Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral

Primavera de 2010

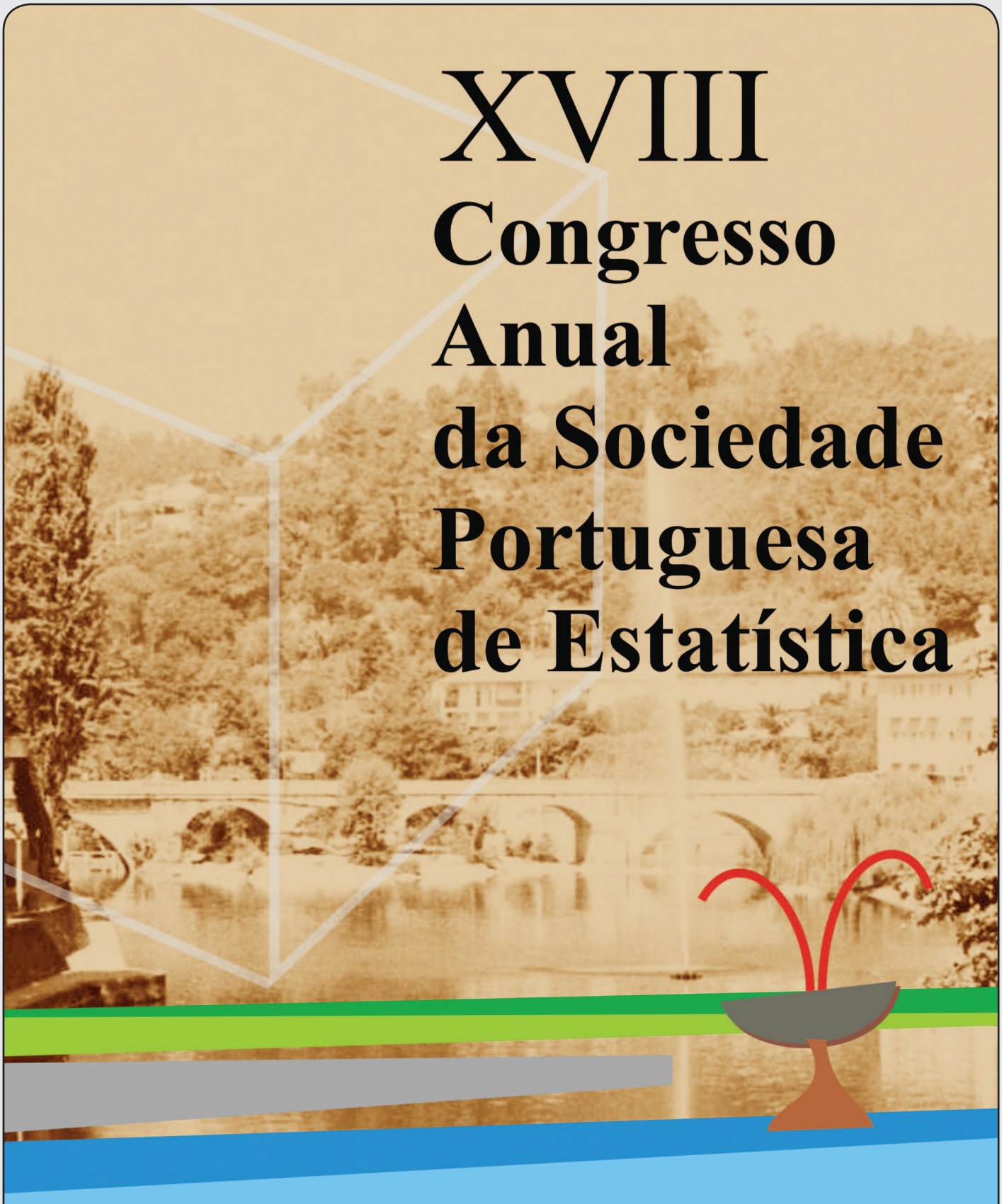
Data Mining - Prospecção (Estatística) de Dados?

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

XVIII Congresso Anual da Sociedade Portuguesa de Estatística



Hotel do Parque
Termas de São Pedro do Sul
29 de Setembro a 2 de Outubro
2010

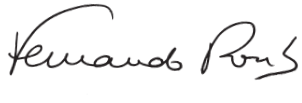
Editorial

… "o futuro da análise de dados?"…

"Data Mining" ou "Statistical Data Mining" são termos que, decerto, estão cada vez mais, a aproximar-se da realidade científica dos estatísticos! De algum modo, revive-se um pouco o ambiente académico que teve o seu apogeu, há cerca de cinquenta anos, com as grandes evoluções informáticas e os consequentes reflexos na evolução da própria ciência Estatística introduzindo novas perspectivas para a teoria e para a prática, através da chamada Análise de Dados.
Entendidas aquelas palavras - na significação mais habitual - como Prospecção de Dados (ou como Mineração em português do Brasil) este é um tema de ponta. Uma das palavras-chave que sempre lhe são associadas são talvez lideradas por "redução de dimensionalidade"; mas, também, "complexidade algorítmica" e "reconhecimento de padrões" são temas afins. Os sistemas de redes neurais também não estão muito longe. Nesta perspectiva a temática em questão é abrangente de um ponto de vista científico, isto é, mantém a tradição estatística de ciência interdisciplinar.
Em 1962, Tukey publicou um histórico artigo sobre o futuro da análise de dados ao qual se seguiu o livro *Exploratory Data Analysis*. Em ambos, podemos verificar novo paradigma para a análise estatística de dados. Passados quase cinquenta anos, simbolicamente, estamos em novo ponto de viragem mas que também é ponto de partida que estimula o avanço científico. Hoje, como ontem, a Prospecção de Dados como uma versão moderna dos desafios da Análise de Dados também se posiciona como um tema que podemos enunciar como contraditório ou talvez mesmo desagregador. Tal como aconteceu com a Análise de Dados também agora a Prospecção de Dados é, ao mesmo tempo, um herói e um vilão. Por um lado, é um tema de sucesso e apelativo porque no limite de diversas áreas científicas nas quais a Estatística consegue penetrar com sucesso para todos os lados - como já é tradição. Mas, por outro lado, os estatísticos clássicos, tal como há quarenta anos atrás sobre a Análise de Dados também de novo e com mais ênfase invocam o abandono ou, pelo menos, o afrouxar das raízes probabilísticas na Estatística.
Como tema recente, a Prospecção de Dados, basicamente, pode ser interpretada como uma extensão da análise exploratória de dados e, na prática, tem os mesmos objectivos. A principal distinção reside na dimensionalidade e no volume dos dados prospectados - examinados minuciosamente e com método. É o caminho científico interdisciplinar mais recente!
Na última década - na ciência da computação, no ambiente científico das bases de dados e dos sistemas de informação - foi iniciado um novo percurso científico, onde os estatísticos são parte importante. É o futuro a continuar a Análise Exploratória de Dados…
Ao abordar a Prospecção de Dados como tema central - com esta edição do Boletim SPE - abre-se um novo campo de partilha entre os estatísticos e a comunidade científica. O espaço editorial do Boletim fica, assim, com um novo domínio aberto à colaboração dos interessados.

Tal como foi anunciado, nesta edição iniciamos uma nova secção: *Pós-doc*. Nela, temos a oportunidade de revelar textos de "novos doutores" cujo doutoramento foi divulgado em anteriores edições do *Boletim SPE* e que, numa consolidação científica, escrevem sobre a sua actualidade. Como editor, agradeço a pronta colaboração dos "primeiros doutores", divulgados no *Boletim SPE Outono de 2006* e que, de imediato, aderiram a esta proposta. Assim, revisitamos o Boletim de há 3 anos que, com esta iniciativa, fica actualizado. Revemos o passado do presente!
É um novo desafio (também) para o formato e inspiração da prosa a utilizar nestas contribuições.
O *Boletim SPE* está mais rico!

O tema central do próximo *Boletim* será *Estatística Espacial.*



Mensagem do Presidente

Caros Colegas:

No momento em que escrevo o tempo está muito pouco primaveril, mas, mais em consonância com o bom trabalho produzido para este Boletim do que com o aspecto plúmbeo que avisto da minha janela, aqui está a edição da Primavera de 2010 do nosso Boletim. Obrigado ao Fernando Rosado e a todos os que contribuíram para este número.

Com algum atraso, de que pedimos desculpa, pusemos a funcionar pela primeira vez o sistema do débito em conta para o pagamento das quotas de 2009 dos sócios que aderiram a esse sistema. Houve necessidade de elaborar e validar os programas informáticos para produzir os ficheiros de processamento e, além disso, houve um atraso (e nesse não tivemos responsabilidade) no registo na SIBS da adesão de alguns sócios, os quais, por isso, só em Janeiro de 2010 é que viram o pagamento da sua quota ser processado. Mas, em 2010, com o sistema já afinado, o processamento far-se-á bem mais cedo e encorajamos os sócios que queiram poupar trabalho a aderirem ao sistema, pois deixam de ter de se preocupar com o pagamento da sua quota.

O nosso XVII Congresso já terá as suas Actas publicadas pelo grupo editorial Springer (na série "Selected Papers in Statistics") no âmbito do acordo estabelecido com este grupo por várias sociedades estatísticas, entre as quais a SPE, acordo que vai contribuir para a difusão internacional da produção científica da nossa comunidade estatística. Isso levou a algum adiamento das datas habituais de submissão dos artigos para se acertarem todas as agulhas. Porventura, os autores não se irão queixar, já que tiveram mais tempo para elaborar os seus artigos, mas, em contrapartida, vamos ter de pedir aos avaliadores que façam o seu importante trabalho num período mais reduzido, o que desde já agradecemos, pedindo a sua compreensão. No próximo ano, com tudo já afinado, já poderemos cumprir o calendário normal.

E, por falar do próximo Congresso, o XVIII, é altura de os colegas iniciarem os preparativos para nele participarem. Lá nos encontraremos, agora num local diferente mas igualmente magnífico, S. Pedro do Sul, e com um programa igualmente aliciante, de 29 de Setembro a 2 de Outubro de 2010, com organização conjunta da Universidade de Coimbra e do Instituto Politécnico de Viseu. A vossa presença nesta celebração anual da Estatística em Portugal é indispensável. A SPE conta convosco.

As actividades da SPE e outras actividades na área da Estatística têm sido noticiadas por e-mail ou através da nossa página *web*. Uma novidade recente que ainda não foi noticiada foi a aprovação pela Comissão Nacional de Matemática de uma proposta que fiz em nome da SPE para que a nossa Sociedade passe a integrar a referida Comissão. A proposta terá, para produzir efeito, de ser aprovada pelo MCTES, que tutela a Comissão Nacional de Matemática. Esperemos que tal venha a suceder, já que é da mais elementar justiça. Agradeço o apoio dos membros da Comissão, e, em particular do seu Presidente, Professor José Francisco Rodrigues, bem como a intervenção da Professora Ivette Gomes, que, estando presente nessa reunião (a representar o CEAUL), gentilmente acedeu a advogar brilhantemente a posição da SPE.

Várias novas iniciativas estão em preparação mas delas lhes daremos notícias quando tivermos informações mais concretas.

Como sabem, a Comissão Especializada de Nomenclatura Estatística, que produziu o glossário estatístico em língua portuguesa (ver páginas *web* da SPE ou do ISI), pediu a vossa ajuda para o alargamento do glossário, contribuindo assim também para o alargamento do próprio glossário internacional do ISI, que está um tanto desactualizado pois não tem reflectido o constante aparecimento de novos termos. É uma iniciativa meritória e muito importante para a nossa actividade. Pede-se, pois, aos sócios que ainda não o fizeram que enviem urgentemente por e-mail os novos termos da sua área de trabalho ao Presidente da Comissão, o nosso Colega Daniel Paulino.

Termino agradecendo a todas as pessoas e entidades que em 2009 apoiaram a SPE e as suas iniciativas. A todos os sócios, e em especial àqueles que exerceram as mais diversas tarefas e funções em prol da SPE, quero agradecer o seu apoio e a sua participação na vida da Sociedade.

Saudações cordiais

Notícias

• Carlos Braumann, Reitor da Universidade de Évora

O Prof. Carlos Braumann foi eleito pelo Conselho Geral como Reitor da Universidade de Évora.
No dia 3 de Fevereiro de 2010, num processo eleitoral que envolveu 7 candidatos admitidos pela Comissão Eleitoral foi proclamado eleito pelo Conselho Geral como Reitor da Universidade de Évora o candidato Carlos Alberto Santos Braumann.
Para além da realização de projectos pessoais, esta eleição é também um acontecimento de júbilo para a Sociedade Portuguesa de Estatística ao registar-se o sucesso naquela eleição de prestígio.
Ao Presidente da SPE e Reitor da Universidade de Évora, formulamos votos do maior sucesso no desempenho das novas funções que lhe são confiadas.

FR

• XVIII Congresso SPE

A Sociedade Portuguesa de Estatística, em colaboração com o Departamento de Matemática da Universidade de Coimbra e a Área Científica de Matemática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Viseu, está a organizar o seu XVIII Congresso Anual, que decorrerá no Hotel do Parque em S. Pedro do Sul, de 29 de Setembro a 2 de Outubro de 2010.

O XVIII Congresso Anual da S.P.E. prevê as sessões habituais de comunicações orais e posters abertos a toda a comunidade estatística de língua portuguesa. Este Congresso Anual pretende dar ênfase especial a dois temas: os métodos não paramétricos em estatística, para o que contribui a apresentação do habitual mini-curso, a cargo de Carlos Tenreiro (CMUC, Universidade de Coimbra) que precede o congresso propriamente dito, e um apelo a contribuições na área da amostragem, dada a aproximação de mais um Censo em Portugal.

No seu programa científico, o XVIII Congresso Anual inclui cinco sessões plenárias para as quais foram convidados os oradores: Alda Carvalho (INE), Enno Mammen (Univ. Mannheim, Alemanha), Mário Figueiredo (Inst. Telecomunicações, I.S.T., Lisboa), Maurizio Vichi, (Univ. La Sapienza, Roma, Itália), Wenceslao González-Manteiga (Univ. Santiago de Compostela, Espanha).
Para as (poucas) horas que o XVIII Congresso Anual deixa livres, S. Pedro do Sul, estando situada em plena Região de Lafões, oferece muitos outros atractivos que começam nas Termas e passam pela excelente gastronomia e vinhos. Para actividades mais saudáveis recomendam-se a visita às aldeias serranas que circundam S. Pedro do Sul.

Informação detalhada sobre o XVIII Congresso Anual da S.P.E. – SPE2010 está disponível na página http://www.mat.uc.pt/~spe2010. Chamamos a atenção para a antecipação da data limite para a submissão dos resumos que é 11 de Abril de 2010.

A Comissão Organizadora Local

• Biometrical Colloquium and Polish-Portuguese Workshop em Biometry

O **40th International Biometrical Colloquium and Second Polish-Portuguese Workshop on Biometry** irá decorrer de 31 de Agosto a 2 de Setembro de 2010, em Bedlewo – Poland (http://www.impan.pl/Bedlewo/), em homenagem ao Professor Doutor João Tiago Mexia.

O objectivo é reunir investigadores que partilhem um interesse comum nas áreas da Estatísticas e Aplicações. São esperadas e bem vindas contribuições em Planeamento de Experiências, Modelos Lineares e Não-lineares, Inferência Estatística, Análise Multivariada, Bioestatística e Biometria, entre outras. Será uma perfeita oportunidade para apresentar e discutir desenvolvimentos recentes nestas matérias e para estabelecer contactos que levarão a frutíferas colaborações.

Teresa Oliveira

• Quarto Workshop em Estatística, Matemática e Computação

O Quarto Workshop em Estatística, Matemática e Computação realizou-se em Novembro de 2009, nos dias 9 e 10, nas instalações da Fundação Calouste Gulbenkian, tendo sido prestada na Sessão Inaugural uma homenagem ao Professor Doutor João Tiago Mexia, que se jubilou em 2009.

Neste encontro reunimos ilustres investigadores, académicos e profissionais, com o objectivo de partilhar resultados de investigação e experiências práticas, com ênfase em áreas como a Bioestatística, a Biometria, a Biomedicina e a Biomatemática. Foram focados muitos avanços na teoria e aplicações, evidenciando os desenvolvimentos a nível Computacional.

