Sanov, I.N. (1957). On the probability of large deviations of random variables (in Russian). *Mat.Sb.* 42, 11-44. English translation in *Selected Translations in Mathematical Statistics an Probability* 1, 213-244, 1961.
Saulis, L. e Statulevicius, V.A. (1991). Limit Theorems for Large Deviations. Kluwer Academic Publishers, Dordrecht, Netherlands.
van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge.
Varadhan, S.R.S. (1966). Asymptotic probability and differencial equations, *Comm. Pure Appl. Math.* 19, 261-289.

Ciência Estatística

• Artigos Científicos Publicados

Andrade, M. e Ferreira, M. A. M. (2010). Object-oriented Bayesian Networks in the evaluation of paternities in less usual environments, *Journal of Mathematics and Technology*, 1(1), pp. 161-164.

Andrade, M. e Ferreira, M. A. M. (2010). Solving civil identification cases with DNA profiles databases using Bayesian networks, *Journal of Mathematics and Technology*, 1(2), pp. 37-40.

Andrade, M. e Ferreira, M. A. M. (2010). Civil Identification Problems with Bayesian Networks using Official DNA Databases, *Aplimat- Journal of Applied Mathematics*, 3(3), pp. 155-162.

Andrade, M., A Note on Foundations of Probability. Journal of Mathematics and Technology, Volume 1 (1), 96-98, 2010.

Caiado, J. (2010). Performance of combined double seasonal univariate time series models for forecasting water demand. *Journal of Hydrologic Engineering*, 15, 215-222.

Costa, C., Scotto, M. G. and Pereira, I. (2010). Optimal alarm systems for FIAPARCH processes. *REVSTAT* 8, 37-55.

Diggle, P., R.Menezes and Ting-Li Su (2010). Geostatistical Inference under Preferential Sampling (with discussion). *Journal of Royal Statistics Society*, series C, vol 59, part 2, 191-232

Ferreira, M. A. M. e Andrade, M. (2010). M/G/oo System Transient Behavior with Time Origin at the Beginning of a Busy Period Mean and Variance, *Aplimat- Journal of Applied Mathematics*, 3(3), pp. 213-221.

Ferreira, M. A. M. e Filipe, J. A. (2010). Solving Logistics Problems using M/G/oo Queue Systems Busy Period, *Aplimat- Journal of Applied Mathematics*, 3(3), pp. 207-212.

Ferreira, M. A. M. (2010). A Note on Jackson Networks Sojourn Times, *Journal of Mathematics and Technology*, 1(1), pp. 91-95.

Ferreira, M. A. M. e Andrade, M. (2010). Looking to a M/G/00 system occupation through a Ricatti equation, *Journal of Mathematics and Technology*, 1(2), pp. 58-62.

Ferreira, M. A. M. e Andrade, M. (2010). M/M/m/m Queue System Transient Behavior, *Journal of Mathematics and Technology*, 1(1), pp. 49-65.

Ferreira, M. and Canto e Castro, L. (2010). Modeling rare events through a pRARMAX process. *J. Statist. Plann. Inference* 140(11), 3552-3566.

Gil, P. M., Fernanda Figueiredo and Óscar Afonso (2010). Equilibrium Price Distribution with Directed Technical Change. *Economic Letters* 108, 130-133.

Guerreiro, G.R., Mexia, J.T., Miguens, M.F. (2010), A Model for Open Populations subject to periodical re-classifications, *Journal of Statistical Theory and Practice*, Vol. 4, Nº2, 303-321.

Hall, A., Scotto, M. G., and Cruz, J. (2010). Extremes of integer-valued moving average sequences. *Test* 19, 359-374.

Macedo, P., Scotto, M. G. and Silva, E. (2010). A general class of estimators for the linear regression model affected by collinearity and outliers. *Communications in Statistics - Simulation and Computation* 39, 981-993.

Menezes, R., P.Garcia-Soidán and C.Ferreira (2010). Nonparametric spatial prediction under stochastic sampling design. *Journal of Nonparametric Statistics*, vol 22, issue 3, 363-377.

Monteiro, M., Scotto, M. G. and Pereira, I. (2010). Integer-valued autoregressive processes with periodic structure. *Journal of Statistical Planning and Inference* 140, 1529-1541.

Pereira, L.N. e Coelho, P.S (2010). Small area estimation of mean price of habitation transaction using time-series and cross-sectional area level models. *Journal of Applied Statistics*, 37:4, 651-666.