Do programa constaram 17 Sessões Convidadas, 7 Comunicações Livres e 2 Sessões de Posters, contando com cerca de 100 de participantes dos quais 16 eram estrangeiros.

O encontro teve o patrocínio da FCT- Fundação para a Ciência e a Tecnologia, da Universidade Aberta, do CMA, da PSE , da Espiral Tours e da Fundação Calouste Gulbenkian.

Teresa Oliveira

• I Dia Mundial da Estatística

A Comissão de Estatística das Nações Unidas (UNSD) declarou o dia 20 de Outubro de 2010 como o primeiro **Dia Mundial da Estatística**.

A UNSD propôs que esse dia foque três sub-temas: o serviço às nações e ao mundo, o profissionalismo e a integridade. Especialmente dirigido às organizações nacionais produtoras de estatísticas, o Dia Mundial da Estatística - como proposto pela UNSD - deve ser um reconhecimento pelos serviços prestados pelo sistema estatístico global, tanto a nível nacional como internacional. É, igualmente, uma proposta de apoio e de reconhecimento pelo trabalho dos estatísticos nos diversos ambientes culturais e domínios onde desempenham funções.

Estão pensadas actividades que envolvem a promoção daquele dia e que vão desde a criação de um logótipo ou a introdução de documentos via YouTube ou Twitter até à nomeação de um Embaixador de Boa Vontade para este evento.

Mais informações em http://unstats.un.org/unsd/ onde, em particular, se pode verificar o Dia Nacional Português da Estatística.

FR

• ”Ensino da Estatística” em homenagem a Eugénia Graça Martins

Na Faculdade de Ciências da Universidade de Lisboa, organizado pelo Departamento de Estatística e Investigação Operacional, no passado dia 27 de Janeiro, decorreu um encontro sobre “O Ensino da Estatística do Básico ao Secundário”. Este evento também foi uma homenagem à Professora Doutora Maria Eugénia Graça Martins, “pelo muito que ela fez e continua a fazer pelo Ensino da Estatística. O nome da Eugénia está associado a muito de bom que se tem feito em Portugal nesta área. Desempenhou e desempenha um papel preponderante no projecto ALEA. Escreveu vários livros didácticos, para alunos do ensino superior e textos de apoio a professores. São muitas as suas publicações relacionadas com o ensino da estatística, em actas de congressos e em revistas de educação. Foram muitos os cursos de formação para professores que promoveu e leccionou. É constantemente requisitada para dar pareceres, fazer parte de comissões com o Ministério da Educação. Tem dedicado muito do seu tempo a fazer a revisão da parte de probabilidades e estatística de livros de Matemática para o ensino”.

No programa, organizado na sequência de um anterior debate sobre “O Ensino da Estatística a Nível Superior”, foram incluídos diversos oradores convidados.

Em dia de homenagem, falou-se “do tempo a passar” e dos diversos caminhos para percorrer; muitas vezes através “das escadas que se sobem e se descem para acompanhar e ajudar aqueles que as desejam subir”. “Torceram-se” números para os mais pequeninos; mas sempre com objectivos “de optimização”. Viajou-se pelo premiado projecto ALEA. E para “domesticar o acaso” olhou-se bem para o futuro, com a apreciação da “Organização e Tratamentos de Dados” - novos programas, caminhos, desafios e “tarefas” para o Ensino Básico.

Foi também, um dia de convívio muito agradável para o numeroso grupo de participantes nesta iniciativa!

FR

Data Mining - Prospecção (Estatística) de Dados?

Avaliação de Agrupamentos – Contribuições em Data Mining

Margarida G. M. S. Cardoso, *margarida.cardoso@iscte.pt*

Departamento de Métodos Quantitativos, Escola de Gestão, ISCTE-IUL

Análise de agrupamento (Clustering analysis)

O objectivo da análise de agrupamento é a constituição de grupos que integram entidades homogéneas e distintas das de outros grupos. Trata-se, geralmente, de uma análise não supervisionada (de interdependência, na terminologia habitual da Estatística Multivariada) sendo a estrutura alvo de aprendizagem – agrupamento – *a priori* desconhecida.

Algumas das decisões básicas na análise de agrupamento referem-se a: 1) Selecção de dados i.e. entidades a agrupar e variáveis base de agrupamento; 2) Pré-avaliação de existência de uma estrutura/agrupamento; 3) Escolha de uma função objectivo e metodologia de agrupamento a adoptar; 4) Determinação do número de grupos.

Após a constituição dos grupos há que avaliar a sua qualidade. No processo de agrupamento o alvo é traduzido por uma função objectivo (f.o.) que tipicamente exprime homogeneidade intra-grupos e/ou heterogeneidade entre-grupos (a constituir). Assim, numa primeira avaliação de agrupamentos, a bondade da solução obtida será referida à f.o. considerada.
No entanto, nem sempre o algoritmo de agrupamento garante um óptimo para a f.o., o que acontece em processos iterativos *gulosos* como o K-Médias (MacQueen, 1967). Sendo garantido um óptimo, este pode não ser global: é o caso de versões diversas do algoritmo EM-Expectation Maximization (Dempster et al., 1977), e.g. (Vermunt e Magidson, 2002), que procuram maximizar a função de verosimilhança associada a um modelo específico de mistura finita. Por esta razão, é comum utilizar instrumentos complementares de avaliação de resultados em agrupamento: um agrupamento é então avaliado com base em indicadores de propriedades desejáveis do mesmo, quer do ponto de vista da análise de dados, quer do ponto de vista substantivo (conhecimento do domínio de aplicação).

Este trabalho refere-se a avaliação de partições simples resultantes de agrupamento (note-se que, na prática, estruturas difusas ou probabilísticas são habitualmente convertidas em partições simples mediante afectação modal).

Paradigmas de análise

Diferentes paradigmas científicos podem conduzir a diferentes propostas na avaliação de resultados de agrupamento (mediante análise de dados): *The computer science data mining paradigm drastically differs from the statistical data analysis paradigm* (Mirkin, 1998). As Figura 1 e Figura 2 ilustram estes dois paradigmas.

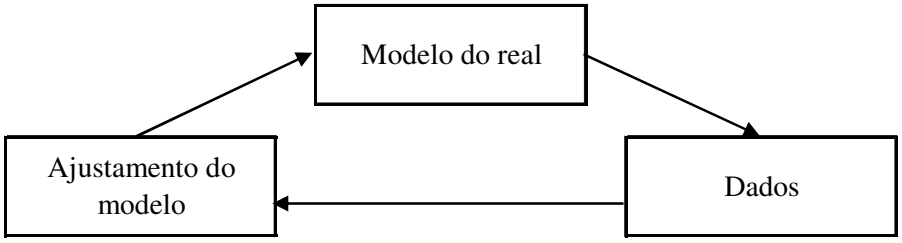


Figura 1 - Análise de Dados: Paradigma Tradicional da Estatística (Mirkin, 1998)

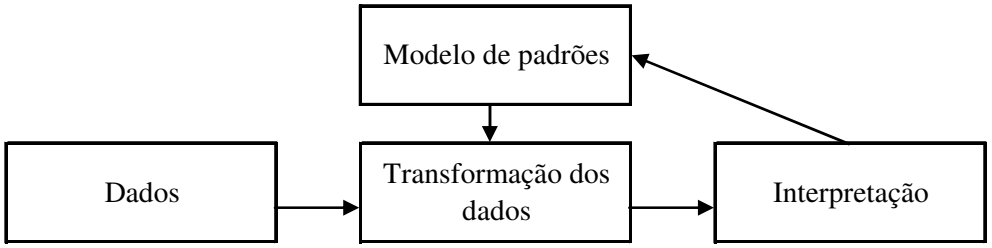


Figura 2- Análise de Dados: Paradigma das Ciências de Computação (Mirkin, 1998)

No domínio da Estatística é comum postular um modelo que se refere à população. Pretende-se, então, ir ao encontro da suposta *verdadeira estrutura* e não somente descobrir uma estrutura compatível com os dados. Nesta linha, uma contribuição específica da Estatística é a realização de testes de hipóteses para aferir a significância de modelos - aferir se uma população é homogénea sendo a sua estrutura de agrupamento simplesmente inexistente, unimodal, por exemplo. Naturalmente, nesse caso, é exigido o cumprimento de alguns pressupostos. Um dos pressupostos mais commumente usado na inferência estatística é o da distribuição normal de uma variável acerca da qual se pretende inferir. Nem sempre, todavia, (mesmo depois de efectuados alguns ensaios de transformação dos atributos) é razoável admitir essa normalidade.

Os processos de aprendizagem de modelos em Data Mining constituem, geralmente, uma abordagem isenta de pressupostos. Associado à proposta de um modelo - por exemplo, uma rede SOM-Self Organizing Maps, (Kohonen, 1995), para agrupamento - há, no entanto, que considerar o *custo* da dimensão da amostra que viabiliza a indução procurando generalizar, para uma população, relações verificadas numa amostra particular. Além disso, ao prescindir de pressupostos habituais no domínio da Estatística, os meios de quantificação da incerteza associada à indução são empíricos. Naturalmente que esta questão não se coloca se a aprendizagem fôr referida à própria população.

No domínio do agrupamento as abordagens são maioritariamente de natureza descritiva pelo que se pode dizer que na avaliação dos seus resultados predomina o paradigma do Data Mining. Esta avaliação concretiza-se, nomeadamente, na verificação de algumas propriedades de bons resultados de agrupamento.

Propriedades de bons resultados de agrupamento

Compacidade, separabilidade e índices de qualidade

A compacidade e separabilidade são propriedades desejáveis num agrupamento, e.g. (Everitt et al., 2009): a compacidade mede a coesão interna das entidades reunidas num mesmo grupo; a separabilidade mede o isolamento de um grupo quando comparado com os outros.

Embora a ideia de compacidade-separabilidade seja geralmente aceite, têm sido propostas várias definições para estas propriedades traduzidas na construção de múltiplos índices de qualidade (IQ), habitualmente quocientes entre uma medida de compacidade (intra-grupos) e uma medida de separabilidade (entre-grupos).

Alguns IQ foram originalmente propostos como critério de paragem para avaliar o número adequado de grupos, e.g. (Davies e Bouldin, 1979). Contudo, todos se propõem medir a qualidade de partições, podendo ser utilizados para a selecção de uma entre várias partições candidatas.

A Tabela 1 apresenta uma lista de IQ. A lista não é exaustiva, mas contém alguns índices muito populares e ilustra bem a diversidade de propostas. O índice de Calinski e Harabask, por exemplo, é uma pseudo estatística F que quantifica a razão entre as distâncias totais entre e intra-grupos.

Alguns IQ foram generalizados de modo a avaliarem estruturas difusas, caso dos índices de Dunn e PBM generalizados - (Bezdek e Pal, 1998) e (Pakhira et al., 2004), respectivamente.

Na prática de agrupamento há ainda a referir que os IQ são, por vezes, considerados como f.o. de um processo de agrupamento: (Hruschka et al., 2006) consideram o Índice Silhouette como f.o., por exemplo. Por outro lado, funções tradicionalmente consideradas como objectivo podem sugerir um novo IQ: (Fred e Jain, 2008) sugerem um índice baseado no critério *Minimum Description Lenght*, por exemplo.

Tabela 1- Alguns índices de qualidade

Hubert's Γ statistic (Hubert e Schulz, 1976)
Dunn (Dunn, 1974)
Calinski e Harabasz (Calinski e Harabasz, 1974)
Hartigan (Hartigan, 1975)
Davies e Bouldin (Davies e Bouldin, 1979)
Silhouette (Rousseeuw, 1987), (Kaufman e Rousseeuw, 1990)
PBM (Pakhira et al., 2004)

Uma vez seleccionado um particular IQ coloca-se a questão de saber que valores limite considerar para que uma partição seja considerada um bom agrupamento. De facto, *it is easy to propose indices of cluster validity. It is very difficult to fix thresholds on such indices that define when the index is large or small enough to be "unusual" or valid…* (Jain e Dubes, 1988).

A estratégia mais comum para abordar este problema é considerar limiares relativos adoptando o seguinte procedimento genérico:

*Para l=1...L (*L pode referir-se ao número de parametrizações alternativas de um algoritmo de agrupamento ou ao nº de diferentes algoritmos considerados*)*
    Obter agrupamento A(l) sobre amostra original
    Calcular IQ[A(l)]
Seleccionar agrupamento Argumento-Melhor (IQ)

No caso particular em que se procura avaliar o melhor número de grupos, uma representação gráfica do número de grupos vs IQ permite identificar um bom compromisso qualidade-complexidade através do *cotovelo* no gráfico. Alguma dificuldade de identificar visualmente o referido *cotovelo - statistical folklore* segundo (Tibshirani et al., 2001) - poderá eventualmente ser contornada com uma metodologia apropriada, nomeadamente o *L-Method* (Salvador e Chan, 2004).

Para lidar com a necessidade de estabelecer limiares pode também recorrer-se ao voto de diversos IQ (Bolshakova e Azuaje, 2003).

Outra possibilidade é confiar nos autores que indicam, eles próprios, limiares para os índices – limiares com autoria - derivados a partir de estudos empíricos. Por exemplo:

- (Hartigan, 1975) sugere uma *crude rule of thumb* – índice de Hartigan superior a 10 - que poderá justificar o aumento do número de grupos de K para K+1;
- (Kaufman e Rousseeuw, 1990) apresentam uma tabela de avaliacão do índice Silhouette resultado de *experience with the silhouette index which has led us to a rather subjective interpretation*…(p. 88).

Uma estratégia alternativa (computacionalmente mais pesada) para a determinação de valores limite para os IQ pode basear-se num modelo que exprime a ausência de estrutura de agrupamento na população – limiares sob hipótese de homogeneidade. Um modelo uniforme ou um modelo unimodal podem, por exemplo, ser utilizados para este fim, (Gordon, 1999), (Bock, 1996). Esta é a proposta de (Cardoso e Carvalho, 2009).

A estatística Gap (Tibshirani et al., 2001) pode ser considerada um primeiro passo neste sentido. Ela lida directamente com a questão do limiar, incorporado-o no próprio IQ. Este valor limite, associado a um número específico de grupos (K), refere-se à média das distâncias intra-grupos correspondendo a M amostras geradas sob a hipótese de homogeneidade (M=20 pode ser usado na prática). Assim, a estatística Gap é a diferença entre esta média e a distância intra-grupos associada à partição que se pretende avaliar.

(Cardoso e Carvalho, 2009) propõem determinar limiares para qualquer IQ utilizando a distribuição empírica de IQ referida a várias amostras geradas sob hipótese de homogeneidade (H_0). O procedimento proposto é o seguinte:

Definir H_0
Para m=1...M
    Gerar amostra aleatória – m - sob H_0
    Obter agrupamento sobre amostra m: A(m)
    Calcular IQ[A(m)]
Obter medidas descritivas associadas à distribuição de IQ: seleccionar valor IQ de referência
Para l=1...L
    Obter agrupamento A(l) sobre amostra original
    Calcular IQ[A(l)]
    Comparar IQ[A(l)] com valor IQ de referência
Seleccionar A(l) correspondente a comparação mais favorável com IQ de referência*

Em experiências realizadas com dados reais, conclui-se que a estratégia proposta tende a concordar com a consideração de limiares relativos na selecção de resultados de agrupamento.

Estabilidade e índices de concordância

A estabilidade é também reconhecida como uma propriedade desejável de uma solução de Agrupamento, e.g. (Mirkin, 1996). Neste sentido, uma solução deve manter-se quando o processo de agrupamento for sujeito a pequenas alterações, tais como parametrizações alternativas do algoritmo utilizado, introdução de ruído nos dados, consideração de variáveis base alternativas e consideração de diferentes amostras.

Na avaliação da estabilidade de um agrupamento que decorre, em particular, da comparação de agrupamentos resultantes de diferentes amostras, pode usar-se um procedimento de validação cruzada, (McIntyre e Blashfield, 1980), (Breckenridge, 1989) (Tabela 2). Trata-se da importação de uma metodologia muito comum na análise supervisionada para o âmbito da análise não supervisionada.