Pereira, L.N. e Coelho, P.S. (2010). Assessing different uncertainty measures of EBLUP: a resampling-based approach. *Journal of Statistical Computation and Simulation*, 80:7, 713-727.

Sáfadi, T. e Pereira, I. (2010). Bayesian Analysis of FIAPARCH. Model: An Application to São Paulo Stock Market, *International Journal of Statistics and Economics,* **5**, 49-63.

Sampaio,P., Calado,C., Sousa,L., Bressler,D., Pais,M. and Fonseca,L. (2010). Optimization of the culture medium composition using response surface methodology for new recombinant cyprosin B production in bioreactor for cheese production. *European Food Research and Technology* 231: 339–346.

Scotto, M. G., Alonso, A. M. and Barbosa, S. M. (2010). Clustering time series of sea levels: extreme value approach. *Journal of Waterway, Port, Coastal, and Ocean Engineering* 136, 215-225.
Sepulveda, N., Paulino, C. D. and Carneiro, J. (2010). Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *J. Immunol. Methods* 353:124-137.

• Teses de Mestrado

Título: *Estudo Hidrogeológico da Sub-Bacia Hidrográfica de Alcântara-Lisboa*
Autora: Margarida Duarte de Oliveira, *margaridaoliveira4@gmail.com*
Orientadoras: Fernanda Diamantino e Catarina Silva

• Livros

Título: Análise Estatística com o PASW Statistics (ex-SPSS).
Autor: João Marôco
Ano: 2010.Editora: ReportNumber, Lda. ISBN: 978-989-967-63-0-5

Título: Classification and Clustering of Time Series
Autor: Jorge Caiado
Ano: 2010.Editora: Lambert Academic Publishing. ISBN: 978-3-8383-4181-1

Título: Controlo Estatístico da Qualidade
Autoras: M. Ivette Gomes, Fernanda Figueiredo e M. Isabel Barão
Ano: 2010.Editora: Sociedade Portuguesa de Estatística. ISBN: 978-972-8890-23-0

Título: Uma Introdução à Estimação Não-Paramétrica da Densidade
Autor: Carlos Tenreiro
Ano: 2010.Editora: Sociedade Portuguesa de Estatística

• Teses de Doutoramento

Título: Contributos para a Análise Estatística de Séries de Contagem
Autor: Magda Sofia Valério Monteiro, *msvm@ua.pt*
Orientadores: Isabel Maria Simões Pereira e Manuel Gonzaléz Scotto

A minha tese incidiu na análise estatística de séries temporais de valores inteiros não negativos descritas por modelos auto-regressivos de primeira ordem baseados no operador *binomial thinning*.
O meu trabalho abordou por um lado, a predição de acontecimentos, aos quais estão associados mecanismos de alarme, e por outro lado estudaram-se modelos auto-regressivos com estrutura periódica e modelos definidos por limiares auto-induzidos.

Relativamente à predição de acontecimentos foram desenvolvidos, nas abordagens clássica e bayesiana, sistemas de alarme óptimos para processos de contagem. A construção destes sistemas de alarme óptimos incidiu na adaptação dos princípios que Lindgren (1975ª, 1975b, 1980) e De Maré (1980) estabeleceram para sistemas de alarme óptimos para modelos de valores contínuos, a modelos de valores inteiros. A abordagem bayesiana sugerida por Amaral Turkman e Turkman (1990) foi também usada e adaptada a modelos de valores inteiros. A aplicação dos sistemas de alarme óptimos focou-se particularmente no modelo auto-regressivo de valores inteiros não negativos com coeficientes estocásticos, designado por DSINAR(1). Para complementar o estudo de simulação realizado fez-se uma aplicação dos sistemas de alarme óptimos ao número de manchas solares observadas pelo Observatório Astrofísico de Catania.