Tabela 2 – Procedimento de validação cruzada para agrupamento

Etapa	Acção	Resultado
1	Particionar amostra original	Amostras de treino e de teste
2	Agrupar amostra de treino	Grupos na amostra de treino
3	Construir um Classificador supervisionado por grupos na amostra de treino. Usar o Classificador na amostra de teste	Classes na amostra de teste
4	Agrupar amostra de teste	Grupos na amostra de teste
5	Calcular índices de concordância entre grupos e classes obtidos sobre amostra de teste.	Valor de referência para avaliação de estabilidade

No final do procedimento de validação cruzada o valor do índice de concordância (IC) entre os grupos (obtidos mediante Agrupamento, na etapa 4) e classes (resultantes de Análise Classificatória/Discriminante, na etapa 3), obtidos na amostra de teste, é usado como indicador de estabilidade.

Na literatura encontra-se múltiplos IC. Na Tabela 3 propõe-se uma tipologia para estes índices (v. também (Cardoso, 2007)). O índice de concordância percentual é, simplesmente, a proporção de observações que duas partições concordam em reunir num mesmo grupo.

Tabela 3 – Alguns índices de concordância (IC)

Tipo de IC	Exemplos
IC simples	Índice de concordância percentual*
	Coeficiente Kappa (Cohen, 1960) *
	Estatística Qui-Quadrado ((Cooper e Weeks, 1983), por exemplo)
	Estatística V de Cramer ((Cooper e Weeks, 1983), por exemplo)
IC pareada	Índice de Jaccard (Jaccard, 1908)
	Índice de Fawlkes e Mallows (Fowlkes e Mallows, 1983)
	Índice de Rand (Rand, 1971)
	Índice de Rand Ajustado (Hubert e Arabie, 1985)
IC de informação mútua	Informação Mútua Normalizada (Strehl e Gohosh, 2002)
	Variação de Informação (Meila, 2007)

*considerando duas partições com o mesmo número de grupos

Em (Martins e Cardoso, 2009) apresenta-se um exemplo de aplicação de validação cruzada aplicado à segmentação de clientes de cartões de crédito. Nas etapas 2 e 4 da validação cruzada utiliza-se o algoritmo *Two-Step* (Chiu et al., 2001) para agrupamento. Este algoritmo apresenta um bom desempenho em conjuntos de dados de grande dimensão e permite a utilização de variáveis mistas (quantitativas e qualitativas), usando uma função de distância apropriada. Na etapa 3 usa-se o algoritmo CART – *Classification e Regression Trees,* (Breiman et al., 1984), para a construção do classificador supervisionado pelos segmentos construídos na amostra de treino e resultantes da aplicação do algoritmo *Two-Step*. Algumas vantagens deste algoritmo são: apresentar bom desempenho em conjuntos de dados de grande dimensão; ser aplicável a qualquer tipo de dados e independente da escala das variáveis; proporcionar resultados fáceis de interpretar, nomeadamente ao nível das regras de negócio. O índide de Rand ajustado por (Hubert e Arabie, 1985) é usado na etapa 5, um índice que de acordo com um estudo comparativo de (Milligan e Cooper, 1986) tem particulares vantagens para avaliar agrupamentos.

O procedimento de validação cruzada (rever Tabela 2) levanta algumas questões:

1) Seleccionar um classificador adequado para exportar o agrupamento do treino para o teste já que, de acordo com (Lange et al., 2004): *by selecting an inappropriate classifier one can artificially increase the discrepancy between solutions* (p. 1304). *…the identification of optimal classifiers by analytical means seems unattainable. Therefore we have to resort to potentially suboptimal classifiers in practical applications* (p.1305).

2) Dispor de uma amostra original com dimensão suficiente para viabilizar a constituição de várias subamostras de treino e teste já que, na prática, é habitual replicar a validação cruzada, usando diversas partições treino-teste alternativas, de modo a melhor avaliar a estabilidade de uma solução, (Tibshirani et al., 2001), (Levine e Domany, 2001), (Dudoit e Fridlyand, 2002), (Law e Jain, 2003), (Lange et al., 2004), (Cardoso, 2007).

A utilização de um procedimento de estimação de um modelo de mistura finita para agrupamento tem a vantagem de evitar a questão **1)**, já que permite não só constituir os grupos, mas também obter um classificador que resulta da própria estimação do modelo e que pode ser utilizado sobre uma amostra de teste (Cardoso, 2007).

Relativamente à questão **2)** (Cardoso et al., forthcoming) fazem uma proposta de metodologia de validação cruzada que radica no uso de uma amostra ponderada. Neste procedimento, a dimensão da amostra deixa de ser uma limitação relevante para a implementação da validação cruzada já que os IC são baseados na amostra global ponderada e não numa amostra de teste. Para além disso, o uso da amostra ponderada imita a constituição de subamostras aleatórias de treino e teste e deixa de haver a necessidade de construir um classificador, pelo que a questão **1)** também não se coloca.

Conclusões e perspectivas

Na avaliação de resultados de agrupamento – partições simples, em particular – são habitualmente utilizados indices de qualidade (IQ) para avaliar a compacidade-separabilidade e índices de concordância (IC) para avaliar a estabilidade (concordância com outras partições resultantes de pequenas modificações no processo de agrupamento).

No estudo dos IQ foi analisada a questão dos limiares que permitem a identificação de uma boa solução de agrupamento, nomeadamente a determinação destes limiares sob hipótese de homogeneidade da população. A determinação de valores limite é também objecto de interesse no que respeita aos IC. Esta é uma questão já abordada implicitamente na construção de alguns IC simples e pareada (Warrens, 2008), mas a merecer mais atenção no futuro. Neste caso, o modelo *nulo* a considerar refere-se à independência das duas partições que se comparam.

Ao determinar os grupos mais compactos e bem separados ou agrupamentos mais estáveis não se identifica, necessariamente, o agrupamento que mais concorda com a estrutura real existente na população - (Brun et al., 2007), (Ben-David e Luxburg, 2008), (Cardoso et al., forthcoming), por exemplo. Neste sentido, a instabilidade observada num resultado de agrupamento ou a falta de compacidade-separabilidade dos grupos poderão ser úteis na eliminação de partições candidatas, mas

não necessariamente na selecção da partição real. Uma questão que requer mais investimento futuro é a da relação entre a concordância com a estrutura real e as propriedades de um agrupamento, supondo que faz sentido referir tal estrutura... De qualquer modo será consensual admitir que podem existir diversos agrupamentos reais revelados consoante o processo de análise utilizado.

Finalmente, convém não esquecer que *Cluster analysis is a very practical subject,* (Kaufman e Rousseeuw, 1990), na primeira linha do seu livro. Assim, em qualquer aplicação, o analista deverá recorrer a conhecimento especializado (no domínio específico) para avaliar um agrupamento. Por exemplo, em Marketing é geralmente pertinente: identificar os segmentos de modo a torná-los acessíveis; analisar modos de resposta diferenciados por parte dos segmentos; estimar o valor dos segmentos; avaliar a estabilidade dos segmentos no tempo. E cabe, ainda, referir que a interpretabilidade de uma solução de agrupamento é, em última análise, a condição da sua utilidade: *Although validation e interpretation are not coincident there are many common features to allow thinking of them as quite intermixed: for instance finding a good interpretation is a part of validation; conversely, if the clusters are invalid, the interpretation seems unnecessary* (p. 160, (Mirkin, 1998)).

Referências

S. Ben-David and U. V. Luxburg. Relating clustering stability to properties of cluster boundaries. *21st Annual Conference on Learning Theory (COLT)*, R. Servedio and T. Zhang, Springer, 2008.

J. C. Bezdek and N. R. Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man & Cybernetics: Part B*, 28 (3): 301-315, 1998.

H. H. Bock. Significance tests in cluster analysis. In: P. Arabie, L. J. Hubert and G. D. Soete (Ed.). *Clustering and Classification*, World Scientific Publishers, 1996.

N. Bolshakova and F. Azuaje. Cluster validation techniques for genome expression data. *Signal Processing*, 83: 825-833, 2003.

J. Breckenridge. Replicating cluster analysis: method, consistency and validity. *Multivariate Behavioral Research*, 24: 147-161, 1989.

L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. *Classification and Regression trees*. Belmont, California, Wadsworth, 1984.

M. Brun, C. Sima, J. Hua, J. Lowey, B. Carrol, E. Suh and E. R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern recognition*, 40: 807-824, 2007.

Calinski and Harabasz. A dendrit method for cluster analysis. *Communications in Statistics*, 3: 1-27, 1974.

M. G. M. S. Cardoso. Clustering and cross-validation. *IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction*, C. L. C. Ferreira, G. Saporta and M. Souto-de-Miranda, 2007.

M. G. M. S. Cardoso and A. P. d. L. F. Carvalho. Quality indices for (practical) clustering evaluation. *Intelligent Data Analysis*, 13 (5): 725-740, 2009.

M. G. M. S. Cardoso, A. P. L. F. Carvalho and K. Faceli. Evaluation of clustering results: the trade-off bias-variability. *Classification as a Tool for Research.11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation*, H. Locarek-Junge and C. Weihs, Springer, forthcoming.

T. Chiu, D. P. Fang, J. Chen, Y. Wang and C. Jeris. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2001.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20: 37-46, 1960.

R. A. Cooper and A. J. Weeks. *Data Models and Statistical Analysis*, Philip Allan Publishers Limited, 1983.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 (2): 224-227, 1979.

A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society. Series B (Methodological)*, 39 (1): 1-38, 1977.

S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a data set. *Genome Biology*, 3 (7): 0036.1-0036.21, 2002.

J. C. Dunn. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4: 95-104, 1974.

B. Everitt, S. Landau and L. M. *Cluster Analysis*. NY, Wiley, 2009.

E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings (with comments and rejoinder). *Journal of the American Statistical Association*, 78: 553-584, 1983.

A. Fred and A. K. Jain. Cluster validation using a probabilistic attributed graph. *ICPR 2008. 19th International Conference on Pattern Recognition*, IEEE, 2008.

L. Hubert and J. Schulz. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical Psychology*, 29: 190-241, 1976.

Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societé Vaudoise de Sciences Naturelles*, 44: 223-370, 1908.

A. K. Jain and R. C. Dubes. *Algorithms for clustering data*, Englewood Cliffs, N.J.: Prentice Hall, 1988.

L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. NY, Wiley, 1990.

T. Kohonen. *Self-Organizing Maps*, Springer-Verlag, 1995.

T. Lange, V. Roth, M. L. Braun and J. M. Buchman. Stability based validation of clustering solutions. *Neural Computation*, 16: 1299-1323, 2004.

M. H. Law and A. K. Jain. Cluster validity by bootstrapping partitions, Department of Computer Science and Engineering. Michigan State University, 2003.

E. Levine and E. Domany. Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13: 2573-2593, 2001.

J. MacQueen. Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. C. J. Neyman, University of California Press, 1967.

C. Martins and M. G. M. S. Cardoso. Evaluation of clusters of credit cards holders. *Revista de Ciências da Computação (Universidade Aberta)*, 3: 1-11, 2009.

R. M. McIntyre and R. K. Blashfield. A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research*, 2: 225-238, 1980.

M. Meila. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98 (5): 873-895, 2007.

G. W. Milligan and M. C. Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21: 441-458, 1986.

B. Mirkin. *Mathematical Classification and Clustering*. Dordrecht/ Boston/ London, Kluwer Academic Publishers, 1996.

B. Mirkin. Data analysis and classification: an overview. *International Summer School on Knowledge Discovery in Databases and Data Mining: Methods and Applications*, 1998.

B. Mirkin. Data Analysis and Classification. *International Summer School on KDD and DM: Methods and Applications - EAIA-98 of APPIA*, 1998.

M. K. Pakhira, S. Bandyopadhyay and U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37 (3): 487–501, 2004.

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66: 846-850, 1971.

Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53-65, 1987.

S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *16th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2004.

A. Strehl and J. Gohosh. Cluster ensembles - a knowledge reuse framework for combining partitions. *Journal of Machine Learning Research*, 3: 583-617, 2002.

R. Tibshirani, G. Walther, D. Botstein and P. Brown. Cluster validation by prediction strength, Department of Statistics, Stanford University, 2001.

R. Tibshirani, G. Walther and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 32 (2): 411-423, 2001.

J. K. Vermunt and J. Magidson. Latent class cluster analysis. In: J.A.Hagenaars; and A. L. McCutcheon (Ed.). *Applied Latent Class Analysis*. Cambridge, Cambridge University Press, 89-106, 2002.

Modelos de Previsão de Valores Extremos e Raros

Luís Torgo e Rita Ribeiro
ltorgo@dcc.fc.up.pt, rpribeiro@dcc.fc.up.pt

LIAAD / Inesc Porto, LA - Faculdade de Ciências
Universidade do Porto

1 Introdução

A previsão de eventos raros é de crucial importância em diversas áreas, incluindo finanças, metereologia, ecologia, etc., para citar algumas. Frequentemente estes fenómenos raros estão associados aos valores de uma variável contínua. Mais concretamente, os tais eventos raros são por diversas vezes associados a valores inexperadamente extremos das referidas variáveis. Neste contexto, muitas destas aplicações beneficiariam de modelos que fossem capazes de prever com eficácia valores raros e extremos desta variável contínua. Estamos portanto em face de problemas de regressão múltipla, uma vez que na maioria das vezes o "estado actual" do sistema para o qual pretendemos previsões, é descrito por uma série de variáveis independentes. Todavia, este é um tipo diferente de problemas de regressão em que, contrariamente ao que é mais comum, não pretendemos obter os parâmetros de um modelo optimizando um qualquer critério de erro médio. De facto, os utilizadores destas aplicações estão essencialmente interessados na performance predictiva dos modelos nas tais situações raras, sendo mais os ou menos irrelevante a sua performance nos casos comuns. Isto acontece porque tipicamente este tipo de eventos raros exigem a tomada de acções (preventivas, correctivas, ou outras), que têm custos e benefícios a elas associadas. Pelo contrário as situações normais não obrigam a qualquer acção não trazendo portanto qualquer custo ou benefício. Em resumo, a classe de aplicações que são abordadas neste artigo envolvem a previsão de uma variável contínua com custos/benefícios diferenciados no seu domínio de valores, sendo que os seus valores extremos e raros são os que de facto importam prever de modo eficaz.

Em data mining (como noutras áreas) existe uma grande quantidade de trabalhos sobre o desenvolvimento de modelos de previsão em situações com custos diferenciados. Todavia, a grande maioria desses trabalhos aborda problemas de previsão de variáveis nominais (i.e. problemas de análise discriminante), e não problemas de previsão numérica. De facto, os problemas de análise discriminante envolvendo custos e envolvendo classes altamente desbalanceadas, são tópicos recorrentes na literatura de data mining (e.g. [3, 4]). Já relativamente ao problema de previsão de valores raros e extremos de variáveis numéricas os trabalhos são escassos embora exista já algum trabalho anterior nesta área (e.g. [7, 5, 8]). Este artigo descreve uma proposta de critério de avaliação de modelos cujo objectivo é prever este tipo de valores. O critério de avaliação é um dos pontos centrais ao desenvolvimento de qualquer técnica de modelação. De facto, a construção de um modelo de previsão pode ser vista como um problema de procura pelos valores dos parâmetros do mesmo que optimizem um dado critério de preferência/avaliação. Assim, o estabelecimento de tais critérios é o primeiro passo para o desenvolvimento de técnicas de modelação que sejam optimizadas para a performance predictiva em valores extremos e raros de uma variável objectivo contínua.

A previsão de eventos raros, nomeadamente num contexto de análise discriminante, é frequentemente avaliada através das medidas de *precision* e *recall*. A principal vantagem destas medidas centra-se no facto de focarem a avaliação dos modelos na sua performance nos ditos eventos, ignorando a sua performance nas restantes classes do problema. A sua definição é facilmente descrita

usando um problema exemplo com duas classes possíveis para a variável objectivo, que iremos denominar "Pos" e "Neg", sendo que "Pos" é a classe associada aos eventos raros que pretendemos prever eficazmente, e que normalmente é muito menos frequente do que as outras classes. Dado tal cenário, a performance predictiva de uma qualquer modelo discriminante pode ser descrita pela matriz de confusão apresentada na Tabela 1.

Tabela 1: Uma matriz de confusão para duas classes.