Em termos da modelação de séries de contagem foram propostos e estudados os modelos $PINAR(1)_T$ e SETINAR(2;1) que até ao momento ainda não tinham sido alvo de análise na literatura.
O primeiro modelo é caracterizado por ter uma estrutura periódica, de período T, em que as inovações constituem uma sucessão periódica de variáveis aleatórias independentes com distribuição de Poisson. Este modelo foi estudado com detalhe ao nível das suas propriedades probabilísticas, métodos de estimação e previsão e através de um estudo de simulação para comparar o desempenho dos estimadores. Foi ainda efectuada uma ilustração deste modelo usando os dados correspondentes ao desemprego mensal de curta duração no Concelho de Penamacor entre 1997 e 2007.
O modelo SETINAR(2;1) é definido por limiares auto-induzidos e com inovações constituídas por uma sucessão de variáveis independentes e identicamente distribuídas com distribuição de Poisson. Este modelo é caracterizado por dois regimes auto-regressivos de ordem um e caracteriza-se pelo facto de a mudança de regimes ser efectuada tendo em conta o valor do próprio processo no instante anterior. Para este modelo estudaram-se as suas propriedades probabilísticas, métodos para estimar os seus parâmetros e foram realizados estudos de simulação para comparar os métodos de estimação apresentados.

Magda Monteiro

Título: Amostragem por distâncias: efeito da distribuição espacial e adaptação para terrenos montanhosos.
Autora: Anabela Cristina Cavaco Ferreira Afonso, *aafonso@uevora.pt*
Orientador: Russell Gerardo Alpizar-Jara

Na minha tese efectua-se uma avaliação do desempenho dos estimadores da amostragem por distâncias com populações distribuídas de forma espacial não homogénea ou com tendência para formar agrupamentos. Avalia-se também a performance destes estimadores em terrenos com diferentes declives e propõe-se um novo estimador para terrenos de declive variável. Sugerem-se ainda novos

planos de amostragem por distâncias adaptativos onde os transectos pontuais são colocados sob os transectos lineares nas zonas de maior abundância.

Verifica-se que segundo o plano de amostragem, a precisão do estimador convencional da densidade depende tanto do número de transectos colocados, como da sua orientação e da distribuição espacial dos animais na área de estudo. Observa-se que em planos de amostragem que combinam a colocação sistemática com aleatória de transectos pontuais em linha ou coluna, o estimador convencional não tem boas propriedades.

Mostra-se que o estimador convencional da probabilidade de detecção, abundância e densidade de animais da amostragem por transectos lineares, não são robustos ao relevo do terreno, podendo apresentar enviesamentos severos e uma grande variabilidade. Face a esta evidência, desenvolvem-se novos estimadores para a probabilidade de detecção tendo em conta que nestes terrenos a distância de observação à esquerda e à direita pode variar significativamente. Um dos estimadores propostos revela ter melhor desempenho que o estimador convencional, numa superfície com declive variável.

Identificam-se algumas condicionantes na aplicação dos planos convencionais em terrenos montanhosos. Sugerem-se planos de amostragem adaptativos que combinam os transectos lineares com os pontuais, uma vez que é habitual neste tipo de terrenos o observador parar para registar as detecções efectuadas e tentar aumentar a dimensão das suas amostras.

Anabela Afonso

Título: A Matemática, a Estatística e o ensino nos estabelecimentos de formação de Oficiais do Exército Português no período 1837-1926: uma caracterização
Autor: Filipe José Loureiro Lopes Papança, *filipe.papanca@gmail.com*
Orientadores: António Manuel Águas Borralho e José Manuel Leonardo de Matos

A minha tese tem como objectivo estudar a evolução da formação de Oficiais do Exército Português no período 1837-1926 em especial nas vertentes da Matemática e da Estatística.
Em particular este trabalho procura responder às seguintes questões:

a) Como se pode caracterizar, em termos de conteúdos, a formação militar nas áreas da Matemática e da Estatística ministradas em cursos de formação de oficiais do exército? Quais os critérios que estiveram na base da escolha desses conteúdos, considerados fundamentais para a sua formação? Como se pode caracterizar o contexto educativo castrense, em particular nas áreas da Matemática e da Estatística em termos de formação de Oficiais do Exército, assim como em outros cursos de formação ministrados nessas instituições?
b) Qual o papel das representações e da Matemática e da Estatística nos momentos solenes?
c) Qual o papel da Estatística no funcionamento da instituição? A dissertação efectua uma análise baseada em fontes documentais (recolhidas em bibliotecas do exército) do quotidiano da formação Oficiais do Exército neste período, em particular das secções de ensino, regulamentos, estatísticas, professores, cerimoniais, visitas de estudo e livros escritos por docentes. Foi feito um levantamento exaustivo da organização curricular, docentes e manuais produzidos por professores da Escola.