	Predicted Pos	Predicted Neg	
Pos	TP	FN	POS
Neg	FP	TN	NEG
	PPOS	PNEG	

Neste contexto, o valor das duas estatísticas mencionadas anteriormente é obtido por:

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{PPOS} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{POS} \tag{2}$$

O objectivo central deste artigo é o de propor estatísticas equivalentes a estas para problemas de previsão numérica.

2 Formalização do Problema

Nos problemas de previsão o objectivo geral é o de obter um modelo de uma função desconhecida que relaciona uma variável objectivo (variável dependente) com um conjunto de predictores (variáveis independentes). Este modelo é normalmente obtido graças a uma amostra de casos exemplo do relacionamento entre estas variáveis. Esta amostra é usada para obter os parâmetros do modelo que minimizam um qualquer critério de erro de previsão. Normalmente em problemas de regressão múltipla este critério é o critério dos mínimos quadrados, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, ou o critério dos desvios absolutos médios, $MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. Ambos os critérios são estimadores do erro médio, quadrado ou absoluto, respectivamente.

Neste artigo estamos interessados numa sub-classe particular dos problemas de regressão múltipla. A característica principal desta classe de problemas é o facto das aplicações que os motivam estarem unicamente interessadas na performance dos modelos numa gama reduzida de valores da variável objectivo. Em particular, na performance dos modelos para valores raros e extremos da variável objectivo. A performance nos valores mais comuns e normais é basicamente irrelevante no contexto deste tipo de aplicações. Esta classe de aplicações é bastante frequente em muitas áreas e está muitas vezes associada a problemas que exigem uma monitorização constante de um determinado fenómeno no sentido de atempadamente tomar acções preventivas ou outras, para evitar certo tipo de eventos raros. Pelos seus custos/benefícios este tipo de eventos são críticos nestes domínios, e a sua previsão/antecipação atempada pode trazer grandes benefícios no contexto destas classe de aplicações. Daí a relevância de termos critérios de avaliação que sejam capazes de identificar os modelos de previsão mais eficazes nesta tarefa tão crítica. Além disso, estes critérios, sendo eficazes, poderão levar ao desenvolvimento de modelos que os optimizem, e portanto mais vocacionados para a previsão destes eventos raros e extremos.

Neste contexto, nós argumentamos que os critérios "clássicos" como o MSE ou o MAD, não são adequados para esta classe de aplicações. De facto, eles olham para todos os erros da mesma forma, independentemente de onde eles ocorrem no domínio da variável objectivo. A magnitude dos erros é o factor decisivo que influencia o custo de uma qualquer previsão. Embora essa magnitude seja

relevante, nós argumentamos que ela deveria ser "pesada" pela relevância dos valores envolvidos numa previsão. Nestas aplicações a relevância não é uniforme para todos os valores do domínio da variável objectivo, e daí a inadequação das medidas "standard". Vejamos um pequeno exemplo artificial que ilustra o problema destas medidas.

Tabela 2: Um exemplo ilustrativo dos problemas das medidas "standard".

Observações	−5.29	−2.65	−2.43	−0.20	−0.03	0.03	0.51	1.46	2.53	2.94
M_1	−4.40	−2.06	−2.20	0.10	−0.23	−0.27	0.97	2.00	1.86	2.15
M_2	−5.09	−2.95	−2.89	0.69	−0.82	0.70	−0.08	0.92	2.83	3.17

Na Tabela 2 apresentamos as previsões de dois modelos artificiais (M_1 e M_2) para um conjunto de 10 casos de teste hipotéticos no contexto de um problema de prever as variações percentuais de um activo financeiro qualquer de um mercado bolsista. Para este tipo de aplicações é muito claro que o que estamos interessados em prever são as grandes variações, uma vez que as pequenas variações não possibilitam qualquer tipo de negociação dados os custos de transacção. Pela observação da tabela, é fácil notar que o modelo M_1 é melhor a prever as menores variações em termos absolutos, enquanto que o modelo M_2 é mais eficaz nos casos das variações maiores em termos de amplitude. Se calcularmos quer o MAD quer o MSE de ambos os modelos vamos ver que os valores são iguais para os dois modelos, respectivamente 0.497 e 0.29893. A indicação dada por estas duas estatísticas é claramente errónea uma vez que é bem óbvio que o modelo M_2 está mais adequado aos objectivos desta aplicação. Este pequeno exemplo artificial ilustra o problema de assumir que as magnitudes dos erros custam o mesmo ao longo do domínio da variável objectivo. Esta assumpção é inadequada para este tipo de aplicações e neste artigo propomos uma medida de performance que tenta resolver este tipo de problemas.

Um outro tipo de problema, não ilustrado no exemplo da Tabela 2, é o facto de que um modelo que tenha performance melhor nos casos extremos e raros, pode muito bem ver essa vantagem diluída por uma performance pior dos casos normais pelo facto de estes serem muito mais frequentes, o que do ponto de vista destas aplicações não deveria acontecer.

3 Algumas Abordagens Existentes

3.1 Uso de Pesos

Dentro do contexto do método para obter modelos de regressão descrito na Secção 2, existem uma série de alternativas às medidas "standard" de erro que podem ser mais adequadas para as nossas aplicações. Uma delas consiste na utilização de pesos associados a cada caso na amostra de dados. Algumas técnicas de modelação podem usar estes pesos para obter os modelos. Se atribuirmos pesos maiores aos casos associados a valores da variável objectivo que são mais relevantes para nós, poderemos obter modelos que sejam enviezados para a correcta previsão destes casos.

Assumindo que é fácil encontrar o peso a atribuír a cada caso, esta poderia eventualmente ser uma solução para os problemas descritos com as medidas "standard". Todavia, esta abordagem vê unicamente um lado do problema. De facto, esta abordagem não iria penalizar de forma adequada uma previsão de um valor extremo para um caso de teste cujo valor é normal. Por ser normal, o seu peso seria baixo e como tal este erro, uma espécie de falso alarme, não seria devidamente penalizado. Este problema resulta do facto de os pesos destas abordagens só levarem em consideração o valor da variável objectivo do caso em análise, não olhando ao valor previsto pelos modelos.

3.2 Medidas Especiais de Erro

Alguns autores das áreas económicas (e.g. [2]) abordaram este tipo de aplicações usando medidas de erro assimétricas. O objectivo principal destes trabalhos é o de distinguir dois tipos de erros

e atribuir-lhes os custos adequados, nomeadamente custos associados a previsões abaixo do valor verdadeiro ($\hat{y} < y$), ou previsões acima do mesmo ($\hat{y} > y$). Esse é o caso da estatística *LINLIN*,

$$LINLIN = \begin{cases} c_o|y-\hat{y}|, & \text{if } \hat{y} > y; \\ 0, & \text{if } \hat{y} = y; \\ c_u|y-\hat{y}|, & \text{if } \hat{y} < y. \end{cases} \quad (3)$$

em que c_o e c_u são os custos das previsões acima e abaixo dos valores verdadeiros, respectivamente.

Apesar das suas vantagens em relação às medidas "standard", a realidade é que esta medida peca por falta de generalidade em relação às nossas aplicações objectivo. De facto, ao considerar todo o tipo de previsões acima ou abaixo dos valores verdadeiros, como sendo iguais, esta abordagem falha em várias situações. Por exemplo, dentro do contexto anterior de previsões ligadas às variações no mercado bolsista, uma previsão de uma variação de -1% para um valor verdadeiro de 1%, tem a mesma amplitude que uma previsão de 7% para um valor verdadeiro de 8%, sendo ambas previsões abaixo dos valores verdadeiros. Como tal, são equivalentes do ponto de visto da medida *LINLIN*. Todavia, a primeira situação fornece uma indicação de descida dos mercados que poderia levar o investidor a vender activos, quando o mercado efectivamente acaba por subir, enquanto na segunda situação a previsão vai no sentido correcto. Como é óbvio ambas as situações têm consequências e logo custos/benefícios bem diferentes e isso não é levado em conta por esta medida.

4 *Precision* e *Recall* para Regressão

Conforme mencionamos anteriormente, em aplicações com custos diferenciados e com um grande desbalanceamento na distribuição dos valores da variável objectivo, as estatísticas *precision* e *recall* são as mais adequadas. A sua principal vantagem é focarem a análise da performance dos modelos nos casos que de facto interessam para a aplicação em causa. Informalmente a estatística *precision* mede a proporção de eventos previstos pelos modelos que de facto se confirmam como eventos. Quanto ao *recall* ele mede a proporção de eventos que ocorrem que são capturados pelos modelos como tal. Normalmente, existe um *trade-off* entre estas duas medidas (prever sempre um evento para qualquer caso de teste garante um *recall* de 100%, mas com muito baixo valor de *precision*), e frequentemente elas são integradas numa só medida, como por exemplo a medida F [6]. A ideia que descrevemos neste artigo é a de tentar transpôr estas medidas, desenvolvidas para problemas de análise discriminante, para problemas de regressão, como é o caso das nossas aplicações objectivo.

4.1 A nossa proposta

O uso dos conceitos de *precision* e *recall* no contexto da análise discriminante pressupõe a definição da "classe objectivo", isto é a classe que pretendemos que o modelo seja eficaz a prever, ou seja a classe que descreve os eventos (c.f. Tabela 1). Esta estratégia de enumeração não é possível em regressão, onde o domínio da variável objectivo é potencialmente infinito. Neste contexto, propomos o uso de uma função de relevância que faça o mapeamento entre o domínio da variável objectivo e uma escala de relevância[1],

$$\phi(Y): \quad]-\infty,\infty[\rightarrow [0,1] \quad (4)$$

Esta função, que é dependente do domínio em causa, permite o estabelecimento de diversos graus de relevância. Note-se que esta mesma estratégia poderia ser usada para os problemas de análise discriminante, definindo a função $\phi()$ da seguninte forma,

[1] Usamos o valor zero para valores irrelevantes, e um para valores maximamente relevantes.

$$\phi(Y) = I(Y = C_E) \tag{5}$$

em que $I()$ é a função indicadora que dá 1 quando o seu argumento é verdeiro e 0 de contrário, e C_E a etiqueta associada com a classe dos eventos.

Especificar a forma analítica da função de relevância poderá não ser uma tarefa fácil em vários domínios. Todavia, é possível arranjar uma especificação heurística que vá de encontro aos objectivos das aplicações. Esse é o caso dos nossos problemas, onde os valores mais relevantes a prever são raros e extremos. Neste contexto, a função de relevância pode ser vista como o complemento da função densidade de probabilidade da variavel objectivo. Os gráficos *box plot* fornecem informação importante sobre estas distribuições. Em particular, eles são a base de um teste de valores extremos, conhecido como a regra do *box plot*. Este teste assume uma distribuição Gaussiana e declara como extremos todos os valores acima de $adj_H = Q_3 + 1.5 \cdot IQR$ e abaixo de $adj_L = Q_1 - 1.5 \cdot IQR$, em que $IQR = Q_3 - Q_1$ e, Q_1 e Q_3 são o primeiro e o terceiro quartis, respectivamente. A nossa forma heurística de especificar a função de relevância, em aplicações em que tal não seja óbvio, consiste em usar uma função de relevância sigmóidal cujos parâmetros são baseados nos valores adj_H e adj_L, que determinam quando um valor é considerado extremo. Nomeadamente, definimos a função de relevância como,

$$f(Y) = \frac{1}{1 + \exp^{-s \cdot (Y - c)}} \tag{6}$$

em que c é o centro da sigmóide e s define a forma da curva. Os valores destes parâmetros são dependentes das estatísticas adj_H e adj_L, e também do tipo de extremos da aplicação. De facto, existem aplicações em que podemos falar da existência tanto de valores extremamente baixos como altos, enquanto que noutras aplicações só um destes tipos de extremos existem. Para aplicações com unicamente um dos tipos de extremos a função de relevância vai ser formada por uma única curva sigmóide, enquanto que em aplicações com os dois tipos de extremos vamos ter uma função com uma dupla sigmóide.

O parâmetro c define o valor para o qual a função vai ter o valor 0.5, i.e. $\phi(Y) = 0.5$. O seu significado é o do valor acima do qual a função de relevância vai começar a considerar os valores minimamente relevantes. Vamos usar como valores centrais para as duas sigmóides (caso ambas existam), os valores $c_L = adj_L$ e $c_H = adj_H$.

Quanto ao parâmetro s o seu valor deve ser escolhido de forma a que $\phi(c - c \cdot k) \simeq 0$ para valores extremos altos, e $\phi(c + c \cdot k) \simeq 0$ para extremos baixos, em que k é uma espécie de rácio de queda que determina quão rápido a curva sigmóide decresce para o valor 0. Seleccionando uma certa precisão Δ (e.g. $1e-04$) e resolvendo a equação em ordem a s obtemos,

$$s = \pm \frac{\ln(\Delta^{-1} - 1)}{|c \cdot k|} \tag{7}$$

em que o sinal $+$ é usado para extremos altos, e o $-$ para extremos baixos. No caso de aplicações com ambos os tipos de extremos, cada curva sigmóide é obtida usando os parâmetros descritos acima.

A Figura 1 ilustra o processo heurístico de obtenção de funções de relevância, que descrevemos acima. São apresentadas as duas funções de relevância obtidas para duas aplicações com dois tipos diferentes de extremos: somente extremos altos, e ambos os tipos de extremos. Foi desenvolvido código na linguagem R[2] que pode ser usado para obter este tipo de funções de relevãncia a partir de uma amostra de dados de um qualquer problema.

Devemos realçar que a nossa proposta de forma alguma é dependente do uso do processo heurístico descrito acima para obter funções de relevância. Tal processo só é útil em situações em que não existe conhecimento sobre o domínio que permita a especificação da função de relevância,

[2]Disponível em http://www.liaad.up.pt/∼ltorgo/DS09 .

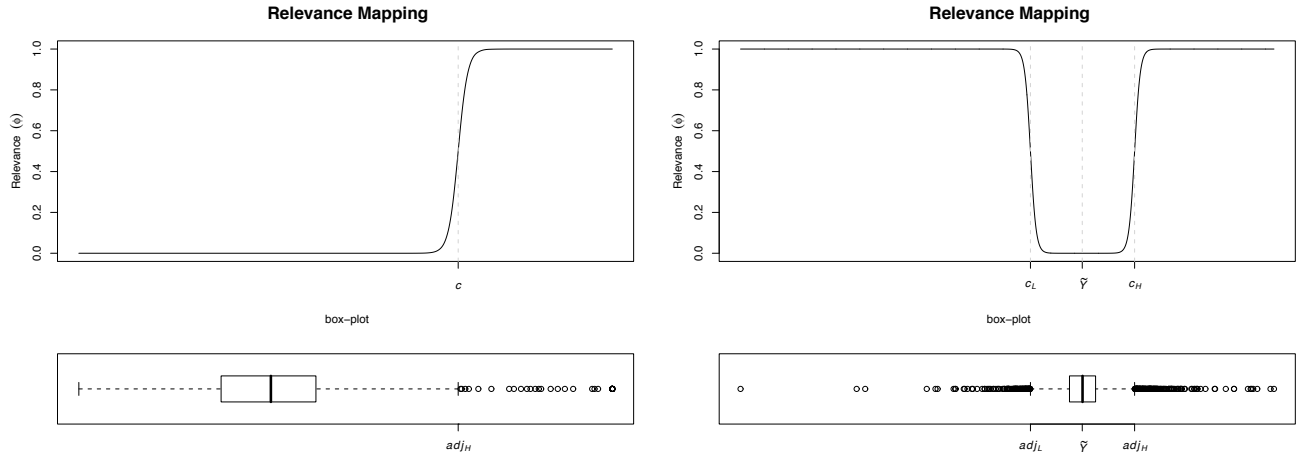


Figura 1: Dois exemplos de funções de relevância geradas com base nos *box plots*.

havendo somente a informação que os valores raros e extremos são os mais importantes/relevantes para os utilizadores.

A estatística *recall* é definida como a proporção de eventos reais que é capturada pelos modelos. Tendo definido a noção de evento como sendo função da relevância, podemos establecer que os eventos relevantes são aqueles cujo valor de relevância está acima de uma valor pré-establecido, i.e. $\phi(Y) \geq t_E$, em que t_E é um limite dependente do domínio de aplicação. Em problemas de análise discriminante, como a relevância é uma função 0/1, este limite é 1. Em regressão, este limite será provavelmente um valor próximo de 1, dependendo dos valores que se pretende considerar como relevantes no domínio em causa.

Importa agora clarificar a noção de "eventos capturados pelos modelos". Em análise discriminantes isto consiste em obter uma previsão correcta que é determinada por uma função 0/1 de erro, isto é $L_{0/1}(\hat{y}_i, y_i) = 0 \Leftrightarrow \hat{y}_i = y_i$. Em regressão as funções de erro são métricas com domínio $[0, \infty[$. Neste contexto impôr que $\hat{y}_i = y_i$ é, na maioria dos casos, demasiado restrictivo. Em geral estamos dispostos a aceitar um pequeno erro de previsão, considerando mesmo assim a previsão como "acertada", ou seja será acertada desde que $L(\hat{y}_i, y_i) \leq t_L$, em que t_L é um limite sobre o domínio da função de erro. Como anteriormente o valor de t_L é dependente do domínio.

Estamos agora em condições de fornecer uma definição formal da estatística *recall* válida tanto para problemas de análise discriminante como para problemas de regressão,

$$Recall = \frac{\sum\limits_{\{i \,|\, \phi(y_i) \geq t_E \wedge \phi(\hat{y}_i) \geq t_E\}} \alpha(\hat{y}_i, y_i) \cdot \phi(y_i)}{\sum\limits_{\{i \,|\, \phi(y_i) \geq t_E\}} \phi(y_i)} \tag{8}$$

em que $\alpha()$ é uma função que define a correcção de uma previsão.

Em análise discriminante a função $\alpha()$ é definida como,

$$\alpha(\hat{y}_i, y_i) = I(L_{0/1}(\hat{y}_i, y_i) = 0) \tag{9}$$

em que $L_{0/1}()$ é uma função 0/1 de erro "standard".

Para problemas de regressão definimos a função $\alpha()$ como,

$$\alpha(\hat{y}_i, y_i) = I(L(\hat{y}_i, y_i) \leq t_L) \tag{10}$$

em que t_L é o limite mencionado anteriormente que define o erro admissível no contexto da função de erro métrica $L()$.

Em alternativa é possível especificar uma noção de correcção de previsão mais "suave", usando uma função contínua no intervalo $[0, 1]$ em vez da função 0/1 da Equação 10. Isto permite aferir

de forma mais precisa a qualidade dos "sinais" dos modelos de regressão. Há muitas formas de mapear um conjunto de valores no intervalo $[0, t_L]$ para uma escala $[1,0]$. Exemplos incluem a interpolação linear ou outras funções lineares. Uma outra alternativa consiste em usar a função de erro complementar [1], que tem uma forma Guassiana que julgamos mais adequada para os objectivos desta aplicação,

$$\alpha(\hat{y}_i, y_i) = I(L(\hat{y}_i, y_i) \leq t_L) \cdot \left(1 - \exp^{-k \cdot \frac{(L(\hat{y}_i, y_i) - t_L)^2}{t_L^2}}\right) \tag{11}$$

em que k é um inteiro positivo que determina a forma da função. Valores maiores levam a decréscimos mais rápidos.

A estatística *precision* mede a proporção de eventos previstos pelos modelos que se confirmam na realidade. Vimos anteriormente qual a noção de evento tanto em análise discriminante como em regressão. A única diferença aqui é que estamos a falar de eventos "previstos". Em análise discriminante um evento previsto acontece quando o modelo prevê a classe objectivo associada aos eventos. Em regressão a previsão de um evento ocorre quando o valor previsto pelo modelo tem uma relevância maior do que o limite de relevância t_E. Neste contexto, propomos a seguinte definição para a estatística *precision*,

$$Precision = \frac{\sum\limits_{\{i \,|\, \phi(y_i) \geq t_E \wedge \phi(\hat{y}_i) \geq t_E\}} \alpha(\hat{y}_i, y_i) \cdot \phi(y_i)}{\sum\limits_{\{i \,|\, \phi(\hat{y}_i) \geq t_E\}} \phi(y_i)} \tag{12}$$

Conforme mencionado anteriormente, é possível agregar as duas estatísticas *precision* e *recall* numa só medida o que facilita a comparação de performance entre modelos. Esse é o caso da medida F [6] definida como,

$$F = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{13}$$

em que $0 \leq \beta \leq 1$ controla a importância relativa do *recall* contra a *precision*.

Em resumo, as definições para as estatísticas *precision* e *recall* aqui propostas, passíveis de serem agregadas numa só medida, podem ser usadas para aferir as capacidades de modelos de regressão múltipla em prever de forma acertada valores extremos e raros da variável objectivo. Experiências levadas a cabo [9] com vários problemas reais e várias técnicas de modelação, confirmam as vantagens do uso destas medidas em alternativa às medidas mais "standard", em aplicações cujo objectivo seja a performance em casos extremos e raros.

5 Conclusões

As medidas que definimos neste artigo possibilitam quer a comparação de modelos em aplicações com os objectivos anteriormente descritos, mas também o desenvolvimento de novas técnicas de modelação que optimizem estes novos critérios, o que levará à obtenção de modelos "afinados" para este tipo de aplicações. Em [9] são descritas várias comparações experimentais usandos estas novas métricas, que confirmam a sua eficácia em produzir rankings de modelos que estão mais adequados aos objectivos operacionais do tipo de aplicações que têm como objectivo único a previsão acertada de valores extremos e raros de uma variável contínua.

As novas medidas propostas para as estatísticas *precision* e *recall* podem ser vistas como generalizações das medidas normalmente usadas em análise discriminante [9]. Estas novas medidas têm como vantagem a possibilidade de serem também usadas em problemas de regressão, portanto extendendo a aplicabilidade das noções capturadas por estas estatísticas que são bastante relevantes para aplicações baseadas em eventos pouco frequentes.

Agradecimentos

Este trabalho tem o apoio dos projectos MORWAQ (PTDC/EIA/68489/2006) e oRANKI (PTDC/EIA/68322/2006) financiados pela FCT, e de uma bolsa de doutoramento da FCT (SFRH/BD/1711/2004) dada a Rita Ribeiro.

Referências

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. New York: Dover, 1972.

[2] P. Christoffersen and F. Diebold. Further results on forecasting and model selection under asymmetric loss. *Journal of Applied Econometrics*, 11:561–571, 1996.

[3] P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pages 155–164. ACM Press, 1999.

[4] C. Elkan. The foundations of cost-sensitive learning. In *Proc. of 7th International Joint Conference of Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.

[5] R. Ribeiro and L. Torgo. Rule-based prediction of rare extreme values. In N. Lavrac L. Todorovski and K. Jantke, editors, *Proceedings of the 9th International Conference on Discovery Science (DS 2006)*, volume 4265 of *Lecture Notes in Artificial Intelligence*, pages 219–230, Barcelona, Spain, October 2006. Springer.

[6] C. Van Rijsbergen. *Information Retrieval*. Dept. of Computer Science, University of Glasgow, 2nd edition, 1979.

[7] L. Torgo. Regression error characteristic surfaces. In R. Grossman, R. Bayardo, K. Bennett, and J. Vaidya, editors, *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2005)*, pages 697–702. ACM Press, 2005.

[8] L. Torgo and R. Ribeiro. Utility-based regression. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)*, pages 597–604. Springer, 2007.

[9] L. Torgo and R. Ribeiro. Precision and recall for regression. In *Proceedings of the 12th International Conference on Discovery Science (DS'2009)*, LNAI 5808, pages 332–346. Springer, 2009.

Classificação sensível aos custos em medicina

Alberto Freitas, *alberto@med.up.pt*

CINTESIS – Centro de Investigação em Tecnologias e Sistemas de Informação em Saúde
Serviço de Bioestatística e Informática Médica, Faculdade de Medicina da Universidade do Porto

Introdução

Este artigo serve como introdução a um tema onde a aprendizagem automática e a estatística desempenham um papel fundamental. É um artigo dedicado a vários aspectos da classificação sensível ao custo.

Nos problemas de classificação é costume medir-se o desempenho dos classificadores através da taxa de erro (*error rate*). A taxa de erro é a proporção dos erros encontrados entre todas as instâncias e é indicativa da performance global do classificador. Grande parte dos algoritmos de classificação assume que todos os erros têm o mesmo custo e são, normalmente, desenhados para minimizar o número de erros (perda 0/1). Nestes casos, a taxa de erro é equivalente à atribuição do mesmo custo para todos os erros de classificação, ou seja, por exemplo, no caso de uma classificação binária, os falsos positivos teriam o mesmo custo que os falsos negativos. No entanto, em muitas situações, cada erro poderá ter um diferente custo associado.

De facto, na maioria das situações do dia-a-dia, as decisões têm custos diferentes, e uma má decisão pode ter consequências graves. É pois importante ter em conta os diferentes custos associados às decisões, ou seja, às classificações obtidas.

Neste contexto pode-se designar por classificação sensível ao custo (*cost-sensitive classification*) quando os custos são ignorados na construção do classificador e são usados na previsão de novos casos. Por outro lado, pode-se falar em aprendizagem sensível ao custo (*cost-sensitive learning*) quando os custos são considerados durante a fase de treino, podendo posteriormente ser ignorados (ou não) na fase de previsão. Por ter em conta os custos na fase de construção do classificador a aprendizagem sensível ao custo será, à partida, uma opção mais ajustada e com melhores resultados. *Cost-sensitive learning* é, na área de aprendizagem automática (*machine learning*), a sub-área relativa a esta questão da não uniformidade dos custos.

Esta é uma área com especial importância em medicina. Basta ver, por exemplo, os casos onde uma determinada doença não é detectada atempadamente pelo facto de um de diagnóstico ter dado negativo, e as possíveis consequências (negativas) desse facto. Neste exemplo há claramente um custo mais elevado para os falsos negativos em relação aos falsos positivos, ou seja, existe um custo mais elevado para as situações onde a doença está presente e o resultado do teste indica que não.

Noutras situações, para confirmar um determinado diagnóstico ou quando um resultado falso positivo tiver implicações negativas no doente (por exemplo num teste para detecção de VIH - Vírus da

Imunodeficiência Humana), haverá necessidade de se usar testes de diagnóstico mais específicos. Neste caso, os custos associados às decisões são muito diferentes comparativamente com os do exemplo anterior.

Na medicina o custo de um falso negativo é normalmente (muito) diferente do custo de um falso positivo. Dado um determinado modelo de decisão, e para além dos custos de má classificação, será também importante considerar outros custos, por exemplo os relacionados com um determinado teste. Para um novo caso poderá ser necessário obter mais informação e existirá normalmente um determinado custo para obtenção dos valores dos atributos desconhecidos (por exemplo a realização de análises ao sangue).

Tradicionalmente a construção de uma árvore de decisão usa unicamente o ganho de informação (*information gain*), ou alternativamente o *gain ratio,* ou o *gini index*, como critério de escolha dos atributos a usar num determinado nó da árvore. Numa perspectiva de minimização dos custos fará sentido que a estratégia para a construção da árvore de decisão minimize simultaneamente os custos de má classificação e os custos dos testes. Se assim não for, não haverá a preocupação de evitar a presença dos atributos (testes) mais dispendiosos nos nós cimeiros da árvore de decisão, o que leva a que todos os novos casos efectuem estes testes, com um consequente aumento no custo total.

Suponhamos, por exemplo, que a classe a prever é se o doente tem ou não tem tumor cerebral. Quando o médico recebe um novo doente com uma dor de cabeça certamente não lhe vai de imediato recomendar que faça um TAC (Tomografia Axial Computorizada). Apesar de este ser um teste com grande poder discriminativo, o médico terá também em mente o custo do teste (ou seja, uma preocupação económica) e normalmente começará, em primeiro lugar, por fazer determinadas perguntas ao doente ou outros testes mais simples (Núñez, 1991).

Em medicina, um determinado teste médico é como um atributo em aprendizagem automática: o teste médico tem um custo que equivale ao custo do atributo; um diagnóstico errado tem um custo associado, que equivale ao custo de má classificação (erro) de uma instância. Na construção de um modelo de classificação para o diagnóstico médico, será necessário considerar os dois tipos de custos, os dos atributos e os da má classificação, ou seja, simultaneamente os custos dos testes médicos e os custos dos diagnósticos errados.

Na secção seguinte é feita uma exposição dos diferentes tipos de custos, onde são destacados os custos de má classificação e os custos dos testes.

Diferentes tipos de custos

Turney (2000) apresentou uma panorâmica geral e propôs uma taxonomia para os diversos tipos de custos que podem ocorrer em problemas de classificação. Assim, dividiu os custos em custos com os erros de má classificação, custos dos testes, custos dos peritos (para classificação de casos com classe desconhecida, antes ou durante a aprendizagem), custos de intervenção (custos necessários para alterar um determinado processo que poderá levar a alterações no custo do atributo), custos de resultados indesejados (resultantes de alterações a um determinado processo), custos de computação (com vários custos de complexidade computacional), custos dos casos (casos adquiridos para treino), custos da interacção homem-computador (na preparação dos casos para treino, selecção de atributos, definição de parâmetros ideais para optimizar a performance do algoritmo de aprendizagem, compreensão dos resultados) e custos da instabilidade (do algoritmo de aprendizagem.

Custo dos erros de má classificação

Para n classes está em geral associada uma matriz n x n, onde o elemento na linha i e coluna j representa o custo de classificar um caso na classe i quando ele de facto pertence à classe j. Normalmente o custo é zero quando $i = j$.

Tipicamente os valores existentes na matriz de custos são constantes, ou seja, o custo é o mesmo para qualquer instância classificada na classe *i* mas pertencente à classe *j*. No caso em que o custo é zero quando $i = j$ e é um para as restantes células, está-se perante a tradicional medida de taxa de erro.

Noutras situações, o custo dos erros de má classificação poderá ser condicional, ou seja, poderá ser dependente de determinadas características da instância ou do momento.

Existem vários exemplos na literatura de situações em que os custos de má classificação são dependentes da instância a classificar, isto é, estão relacionados com as características da instância. Na banca, o custo associado à não detecção de fraude está claramente associado ao montante envolvido em cada caso (Elkan, 2001). O mesmo se passa nas fraudes nas telecomunicações (Hollmén et al., 2000). Na medicina, o custo de prescrever um determinado medicamento a um doente alérgico pode ser diferente quando comparado com a sua prescrição a um doente não alérgico. Um diagnóstico errado pode ter consequências (custos) diferentes para cada doente, como por exemplo nos doentes mais idosos, ou que tenham determinadas comorbilidades.

Por outro lado, existem situações onde o custo de má classificação está associado ao momento em que ocorre. Um aparelho médico que monitorize um doente poderá emitir sinais de alarme perante um determinado problema. Os custos de classificação neste exemplo dependem não só do facto de a classificação estar ou não correcta, mas também da altura no tempo em que o alarme foi accionado, isto é, o sinal de alarme só será útil se houver tempo para uma resposta adequada ao problema existente (Fawcett e Provost, 1997).

Custo dos testes

A cada teste (atributo) pode estar associado um determinado custo. Na medicina, a maioria dos testes de diagnóstico tem um custo associado (ex.: uma ecografia ou um exame ao sangue). Estes custos podem variar muito entre diferentes testes. Por exemplo, em termos monetários, de acordo com a Portaria nº 567/2006 (Ministério da Saúde, Nº 113 – 12 de Junho de 2006, Anexo III), uma tomografia de emissão (SPECT) tem o preço de 107,4 € enquanto uma tomografia por emissão de positrões tem o preço de 1.392,8 €.

Os custos dos testes podem ser constantes para todos os doentes, mas podem também mudar conforme as características dos doentes. Por exemplo, de acordo com a mesma Portaria, uma prova de broncodilatação tem o preço de 36,1 €, mas se for para crianças com menos de 6 anos já custa 95,3 €, ou seja, a característica idade tem influência no custo do teste.

Por outro lado, os testes médicos podem ser muito díspares em termos de influência na qualidade de vida dos doentes. Se alguns testes são praticamente inócuos para os doentes (por exemplo a ecografia obstétrica), existem outros que podem colocar a sua vida em risco (por exemplo o cateterismo cardíaco), ou podem porventura ser muito desconfortáveis (por exemplo a endoscopia digestiva). Estas características dos testes deveriam poder ser combinadas com os custos monetários.