Ao longo do período estudado, a organização curricular procurou-se adaptar às novas exigências tecnológicas. Procurou ainda dotar a Escola de um ensino prático e laboratorial. Nos momentos solenes a Matemática desempenha diversas funções. A Estatística desempenha um papel importante como factor organizativo e de coesão. Em termos curriculares constata-se a não existência de uma visão unificada da Estatística.

Filipe Lopes Papança

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIOS "ESTATÍSTICO JÚNIOR 2010"

Trabalho classificado em 1º lugar (Ensino Básico)

Título: *"Atitude Ecológica = Escola com menos Resíduos"*
Autores: Eva Pinheiro Lisboa, Ana Carolina M. Fontes, Ana Sofia Tavares J. Alves
Professora orientadora: Marta Maria Direito P. Caseiro
Estabelecimento de Ensino: Colégio Senhor dos Milagres, Leiria

Trabalho classificado em 2º lugar (Ensino Básico)

Título: *"Ardósia Interactiva"*
Autor: Carlos Moura P. Lucas Teixeira
Professora orientadora: Maria Augusta Carvalho de Azevedo
Estabelecimento de Ensino: Escola Básica D. Manuel I, Tavira

Trabalho classificado em 3º lugar (Ensino Básico)

Título: *"Comportamentos Sociais face a uma possível pandemia e impacto da "Gripe A" na escola"*
Autora: Ana Beatriz Correia Lopes
Professora orientadora: Teresa de Jesus Poço Isabel
Estabelecimento de Ensino: Agrupamento de Escolas Artur Gonçalves, Torres Novas

Trabalho classificado em 1º lugar (Ensino Secundário)

Título: *"A importância de sermos saudáveis: um estudo com alunos do 7º e 12º anos"*
Autores: Susana Alice F. Cunha, Andreia Filipa G. Vale
Professora orientadora: Carla Manuela Navio Dias
Estabelecimento de Ensino: Escola Secundária Padre Benjamim Salgado, Vila Nova de Famalicão

Trabalho classificado em 2º lugar (Ensino Secundário)

Título: *"Jogo da Glória - Aleatórios e simulações"*
Autor: Emanuel da Silva Boavista
Professor Orientador: José Eduardo F. Cunha
Escola:Escola Secundária Barcelos

Trabalhos classificados em 3º lugar (exæquo) (Ensino Secundário)

Título: *"Caracterização socioeconómica dos alunos da EPAFB"*
Autores: Charlene Ribeiro Alves, Maria João Teixeira de Sousa, Sónia Catarina Gonçalves
Professora orientadora: Patrícia Alexandra S. Ribeiro Sampaio
Estabelecimento de Ensino: Escola Profissional do Fermil de Basto, Molares, Celorico de Basto

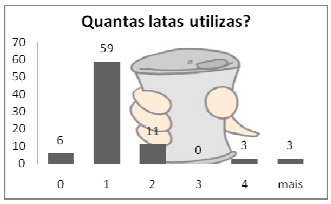
Título: *"As obras no liceu Camões"*
Autores: Ana Rita Silva Costa, André Gonçalo C. F. Almeida Costa
Professora orientadora: Luísa Margarida Cunha
Estabelecimento de Ensino: Escola Secundária de Camões, Lisboa

Trabalho classificado em 1.º lugar (Ensino Básico)

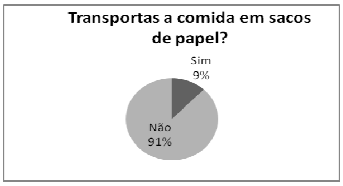
Atitude Ecológica:

Escola com menos Resíduos

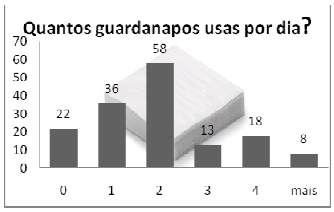
As latas de alumínio levam 100 a 500 anos a decompor-se.



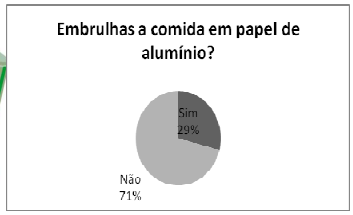
As embalagens de papel levam de 1 a 4 meses a decompor-se.



Os guardanapos levam cerca de 3 meses a decompor-se.



O papel de alumínio leva mais de 500 anos a decompor-se.