A existência de determinadas características comuns entre testes pode também possibilitar a sua realização em grupo. Assim, poderão existir testes que, feitos em grupo, sejam menos dispendiosos (e menos demorados) do que feitos individualmente (por exemplo (a) os três testes: ecografia renal, digestiva e ginecológica ou (b) os testes sanguíneos). Alguns testes podem também ter custos em comum e, por isso, poderão ser conjugados de forma a diminuir o custo total. Nestas circunstâncias estarão por exemplo os testes sanguíneos, onde existe um custo comum, a colheita de sangue, que pode ser único para todos os testes, ou seja, o sangue é recolhido uma única vez para a realização do primeiro teste não havendo necessidade de o fazer novamente para posteriores testes sanguíneos (se a colheita for suficiente). Neste exemplo, para além da poupança na recolha do sangue (poupança monetária e de incómodo para o doente), podem existir outras poupanças no caso de os testes serem feitos em grupo: poupança por o preço de grupo ser mais baixo e poupança no tempo (resultados poderão estar disponíveis mais rapidamente).

Os custos de alguns testes podem também depender dos resultados de outros testes (por exemplo o teste "idade" condiciona o custo do teste "prova de broncodilatação"). Poderão existir ainda testes com custos que sejam específicos para cada doente ou que tenham custos diferenciados consoante o momento ou a urgência com que são realizados. Uma estratégia de aprendizagem sensível aos custos deve também considerar e combinar estes tipos de custos.

Conhecendo de antemão os custos associados aos erros de má classificação, só fará sentido realizar determinados testes se os seus custos não forem superiores aos custos dos erros. Assim, qualquer teste que tenha custo superior aos custos de má classificação não fará sentido que se realize. Por outro lado, se os custos de todos os testes forem inferiores aos custos de má classificação fará sentido que se realizem todos os testes, a menos que existam testes claramente irrelevantes. Estes aspectos devem similarmente ser considerados numa estratégia de aprendizagem sensível aos custos (Freitas et al., 2009a).

Na secção seguinte é feita uma breve revisão ao trabalho feito nesta área da aprendizagem sensível ao custo.

Trabalho na área

Na aprendizagem indutiva (a aprendizagem a partir de exemplos), a maioria dos trabalhos preocupa-se com a taxa de erro ou taxa de sucesso dos modelos.

Existem, no entanto, alguns trabalhos que se preocupam com os custos não uniformes associados à má classificação, ou seja, diferentes custos conforme o tipo de erro (Breiman et al., 1984; Provost e Buchanan, 1995; Domingos, 1999; Elkan, 2001), mas sem abordar a questão dos custos dos atributos. Este tipo de trabalho pode servir também para resolver problemas de aprendizagem em dados não balanceados (por exemplo com rácio entre classes de 1 para 100, ou até mais) (Japkowicz e Stephen, 2002; Chawla et al., 2002; Liu e Zhou, 2006), dada a relação existente entre classes não balanceadas e custos assimétricos (Drummond e Holte, 2000). Ou seja, o problema das classes não balanceadas pode ser atacado aumentando o custo da classe que está em minoria e, por outro lado, é possível tornar um algoritmo sensível ao custo balanceando os dados de treino.

Outros trabalhos preocupam-se com os custos dos testes, sem considerar os custos de má classificação (Núñez, 1991; Tan, 1993; Melville et al., 2004).

E existem ainda alguns trabalhos que se preocupam simultaneamente com vários tipos de custos, descritos de seguida.

Trabalhos com custos dos testes e das más classificações

Turney (1995) implementou o sistema ICET que utiliza um algoritmo genético para construir uma árvore de decisão que minimize os custos dos testes e das más classificações. Os resultados apresentados mostram que o sistema ICET, apesar de robusto, é computacionalmente pesado quando comparado com outros algoritmos (por exemplo o C4.5). Turney considera que o problema de classificação sensível ao custo, na forma como o abordou é, essencialmente, um problema de *reinforcement learning*.

Zubek e Dietterich (2002) associaram o problema dos custos a um processo de decisão de Markov, com uma estratégia de pesquisa óptima (com uma heurística para o algoritmo AO*) que tem a desvantagem de poder ser computacionalmente muito dispendiosa. Anteriormente, outros autores descreveram também a utilização de processos de decisão de Markov, mas na versão generalizada onde são permitidos estados parcialmente observáveis (POMDP - Partially Observable Markov Decision Process), para a indução de árvores de decisão (Bonet e Geffner, 1998) e em conjunto com um classificador naive Bayes (Guo, 2003).

Arnt e Zilberstein (2004), consideram no seu trabalho, para além dos custos de má classificação e dos custos de teste, um custo de utilidade relacionado com o tempo necessário para obter o valor de um teste. Como no trabalho de Zubek e Dietterich (2002), usam também um processo de decisão de Markov e a heurística de pesquisa AO*.

Greiner et al. (2002) analizaram o problema de "aprendizagem activa" (*active learning*) para classificadores óptimos, sensíveis ao custo, usando uma variante do modelo *probably-approximately-correct* (PAC). Lizotte et al. (2003) estudou uma situação de aprendizagem activa, onde o classificador (naïve Bayes) tinha um orçamento fixo, limitado, que podia usar para "comprar" dados durante a fase de treino. No modelo estudado, sequencialmente eram escolhidas as características a comprar (com um determinado custo), tendo em consideração o limite imposto pelo orçamento e determinados parâmetros do modelo naïve Bayes.

Chai et al. (2004) propuseram um algoritmo sensível ao custo baseado em naïve Bayes, que reduzia o custo total dos atributos e das más classificações. Posteriormente (Sheng et al., 2005), apresentaram uma abordagem para árvores de decisão sensíveis ao custo onde é construída uma árvore para cada novo caso a testar. Nesse processo, dito *lazy*, só são considerados os custos dos atributos com valor desconhecido (para os atributos conhecidos o valor entra a zero) e, consequentemente, é construída uma árvore diferente para cada caso diferente. Zhang et al. (2005), comparam estratégias para averiguar se determinados valores desconhecidos (omissos) deverão ser obtidos ou não. Para testes que tenham um custo muito elevado (ou que pressuponham um risco) será mais vantajoso não obter os valores desconhecidos. Mais recentemente, Sheng et al. (2006), actualizaram a sua estratégia de construção de árvores de decisão sensíveis ao custo, com a inclusão de estratégias de teste sequenciais, num único grupo ou em vários grupos.

Meta-classificadores

É possível manipular as instâncias de treino ou manipular resultados de forma a obter classificadores sensíveis ao custo. Este tipo de algoritmo é conhecido por meta-classificador. Os meta-classificadores representam um componente que pré-processa os dados de treino ou que pós-processa os resultados dos classificadores não sensíveis aos custos. Os meta-classificadores podem ser aplicados sobre qualquer classificador já construído e fornecer previsões alteradas no sentido de minimizar os custos de má classificação (*cost-sensitive meta-learning*).

Os meta-classificadores podem utilizar, ou não, amostragem sobre os dados de treino. Exemplo da utilização de amostragem é o sistema Costing (Zadrozny et al., 2003), onde são atribuídos pesos às instâncias tendo por base custos. A outra abordagem, onde não é utilizada amostragem, pode ainda ser subdividida em três categorias, nomeadamente *relabeling,* através do critério do custo mínimo esperado (Michie et al., 1994), *weighting* (Fan et al., 1999) e *threshold adjusting* (Elkan, 2001).

O meta-classificador *CostSensitiveClassifier* (Witten e Frank, 2005) permite duas abordagens para tornar um classificador sensível aos custos de má classificação, considerando o custo total atribuído a cada classe para atribuir diferentes pesos às instâncias de treino (*weighting*), ou ajustando o modelo para que a classe prevista seja a que tem menores custos esperados de má classificação, e não a mais frequente (*relabeling*).

Dado que estes algoritmos trabalham sobre classificadores já construídos não têm influência na escolha dos atributos de teste em cada nó da árvore, tendo somente em consideração os custos de má classificação.

Na secção seguinte é apresentada uma divisão das várias abordagens possíveis para a aprendizagem sensível ao custo.

Avaliação, sensível ao custo, dos classificadores

Para a avaliação de classificadores sensíveis ao custo, Margineantu e Dietterich (2000) propuseram dois métodos estatísticos, um para construir um intervalo de confiança para o custo esperado de um

classificador em termos individuais, e outro método para construir um intervalo de confiança para a diferença esperada nos custos de dois classificadores. Em ambos os casos, a ideia base foi a de separar o problema de estimação das probabilidades de cada célula na matriz de confusão relativamente ao problema de computação do custo esperado.

Adams e Hand (1999) propuseram o *LC index*, que resulta de uma transformação das curvas *Receiver operating characteristic* (ROC), para facilitar a comparação de classificadores, através dos custos. Nesse trabalho, argumentam que normalmente não existe informação precisa sobre os custos, mas existe pelo menos uma ideia sobre a relação de um erro (Falso Negativo [FN]) sobre o outro (Falso Positivo [FP]) (por exemplo, os FN podem custar dez vezes mais do que os FP). O método proposto faz um mapeamento dos rácios dos custos dos erros num intervalo de 0 a 1 e transforma as curvas ROC em linhas paralelas, mostrando quais os classificadores dominantes em determinadas regiões do intervalo. O *LC index* é uma medida de confiança que permite indicar se um classificador é superior a outro num determinado intervalo. O *LC index* não expressa as diferenças de custos, sendo somente uma medida de superioridade (Fawcett, 2004).

Uma outra possibilidade, que surgiu como alternativa às curvas ROC (Fawcett, 2004), foi as curvas de custos (*cost curves*) (Drummond e Holte, 2006). Nas curvas de custo, a cada classificador está associada uma linha recta que mostra como a performance do classificador muda com alterações na distribuição dos custos das classes. As curvas de custos têm a maioria das características das curvas ROC mas permitem, adicionalmente, que se visualize determinadas medidas de performance que não eram facilmente possíveis com as curvas ROC. Entre essas medidas estão a possibilidade de visualizar os intervalos de confiança para o desempenho dos classificadores assim como a significância estatística da diferença de desempenho entre dois classificadores.

Um classificador simples, que origina uma matriz de confusão, é representado por um ponto (*FP*, *VP*) no espaço ROC. No espaço de custos esse ponto é representado por uma linha que liga os pontos (0, *FP*) e (1, *FN*), como no exemplo da Figura 1. Um conjunto de pontos no espaço ROC é um conjunto de linhas no espaço de custos. Assim como uma curva ROC é construída por um conjunto de pontos ROC ligados, a curva de custos é definida pelo conjunto de linhas de custos. Para informações mais detalhadas sobre as curvas de custos, consultar Witten e Frank (2005) ou Drummond e Holte (2006).

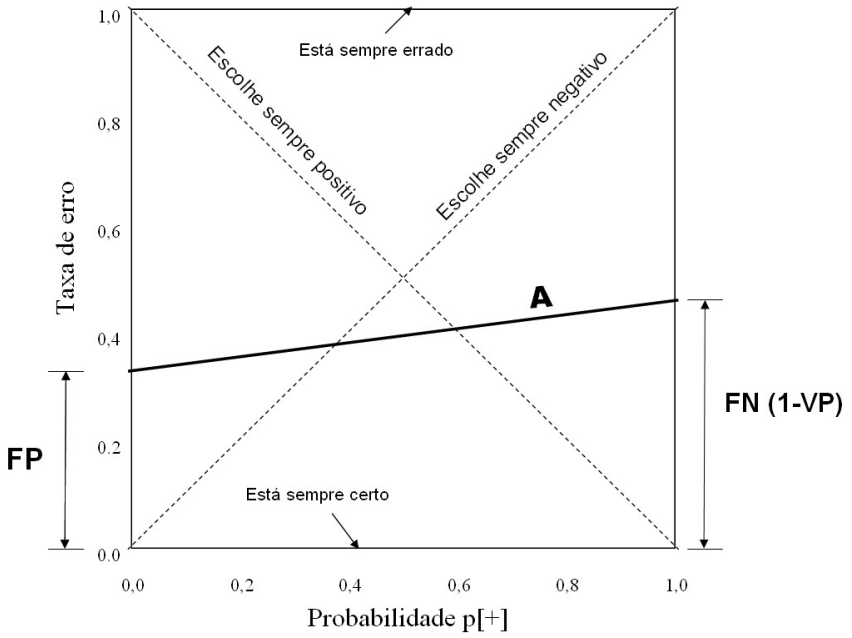


Figura 1 – Classificador **A** representado no espaço de custos

As curvas de custo são úteis por exemplo em situações onde há uma variação nas classes/custos ao longo do tempo. Conforme o balanceamento das classes, ou seja, conforme a função de probabilidade

de custos, um classificador *A* poderá ser melhor do que outro *B* e vice-versa. As curvas de custo permitem visualizar facilmente qual o melhor classificador para uma determinada probabilidade de valores de custo. Nessa visualização é possível também verificar se têm melhor ou pior desempenho do que os classificadores triviais (os que escolhem sempre "sim" e os que escolhem sempre "não").

Na secção seguinte é apresentada uma estratégia para a aprendizagem sensível a vários tipos de custos, incluindo custos relacionados com prejuízo na qualidade de vida associado a testes/exames médicos.

Uma nova estratégia de classificação sensível ao custo

Recentemente, Freitas et al. (2007, 2009b), propuseram uma estratégia de aprendizagem e de utilização sensível a vários tipos de custos. A estratégia adoptada incidiu essencialmente na modificação do algoritmo de treino, em especial no seu processo de partição do conjunto de exemplos de treino, pois só por este meio é possível que os custos dos testes sejam eficazmente considerados. Outras possíveis abordagens, com manipulação dos dados de treino ou com manipulação dos resultados, são mais indicadas para situações onde se pretende considerar os custos de má classificação.

Função de custo

A heurística utilizada foi definida para que a informação relativa a vários tipos de custos fosse usada na construção do classificador. Ou seja, foi definida uma nova função de custo de modo que a construção de árvores de decisão contemplasse os custos, em detrimento da tradicional função de ganho de informação. Esta heurística representa um rácio custo/benefício, muitas vezes utilizado em gestão.

Na definição desta estratégia foi utilizada também a abordagem com manipulação dos dados de treino, através da repesagem, para que os custos de má classificação fossem considerados. Desta forma é possível contemplar as diferenças de custos existentes entre a classificação de um falso negativo e a de um falso positivo.

A função de custo proposta altera o critério de escolha dos atributos, considerando por este meio os vários custos associados a cada atributo (teste). Esta função utiliza, conjuntamente, o tradicional ganho de informação e parâmetros associados aos custos:

$$\frac{\Delta I_i}{\left(C_i \phi_i\right)^{\omega}}$$

Heurística usada na selecção de atributos

Com,

- ΔI_i Ganho de informação para o atributo *i*
- C_i Custo do atributo *i*, com $C_i \geq 1$
- ϕ_i Factor de "risco" associado ao atributo *i*, com $\phi_i \geq 1$
- ω Factor escala de custos, com $\omega \geq 0$

Na construção da árvore, em cada nó será escolhido o atributo que maximize a heurística definida para a função de custo.

A função de custo não considera os custos de má classificação, por estes não terem influência no critério de divisão da árvore de decisão. Drummond e Holte (2000) mostraram que as divisões da árvore de decisão podem ser feitas independentemente dos custos de má classificação.

Os ***custos*** devem ser expressos na mesma unidade para todos os atributos. Atributos que não tenham custo, por exemplo a idade ou o sexo, ficam com o valor 1 ($C_i = 1$). Quanto maior for o custo de um determinado atributo, menor será o resultado da função de custo e terá, por conseguinte, menores possibilidades de ser escolhido.

O ***factor de risco*** (ou *prejuízo na qualidade de vida*), associado a um dado atributo, é uma parcela influente na função de custo. Este factor pretende dar um determinado peso a atributos que possam ser invasivos, possam causar desconforto ou incómodo, ou que possam de alguma forma contribuir para uma baixa na qualidade de vida dos pacientes (por exemplo, o teste invasivo "angiografia coronária"). O valor 1 ($\phi_i = 1$) significa ausência de influência, o que equivale a um teste completamente inócuo, e os valores superiores a 1 significam que existe influência, que será tanto maior quando maior for o valor do factor. O factor de risco poderá, por outro lado, ser também utilizado para penalizar os testes que sejam mais demorados.