Os recipientes de vidro levam cerca de 1 milhão de anos a decompor-se.

O plástico leva cerca de 100 anos a decompor-se.

Sabes que:

Com a reciclagem de uma lata de alumínio economiza-se a energia suficiente para manter ligada uma televisão durante três horas.

A produção de papel reciclado consome menos cerca de 50% de energia, comparativamente com a produção a partir das árvores. Para além disso, a poluição do ar é reduzida em 95%.

Na produção de uma tonelada de papel reciclado são necessários apenas 2000 litros de água, enquanto no processo tradicional esse volume de água pode chegar aos 100 000 litros por tonelada de papel fabricado.

Ana Alves, Ana Fontes e Eva Lisboa - 7ºA
Concurso Nacional de Estatística
Colégio Senhor dos Milagres - Ano Lectivo 2009/2010

PRÉMIOS "ESTATÍSTICO JÚNIOR 2010"

Trabalho classificado em 1.º lugar (Ensino Secundário)

Escola Secundária Padre Benjamim Salgado

A importância de sermos saudáveis: Um estudo com alunos do 7º e 12º anos

Andreia Vale & Susana Cunha
Professora Orientadora: Carla Manuela Navio Dias

Valores de Referência

Classificação	IMC	Risco Co-Morbilidades
Baixo peso	<18,5	Baixo
Normal	18,5 – 24,99	Médio
Excesso de peso	≥ 25,00	
Pré-obesidade	25 – 29,99	Aumentado
Obesidade classe I	30 – 34,99	Moderado
Obesidade classe II	35 – 39,99	Elevado
Obesidade classe III	≥ 40	Muito elevado

Tabela 1. Classificação da obesidade (OMS, 2000)

	Aumentado	Muito Aumentado
Homem	≥ 94	≥ 102
Mulher	≥ 80	≥ 88

Tabela 2. Risco de complicações metabólicas associadas ao perímetro abdominal (OMS, 2000)

Resumo

Este estudo pretendeu, a partir da utilização de conceitos matemáticos (estatística), perceber melhor como o conhecimento e a manipulação de indicadores como o índice de massa corporal, o perímetro da cintura, a pressão arterial e o número de batimentos cardíacos por minuto, pode ajudar a um melhor conhecimento de nós próprios, prevenindo problemas de saúde e garantindo uma melhor qualidade de vida.
Para tal, consideramos como amostra os alunos que frequentam o 7º ano e o 12º ano na nossa escola.

Idade	Pressão máxima ou sistólica (mm Hg)	Pressão mínima ou diastólica (mm Hg)
3 meses	80	55
12 meses	90	65
4 anos	110	70
10 anos	120	75
15 anos	130	80
20 anos	135	85
30 anos	145	90
40 anos	150	90
50 anos	160	95
60 anos	165	95
70 anos	170	98
mais de 70 anos	175	100

Tabela 3. Variação da pressão arterial com a idade (Fonte: Enciclopédia da Saúde (2005, p.66))

Resultados

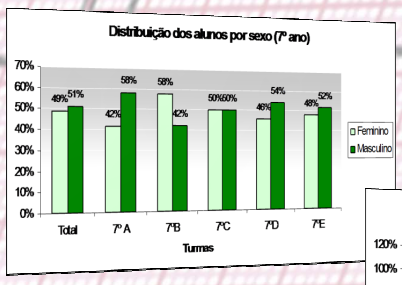


Gráfico 1. Distribuição dos alunos por sexo (7º ano)

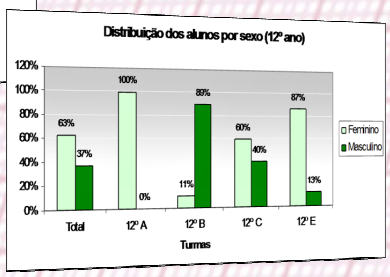


Gráfico 2. Distribuição dos alunos por sexo (12º ano)

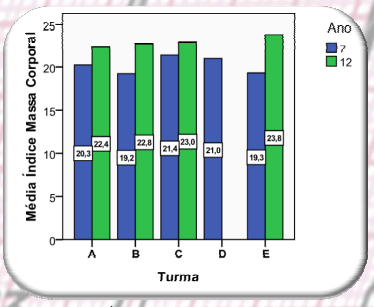


Gráfico 3. Índice de Massa Corporal Médio por turma, segundo Nível de Escolaridade