Se $C_i = 1$ e $\phi_i = 1$ então a participação do custo do atributo e do seu factor de risco são ambos neutras e portanto a função de custo é igual ao tradicional ganho de informação.

O custo do atributo pode também ser ajustado através de um parâmetro geral, que se aplica por igual a todos os atributos candidatos, o ***factor escala de custos***. Este factor permite aumentar ou diminuir a influência dos custos na selecção dos atributos, servindo assim para regular a influência dos custos na escolha dos atributos para os nós da árvore. Desta forma pode-se tornar a árvore mais ou menos sensível aos custos. Para um factor zero os custos não são considerados, o que equivale a considerar o original ganho de informação. Com o aumento do factor o custo dos atributos terá cada vez mais influência na sua escolha, com os atributos mais caros a serem preteridos em função dos mais baratos.

Implementação

A implementação baseou-se no algoritmo de indução de árvore de decisão C4.5 (Quinlan, 1993), que foi alterado para contemplar custos e consequentemente gerar árvores de decisão sensíveis ao custo. Em relação aos custos de má classificação, o meta-classificador *CostSensitiveClassifier* (CSC) (Witten e Frank, 2005) foi alterado para incluir nos seus métodos a possibilidade de testar o modelo obtido considerando os custos dos atributos. Desta forma, utilizando em simultâneo o C4.5 alterado para considerar os custos dos atributos e o meta-classificador CSC, é possível obter modelos (árvores) que sejam sensíveis tanto às diferenças nos vários custos dos atributos como aos custos de má classificação.

Estratégias de teste da árvore de decisão

Depois de construída uma árvore de decisão é possível testá-la utilizando casos novos, não usados na fase de treino da árvore de decisão. Esta fase de teste serve para avaliar o desempenho do modelo construído, ou seja, quão bem o modelo consegue classificar um conjunto de casos novos. Depois de validado e de se conhecer o erro médio do modelo é possível utilizá-lo para a classificação de novos casos individuais.

De igual forma, depois de utilizado o sistema (C4.5-MCost) para a construção de uma árvore de decisão sensível aos custos, é possível utilizar um subconjunto dos dados para averiguar qual o custo médio dos novos casos classificados (nestas situações, não será tão habitual falar em erro médio mas sim em custo médio). Os custos associados aos erros de classificação estão também presentes neste custo médio. Desta forma, podemos comparar diferentes árvores de decisão.

Esta estratégia de teste permite a existência de custos para *testes*[1] (atributos) individuais e custos para grupos de testes. Para os grupos de testes há a possibilidade de distinção entre testes imediatos e testes

[1] Testes "médicos", atributos no conjunto de dados a testar/medir; exemplo: o teste idade, teste insulina sérica, teste da albumina (análises sanguíneas).

não imediatos. Nos testes não imediatos o médico tem que decidir, num determinado instante (nó da árvore), se pede todos os testes do grupo (sendo imputado o custo de grupo), ou se pede um teste de cada vez (sendo imputado o custo individual).

Custos individuais

Para além da estratégia de teste descrita, este sistema permite outras vertentes de teste, nomeadamente considerando características específicas de cada doente, modificando os custos dos atributos nas situações em que os valores destes já tenham sido obtidos anteriormente, e considerando a disponibilidade e demora de certos testes.

Alguns casos práticos com árvores de decisão sensíveis a custos mostram claramente os ganhos possíveis face à abordagem tradicional, ou seja, sem os custos serem considerados no processo de indução das árvores de decisão (Freitas et al., 2007b).

Conclusões

Na área da saúde os custos estão, directa ou indirectamente, presentes na maioria das situações. A um determinado teste diagnóstico podem estar associados uma variedade de custos de natureza económica ou não económica, como por exemplo o risco. A utilização de métodos de aprendizagem para a construção de modelos de diagnóstico ou prognóstico sensíveis a vários tipos de custos é um passo importante para que a aquisição de conhecimento por meios informáticos seja um processo cada vez mais natural e tendencialmente parecido com os processos mentais utilizados pelos médicos. Por outro lado, este tipo de estratégias pode permitir grandes poupanças económicas e melhorias nos custos relacionados com a qualidade de vida.

Nas suas decisões do dia-a-dia os profissionais de saúde têm normalmente em consideração vários tipos de custos. Para além dos custos económicos imediatos, consideram também os custos relacionados com o risco, com a demora ou com a possibilidade de realização de um determinado teste numa data altura. A construção de modelos de decisão utilizando técnicas de *data mining*, sejam elas do foro da estatística, de aprendizagem automática ou de base de dados, deve por isso ser, tendencialmente, o mais possível ajustada à realidade e à forma de pensamento dos seus utilizadores nos processos de tomada de decisão.

Referências

Adams, N. M. e D. J. Hand (1999), "Comparing classifiers when the misallocation costs are uncertain", Pattern Recognition, Vol. 32, Nº 7, pp. 1139-1147.

Arnt, A. e S. Zilberstein (2004), "Attribute Measurement Policies for Time and Cost Sensitive Classification", in Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04), pp. 323-326.

Bonet, B. e H. Geffner (1998), "Learning sorting and decision trees with POMDPs", in Proceedings of the 15th International Conference on Machine Learning (ICML), pp. 73-81.

Breiman, L., J. H. Freidman, R. A. Olshen e C. J. Stone (1984). Classification and Regression Trees, Belmont, California: Wadsworth.

Chai, X., L. Deng, Q. Yang e C. X. Ling (2004), "Test-Cost Sensitive Naive Bayes Classification", in Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'2004).

Chawla, N. V., K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer (2002), "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357.

Domingos, P. (1999), "MetaCost: A general method for making classifiers cost-sensitive", in Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99), pp. 155-164.

Drummond, C. e R. C. Holte (2000), "Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria", in Proceedings of the 17th International Conference on Machine Learning (ICML), pp. 239-246.

Drummond, C. e R. C. Holte (2006), "Cost curves: An improved method for visualizing classifier performance", Machine Learning, Vol. 65, pp. 95-130.

Elkan, C. (2001), "The Foundations of Cost-Sensitive Learning", in Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01), pp. 973-978.

Fan, W., S. J. Stolfo, J. Zhang e P. K. Chan (1999), "AdaCost: Misclassification Cost-Sensitive Boosting", in Proceedings of the 16th International Conference on Machine Learning (ICML), pp. 97-105.

Fawcett, T. (2004), "ROC Graphs: Notes and Practical Considerations for Researchers", Technical report, HP Laboratories, Palo Alto.

Fawcett, T. e F. Provost (1997), "Adaptive fraud detection", Data Mining and Knowledge Discovery, Vol. 1, Nº 3, pp. 291-316.

Freitas, J.A. (2007), "Uso de técnicas de data mining para análise de bases de dados hospitalares com finalidades de gestão", tese de doutoramento, Universidade do Porto.

Freitas, J.A., A. Costa-Pereira, P. Brazdil (2007b), "Cost-sensitive decision trees applied to medical data", In: Song, I.Y., Eder, J., and Nguyen, T.M. (Eds.): 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2007), LNCS Vol. 4654, 303-312, Springer-Verlag Berlin Heidelberg.

Freitas, J.A., A. Costa-Pereira, P. Brazdil (2009b). "Learning cost-sensitive decision trees to support medical diagnosis", In: Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications, Edited by Nguyen Manh Tho: Idea Group Inc (IGI): 287-307: Advances in Data Warehousing and Mining (ADWM) Book Series.

Freitas, J.A., P. Brazdil, A. Costa-Pereira (2009a), "Cost-sensitive learning in medicine", In: Data Mining and Medical Knowledge Management: Cases and Applications, Edited by Petr Berka, Jan Rauch, Djamel Abdelkader Zighed: Medical Information Science Reference; 57-75.

Greiner, R., A. J. Grove e D. Roth (2002), "Learning cost-sensitive active classifiers", Artificial Intelligence, Vol. 139, Nº 2, pp. 137-174.

Guo, A. (2003), "Decision-theoretic active sensing for autonomous agents", in Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 1002-1003.

Hollmén, J., M. Skubacz e M. Taniguchi (2000), "Input dependent misclassification costs for cost-sensitive classifiers", in Proceedings of the 2nd International Conference on Data Mining, pp. 495-503.

Japkowicz, N. e S. Stephen (2002), "The Class Imbalance Problem: A Systematic Study", Intelligent Data Analysis, Vol. 6, Nº 5, pp. 429-449.

Ling, C. X., Q. Yang, J. Wang e S. Zhang (2004), "Decision Trees with Minimal Costs", in Proceedings of the 21st International Conference on Machine Learning (ICML).

Ling, C. X., V. S. Sheng e Q. Yang (2006), "Test Strategies for Cost-Sensitive Decision Trees", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, Nº 8, pp. 1055-1067.

Liu, X. Y. e Z. H. Zhou (2006), "The influence of class imbalance on cost-sensitive learning: An empirical study", in Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06).

Lizotte, D. J., O. Madani e R. Greiner (2003), "Budgeted Learning of Naive-Bayes Classifiers", in Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence (UAI'03), pp. 378-385.

Margineantu, D. D. e T. G. Dietterich (2000), "Bootstrap Methods for the Cost-Sensitive Evaluation of Classifiers", in Proceedings of the 17th International Conference on Machine Learning (ICML-2000), pp. 583-590.

Melville, P., M. Saar-Tsechansky, F. Provost e R. Mooney (2004), "Active Feature-Value Acquisition for Classifier Induction", in Proceedings of the 4th IEEE International Conference on Data Mining (ICDM'04), pp. 483-486.

Michie, D., D. J. Spiegelhalter e C. C. Taylor (1994) (eds). Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence, Prentice Hall.
Núñez, M. (1991), "The use of background knowledge in decision tree induction", Machine Learning, Vol. 6, pp. 231-250.
Provost, F. e B. G. Buchanan (1995), "Inductive Policy: The Pragmatics of Bias Selection", Machine Learning, Vol. 20, pp. 35-61.
Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers.
Sheng, S., C. X. Ling e Q. Yang (2005), "Simple Test Strategies for Cost-Sensitive Decision Trees", in Proceedings of the 16th European Conference on Machine Learning (ECML), pp. 365-376.
Sheng, V. S. e C. X. Ling (2006), "Feature value acquisition in testing: a sequential batch test algorithm", in Proceedings of the 23rd International Conference on Machine Learning (ICML), pp. 809-816.
Sheng, V.S., C. X. Ling, A. Ni e S. Zhang (2006), "Cost-Sensitive Test Strategies", in Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06).
Tan, M. (1993), "Cost-sensitive learning of classification knowledge and its applications in robotics", Machine Learning, Vol. 13, pp. 7-33.
Turney, P. (1995), "Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm", Journal of Artificial Intelligence Research, Vol. 2, pp. 369-409.
Turney, P. (2000), "Types of cost in inductive concept learning", in Proceedings of the Workshop on Cost-Sensitive Learning at the 17th International Conference on Machine Learning (WCSL at ICML-2000), pp. 15-21.
Witten, I. H. e E. Frank (2005), Data mining: Practical machine learning tools and techniques, San Francisco: Morgan Kaufmann, 2nd Edition.
Zadrozny, B., J. Langford e N. Abe (2003), "Cost-Sensitive Learning by Cost-Proportionate Example Weighting", in Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM'03).
Zhang, S., Z. Qin, C. X. Ling e S. Sheng (2005), ""Missing Is Useful": Missing Values in Cost-Sensitive Decision Trees", IEEE Transactions on Knowledge and Data Engineering. Vol. 17, Nº 12, pp. 1689-1693.
Zubek, V. B. e T. G. Dietterich (2002), "Pruning improves heuristic search for cost-sensitive learning", in Proceedings of the 19th International Conference of Machine Learning (ICML), pp. 27-35.

CRM e Prospecção de Dados – ao seu serviço.

Marília Antunes, *marilia.antunues@fc.ul.pt*

DEIO-FCUL e CEAUL

Universidade de Lisboa

Customer Relationship Management (ou o Cliente Realmente Manda)

"Os clientes podem escolher qualquer cor, desde que seja preto" é, possivelmente, a frase mais conhecida de Henry Ford, fundador da Ford Motor Company, numa época em que eram as empresas que decidiam o que os clientes eram e em que classe cada cliente encaixava e em que as técnicas de marketing eram montadas em função da procura do produto e não do desejo do potencial cliente.

Mas foi o mesmo Henry Ford que, não só passou, em pouco tempo, a produzir automóveis doutras cores, como também afirmou que "Não é o empregador que paga os salários, mas o cliente.". De facto, o cliente exige que o fornecedor do bem ou serviço ofereça flexibilidade, criatividade, disponibilidade e preço vantajoso. Assim sendo, são necessárias técnicas que permitam descobrir esses atributos, para que as empresas tenham sucesso num mundo de desejos, preferências, comportamentos e lealdades de clientes em constante mutação.

O produto (ou serviço) certo, para o cliente certo, na hora certa, pelos canais certos, para satisfazer os desejos ou as necessidades dos clientes.

A sobrevivência e o sucesso de uma empresa dependem do cumprimento destes requisitos no tratamento do seu "bem" mais precioso – o cliente. Centradas no cliente, as empresas necessitam de atrair novos clientes, reter os clientes que já possuem e maximizar o lucro a obter por cliente. A palavra fidelização ganha uma nova importância. As empresas percebem que não basta satisfazer o cliente da primeira vez – é preciso fazer não só com que o cliente volte mas também com que dê lucro, numa relação de longo prazo. Para que aconteça a fidelização, é preciso conhecer o cliente identificando as suas características, necessidades e desejos, utilizando essas informações para estreitar o relacionamento. O Cartão de Cliente, inicialmente criado para distinguir os melhores clientes populariza-se e torna-se num elemento precioso para a empresa - vem permitir que sejam identificadas todas as transacções realizadas pelo cliente, passando a ser possível saber o que cada cliente compra, como o faz e com que regularidade.

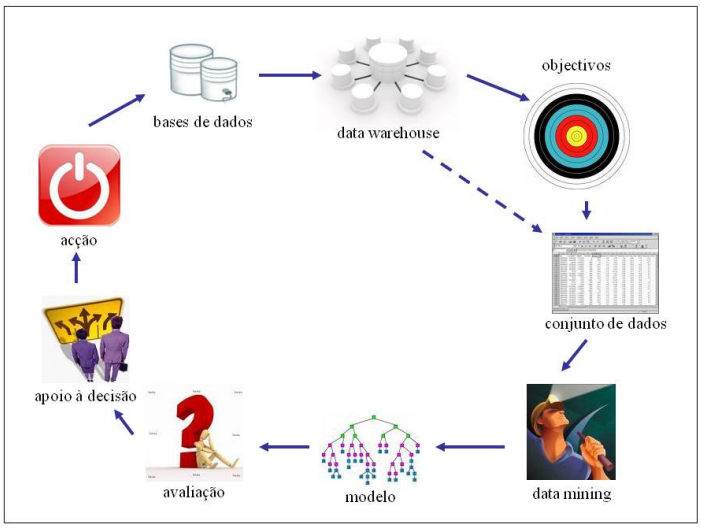
A tecnologia da informação, na forma de sofisticadas bases de dados alimentadas pelo comércio electrónico, por dispositivos instalados em pontos de venda, ATM's e outros pontos de contacto com o cliente, vêm proporcionar a evolução nas técnicas de marketing, de gestão e de relação com o cliente -- a informação contida nas bases de dados deve ser transformada em conhecimento que permita melhorar a relação com o cliente e aumentar os lucros.

Prospecção de Dados

Afinal, o que é Prospecção de Dados? Também apelidada de *mineração de dados,* por tradução mais imediata do termo inglês *data mining*, prospecção de dados é o processo de explorar grandes conjuntos de dados na busca de padrões consistente e úteis, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos conjuntos de dados. É uma área relativamente recente em ciência da computação mas utiliza várias técnicas da Estatística, recuperação da informação, inteligência artificial e reconhecimento de padrões. Frequentemente, prospecção de dados é entendida como sinónima de Knowledge Discovery in Databases (KDD) ou Descoberta de Conhecimento em Bases de Dados. Na verdade, KDD é um processo mais amplo consistindo das seguintes etapas:

1. Limpeza dos dados: etapa onde são eliminados ruídos e dados inconsistentes.
2. Integração dos dados: etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
3. Selecção: etapa onde são seleccionados os atributos que interessam ao utilizador. Por exemplo, o utilizador pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador ou não.
4. Transformação dos dados: etapa onde os dados são transformados num formato apropriado para aplicação de algoritmos de prospecção (por exemplo, através de operações de agregação).
5. Prospecção: etapa essencial do processo consistindo na aplicação de técnicas inteligentes a fim de se extrair os padrões de interesse.
6. Avaliação ou Pós-processamento: etapa onde são identificados os padrões interessantes de acordo com algum critério do utilizador.
7. Visualização dos resultados: etapa onde são utilizadas técnicas de representação de conhecimento a fim de apresentar ao utilizador o conhecimento extraído.