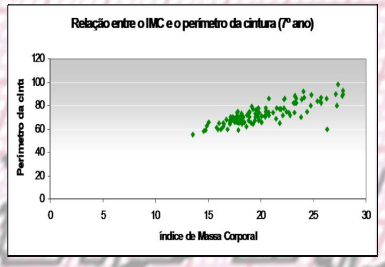


Gráfico 4. Distribuição bidimensional (IMC e Perímetro da Cintura) – 7º ano

Conclusão

O estudo estatístico aqui realizado permitiu aplicar conhecimentos estatísticos ao nível da análise descritiva e correlacional.
Pela análise efectuada, podemos concluir que estes alunos apresentam, em termos médios, um índice de massa corporal (IMC) e um perímetro da cintura normais. Ao nível da pressão arterial máxima e mínima registam valores inferiores aos de referência e em relação ao número de batimentos cardíacos por minuto os valores observados situam-se no intervalo desejável.

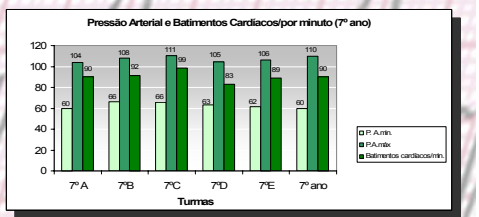


Gráfico 6. Pressão Arterial e Batimentos Cardíacos/por minuto (média por turma) de 7º ano

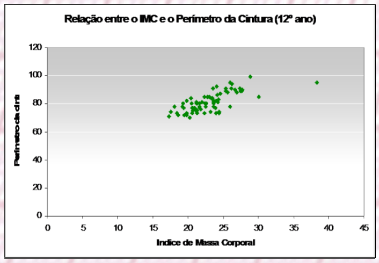


Gráfico 5. Distribuição bidimensional (IMC e Perímetro da Cintura) – 12º ano

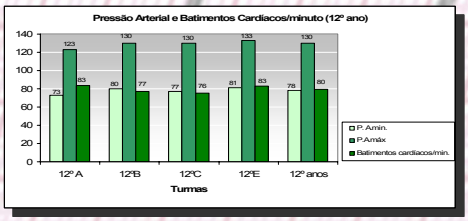


Gráfico 7. Pressão Arterial e Batimentos Cardíacos/por minuto (média por turma) de 12º ano

"ESTATÍSTICOS JÚNIOR 2010"

PRÉMIOS "ESTATÍSTICO JÚNIOR 2011"

Está aberto, até 27 de Maio de 2011, o concurso para atribuição de prémios **"Estatístico Júnior 2011"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2011"** é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio da Porto Editora, e tem como objectivo estimular e desenvolver o interesse dos alunos do ensino básico e secundário pelas áreas da Probabilidade e Estatística.

2. Os candidatos a prémios **"Estatístico Júnior 2011"** devem ser alunos do 3.º Ciclo do Ensino Básico, do Ensino Secundário, dos Cursos de Educação e Formação (CEF), ou dos Cursos de Educação e Formação de Adultos (EFA) no ano lectivo 2010/2011.

3. As candidaturas podem ser individuais ou em **grupo com um máximo de 3 alunos**. Do grupo pode ainda fazer parte um professor do ensino básico ou secundário ao qual caberá o papel de orientador.

4. Os candidatos devem apresentar um trabalho cuja temática deve estar relacionada com a teoria da Probabilidade e/ou Estatística.

5. O trabalho deverá ser constituído por um texto escrito em Português com um máximo de 10 páginas A4 dactilografadas e um poster formato A2 que resuma os principais aspectos do trabalho. O trabalho (poster e texto escrito) deverá ser **enviado impresso em papel para efeitos da avaliação**.

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2011"** a 7 trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.º Ciclo do Ensino Básico, aos três primeiros classificados de entre os trabalhos candidatos do Ensino Secundário, e um primeiro classificado de entre os trabalhos candidatos dos Cursos CEF-EFA. Os prémios são constituídos por produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 600 euros, 300 euros e 200 euros, a atribuir, respectivamente, aos grupos cujos trabalhos sejam classificados em 1.º, 2.º e 3.º lugar para as categorias Ensino Básico e Secundário e 600 euros para a categoria dos Cursos CEF-EFA.

7. Ao professor orientador do trabalho classificado em 1º lugar, em cada categoria, é ainda atribuída uma anuidade grátis como sócio da SPE, ajudas de custo para participação no XVII Congresso Anual da SPE e produtos pedagógicos editados pela Porto Editora (à excepção de manuais escolares) no valor de 500 Euros.

8. Aos grupos proponentes dos trabalhos classificados em 1º lugar será também oferecida uma ampliação do correspondente poster que será colocado na Sessão de Posters do XIX Congresso Anual da SPE.

9. O boletim de candidatura, acompanhado do trabalho concorrente, deverá ser dirigido ao Presidente da SPE para a morada abaixo indicada. O carimbo do correio validará a data de entrega.

Sociedade Portuguesa de Estatística – Bloco C6, Piso 4 – Campo Grande – 1749-016 Lisboa

O boletim de candidatura e este regulamento podem ser obtidos em
http://www.spestatistica.pt/static/docs/BoletimCandidaturaPEJ11.pdf
http://www.spestatistica.pt/static/docs/RegulamentoPEJ11.pdf

10. A admissibilidade e apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição e nomeação será da responsabilidade da Direcção da SPE.

11. O júri é soberano nas decisões, não havendo lugar a impugnação ou recurso.

12. A atribuição dos prémios **"Estatístico Júnior 2011"** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será realizada no XVII Congresso Anual da SPE.

13. Os prémios **"Estatístico Júnior 2011"** poderão não ser atribuídos.

Apoio da Porto Editora

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO SPE 2010

Modelação dos Tempos de Duração de Rotas em Redes Móveis Ad Hoc

Gonçalo Jacinto, *gjcj@uevora.pt*

Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática e CIMA-UE

As redes de telecomunicações móveis têm crescido de forma exponencial nos últimos anos. Para fazer face às novas velocidades e quantidades de tráfego exigidas pelos utilizadores, uma das modernizações mais promissoras das actuais redes de telecomunicações móveis é a sua integração com as redes de telecomunicações móveis ad hoc. Estas redes são caracterizadas por terem nós que comunicam entre si através de ligações sem fios, organizando-se de forma autónoma e sem qualquer infra-estrutura. Apesar dos nós se movimentarem de forma independente, estes cooperam entre si de forma a criarem ligações entre um nó fonte e um nó destino. Quando os nós fonte e destino se encontram a uma distância superior ao seu alcance de transmissão, a comunicação entre ambos é feita através de rotas com múltiplos passos, utilizando os nós vizinhos para encaminhar o tráfego desde o nó fonte até ao nó destino. A mobilidade dos nós e a possibilidade de comunicação por rotas com múltiplos passos torna a topologia destas redes dinâmica e imprevisível, sendo necessário modelar a dinâmica destas rotas e descrever as relações geométricas que governam o movimento aleatório dos nós que compõem uma rota, com o estado de cada ligação limitado pelo alcance de transmissão de cada nó.

Neste trabalho foi desenvolvido um modelo para caracterizar a dinâmica das rotas de comunicação e para derivar o tempo médio de duração dessas rotas. Para tal, a dinâmica das rotas de comunicação é modelada por um processo de Markov determinístico por troços, onde o tempo médio de duração das rotas é obtido como solução única de um sistema de equações integro-diferenciais. Um método recursivo é proposto de forma a tornar possível a sua computação. O modelo proposto considera que as ligações que compõem uma rota são dependentes entre si, conseguindo integrar os requerimentos de mobilidade e conectividade dos nós de uma rede móvel ad hoc, sendo uma extensão dos modelos já existentes que assumem que as diferentes ligações se comportam de forma independente. Resultados numéricos ilustram o cálculo do tempo médio de duração das rotas, os quais são comparados com os obtidos quando se assumem rotas com ligações independentes.

Gonçalo Jacinto, **galardoado com o Prémio SPE 2010**, é Licenciado em Matemática Aplicada pela Universidade de Évora e Mestre em Matemática Aplicada pelo Instituto Superior Técnico da Universidade Técnica de Lisboa.
Está na fase final do Doutoramento em Matemática - aguardando a realização de provas - no Instituto Superior Técnico sob a orientação dos Professores António Pacheco e Nelson Antunes.
É Assistente na Universidade de Évora, Escola de Ciências e Tecnologia, Departamento de Matemática e é membro do CIMA-UE.

Índice