A etapa 7 deverá conduzir a uma reflexão sobre o conhecimento obtido, que servirá de apoio à tomada de decisões. Estas, por sua vez levarão a acções donde resultarão novos dados que alimentarão mais um ciclo do processo ilustrado na figura abaixo.

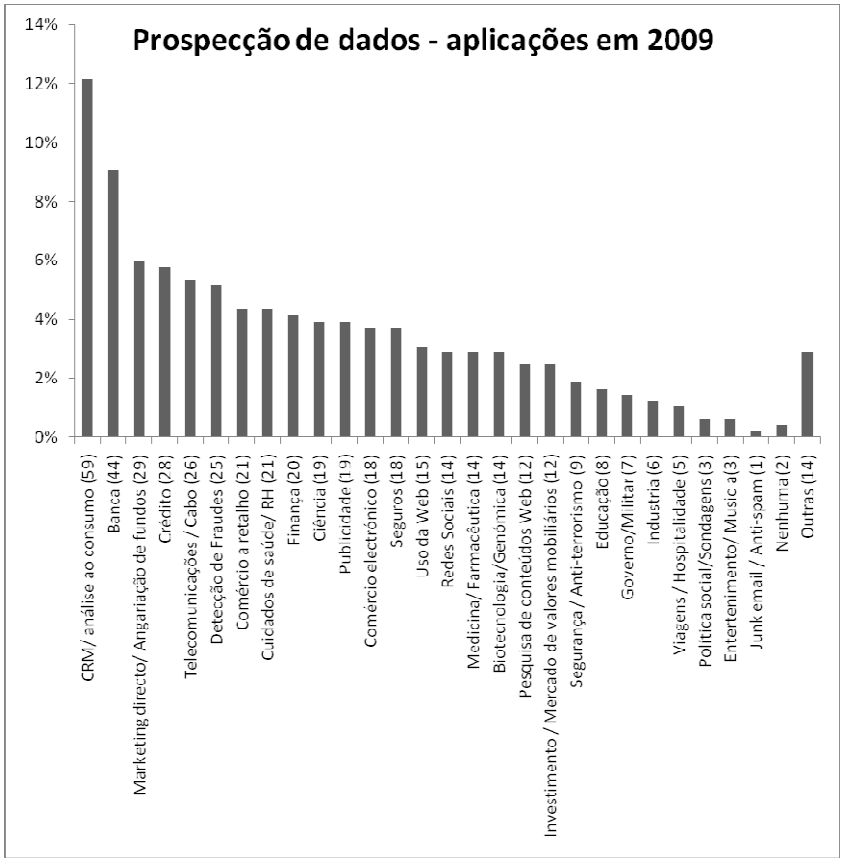


Por razões óbvias, a aplicação de metodologias de prospecção de dados popularizou-se rapidamente nas áreas de negócios. De acordo com um inquérito realizado no sítio da internet http://www.kdnuggets.com, que centraliza informação relacionada com a prospecção de dados, os sectores do consumo, da banca e das comunicações reúnem cerca de 60% das aplicações de técnicas de prospecção de dados.

Prospecção de dados dirigida e não dirigida

O estilo de abordagem a adoptar dependerá, essencialmente, do problema ao qual se pretende dar uma resposta. A prospecção de dados dirigida é a abordagem mais usual. O termo *dirigida*, aqui, significa que o alvo é conhecido e que se dispõe de dados ou conhecimento do passado que permitiram aprendizagem. Assenta no pressuposto que padrões identificados no passado continuam a ser, na sua essência, válidos. É o caso em que os objectivos da prospecção de dados estão bem definidos. Trata-se de uma abordagem de tipo topo-base e tem-se como exemplo os modelos preditivos, construídos com base em informação passada. Estes modelos são usados para prever desfechos de determinado tipo entre os elementos do novo conjunto de dados, como, por exemplo identificar quais os clientes que serão potenciais compradores de um determinado modelo de automóvel.

A prospecção de dados não dirigida recorre a técnicas sofisticadas para determinar padrões nos dados ou identificar grupos constituídos por elementos que, muitas vezes, à partida nem se supunham relacionados. Nesta abordagem, não há um alvo (objectivo concreto) a atingir, a não ser encontrar grupos que poderão ou não fazer sentido em termos práticos. Por partir de uma situação de desconhecimento, é considerada uma abordagem de tipo base-topo, de que as regras de associação são um exemplo. São utilizados sobretudo nas fases exploratórias da análise dos dados e exigem a intervenção humana para determinação da significância e importância dos padrões encontrados.



Aplicações da prospecção de dados em 2009 segundo Kdnuggets.

Como devo arrumar a loja? – regras de associação

A análise do cabaz de compras, que consiste na identificação dos produtos que são habitualmente comprados em conjunto, pode melhorar a política de stocks, a estratégia de disposição dos produtos, a definição de promoções, etc. A descoberta de padrões úteis e de regras de associação podem dar resposta a estas questões. Um padrão é um conceito à escala local, descrevendo um aspecto particular dos dados, ao contrário do que se pretende com os modelos, que é a descrição do conjunto de dados na sua globalidade.

Num conjunto de dados referente aos clientes de um supermercado, um padrão poderá ser "dez por cento dos clientes compra queijo e vinho"; para um conjunto de dados de alarmes em telecomunicações, um padrão poderá ser "se os alarmes A e B são accionados com um intervalo de

tempo inferior a 30 segundos, então o alarme C será accionado num espaço de tempo de 60 segundos, com probabilidade 0.5"; em dados referentes a tráfego de internet, um padrão poderá ser "se uma pessoa visita o site de notícias do canal de televisão A, então existe uma probabilidade de 0.6 de que visite o site de notícias do canal B no mesmo mês". Em cada um destes casos, o padrão é uma quantidade de informação relevante, da informação existente nos dados. A questão que se coloca é a de como encontrar os padrões interessantes. Dada uma forma de representar os possíveis padrões, uma forma de identificar os padrões presentes nos dados seria construir todos os padrões possíveis para o conjunto de dados e, para cada um, verificar se está presente nos dados e se essa presença é significativa. Nesta abordagem, surge de imediato o problema do número de padrões possíveis se tornar rapidamente incomportável. Por exemplo, se definirmos padrão como um conjunto de artigos dos vendidos num supermercado, se existirem 1000 artigos diferentes, então o número de padrões possíveis é 2^{1000}. Se os diferentes padrões forem completamente não relacionados entre si, então não existiria outra forma de identificar os padrões interessantes a não ser pelo método trivial atrás descrito. No entanto, o conjunto de padrões é, em geral, estruturado, usando-se precisamente essa estrutura para guiar a busca. Em geral, fala-se numa relação de generalização/especialização entre padrões. Um padrão α diz-se mais geral do que um padrão β, se sempre que β ocorrer nos dados, α também ocorrer. Por exemplo, o padrão "o cliente compra queijo e vinho" é mais geral do que o padrão "o cliente compra pão, queijo e vinho". A utilização deste tipo de relações de generalização entre padrões conduzem à construção de algoritmos simples para a identificação de padrões de certo tipo que ocorrem nos dados.

Representação de regras: Uma regra é constituída por duas proposições: a primeira, do lado esquerdo, chamada *antecedente* ou *condição,* e a segunda, do lado direito, chamada *consequente*. Por exemplo, "Se chover, então o chão ficará molhado". Uma regra é, pois, um par de proposições ligadas por uma implicação. Assim sendo, se a proposição antecedente for verdadeira, a proposição consequente também o será. Uma regra diz-se probabilística se à implicação acima se juntar uma probabilidade: se a proposição antecedente for verdadeira, então a proposição consequente será verdadeira com probabilidade *p*. A probabilidade *p* é a probabilidade condicional da veracidade da proposição consequente dada a veracidade da proposição antecedente.

As regras assim construídas apresentam a vantagem de serem facilmente compreendidas e aplicadas, representando uma base para a construção de modelos cognitivos. Note-se que as regras são de natureza intrinsecamente discreta; isto é, os seus dois termos são proposições a que se atribui um valor lógico (verdadeiro ou falso). Com certeza que poderemos utilizar variáveis de natureza contínua para construir tais proposições - nestes casos, as condições a considerar serão tais que aos valores possíveis para a variável corresponderá apenas "verdadeiro" ou "falso" - os valores são dicotomizados. Por exemplo, "Se X>10.2, então Y<1". Tipicamente, o termo antecedente é uma conjunção de proposições. Quando envolvem variáveis de natureza contínua, a conjunção de condições como "X_1>10.2 e X_2<3.6", conduz a regiões cujas fronteiras são paralelas aos eixos e, portanto, são hiper-rectângulos no espaço multidimensional gerado pelas variáveis consideradas.

Conjuntos frequentes e regras de associação: As regras de associação são uma forma simples de identificar padrões num contexto de prospecção de dados. Consideremos, por exemplo, os dados representados na tabela abaixo, descrevendo a composição de dez cestos de compras fictícios. São considerados cinco artigos (A, B, C, D e E) e dez clientes. A matriz de dados é uma matriz de tipo indicatriz, isto é, o valor 1 na célula *(i,j)* significa que o cliente *i* compra o artigo *j*, e o valor 0 significa que o cliente *i* não compra o artigo *j*. Em cada linha podem ser vistos os artigos que são comprados conjuntamente.

O objectivo é encontrar regras úteis a partir deste conjunto de dados. Dado um conjunto de valores 0, 1 correspondendo a observações de um conjunto de variáveis $A_1,...,A_p$, uma regra de associação tem a forma $((A_{i_1} = 1) \wedge \ldots \wedge (A_{i_k} = 1)) \Rightarrow (A_{i_{k+1}} = 1)$, onde $1 \leq i_j \leq p$ para todo o *j*. Esta regra de associação pode ser escrita de forma mais abreviada como $A_{i_1} \wedge \ldots \wedge A_{i_k} \Rightarrow A_{i_{k+1}}$. Um padrão como $(A_{i_1} = 1) \wedge \ldots \wedge (A_{i_k} = 1)$ designa-se por conjunto de itens, a que nos referiremos, também, como padrão. Assim, as regras de associação podem ser vistas como sendo da forma $\theta \Rightarrow \varphi$, onde θ é um conjunto de itens e φ é um item. Este conceito pode ser generalizado considerando-se que φ é, também, um conjunto de itens. Dado um padrão θ, a sua frequência, $fr(\theta)$, é igual ao número de casos

no conjunto de dados que satisfazem θ. Dada uma regra de associação $\theta \Rightarrow \varphi$, a sua acurácia ou confiança, $c(\theta \Rightarrow \varphi)$, é a proporção de casos que satisfazem φ entre aqueles que satisfazem θ, isto é, $c(\theta \Rightarrow \varphi)=fr(c(\theta \Rightarrow \varphi)/fr(\theta)$.

	A	B	C	D	E
t_1	1	0	0	0	0
t_2	1	1	1	1	0
t_3	1	0	1	0	1
t_4	0	0	1	0	0
t_5	0	1	1	1	0
t_6	1	1	1	0	0
t_7	1	0	1	1	0
t_8	0	1	1	0	1
t_9	1	0	0	1	0
t_{10}	0	1	1	0	1

Os padrões frequentes são padrões em geral muito simples, indicando quais os objectos que surgem conjuntamente com uma frequência razoável, mas conhecer apenas um destes padrões não fornece muita informação sobre os dados. Da mesma forma, conhecer apenas uma regra de associação dá-nos informação apenas sobre uma probabilidade condicional, não fornecendo informação sobre a distribuição de probabilidade conjunta das variáveis.

A tarefa de descoberta de padrões frequentes é simples: dado um nível de referência, *s*, para a frequência, determinam-se todos os padrões e calculam-se as respectivas frequências. No exemplo dado, considerando o nível 0.4 para a frequência, os padrões (ou conjuntos) frequentes são {A}, {B}, {C}, {D}, {AC} e {BC}. A partir daqui, extraem-se facilmente algumas regras: A⇒C com confiança 4/6=2/3; e B ⇒C com confiança 5/5=1.

Os algoritmos para descobrir regras de associação identificam todas as regras que cumpram os níveis estabelecidos para a frequência e a confiança (acurácia). Considerar níveis muito baixos para estas medidas leva à obtenção de um conjunto de itens frequentes muito elevado e, consequentemente, à identificação de muitas regras de associação. O desafio consiste, então, em identificar, de entre as regras descobertas, quais as mais interessantes, tendo em atenção que a acurácia de uma regra não é, necessariamente, um bom indicador do seu interesse. Por exemplo, em dados médicos, uma regra é que gravidez implica que o paciente é do sexo feminino com acurácia igual a 1.

A significância estatística de uma regra $A \Rightarrow B$ pode ser avaliada utilizando-se testes de hipóteses clássicos para testar se a probabilidade estimada $P(B=1|A=1)$ difere da probabilidade estimada $P(B=1)$, e quão provável é que esta diferença tenha ocorrido por mero acaso. Testar esta hipótese equivale a testar se existe diferença significativa entre $P(B=1|A=1)$ e $P(B=1|A=0)$. No entanto, a realização conjunta de um número elevado de testes acarreta todos os problemas da inferência simultânea, levando a que seja elevada a probabilidade de se considerar significativa, por mero acaso, uma diferença que na verdade não o é.

Descoberta de conjuntos frequentes e regras de associação: Os típicos conjuntos de dados relativos à venda a retalho contêm frequentemente um número de linhas (indivíduos ou transacções) que se situa entre 10^5 e 10^8 e um número de variáveis (itens ou artigos) que se situa entre 10^2 e 10^6. Naturalmente, as matrizes de dados são muito esparsas, com a probabilidade de encontrar um 1 numa determinada entrada, em geral da ordem de 0.1% ou inferior. O objectivo é encontrar todas as regras que cumpram os níveis pré-especificados de frequência e acurácia. Esta tarefa pode parecer pouco animadora uma vez que o número de conjuntos possíveis cresce exponencialmente com o número de

variáveis, o qual tende a ser também elevado, em particular no caso dos cestos de compras do comércio a retalho. No entanto, cada cliente compra um número reduzido de artigos (quando comparado com o número de artigos disponíveis), o que faz com que na prática o número de conjuntos de itens não seja tão elevado como teoricamente se admite que possa ser. Se o conjunto de dados for suficientemente grande, rapidamente surgem dificuldades de memória no tratamento do problema. Uma solução passa por dividir o problema em duas partes: primeiro encontram-se os conjuntos de itens frequentes e depois identificam-se as regras para os conjuntos identificados.

A partir do momento em que os conjuntos frequentes são conhecidos, descobrir as associações é uma tarefa simples. Se uma regra $X \Rightarrow B$ tem frequência pelo menos s, então o conjunto X tem frequência no mínimo igual a s. Então se todos os conjuntos frequentes são conhecidos, podemos gerar as regras da forma $X \Rightarrow B$, onde X é frequente, e avaliar a acurácia de todas as regras numa única passagem pelos dados.

Identificar de forma exaustiva todos os conjuntos frequentes é uma tarefa que se torna muito lenta. A solução é começar por identificar todos os conjuntos frequentes compostos por apenas um item. A ideia chave é que todos os conjuntos frequentes são compostos por elementos que são, também eles, conjuntos (elementares) frequentes. Em suma, identificam-se todos os conjuntos frequentes contento apenas uma variável. Depois, constroem-se todos os conjuntos compostos por duas variáveis: conjuntos *{A,B}* tais que *{A}* é frequente e *{B}* é frequente. Uma vez construído o grupo, identificam-se quais destes conjuntos são realmente frequentes, obtendo-se o conjunto dos conjuntos frequentes de dimensão 2. A partir destes constroem-se todos os conjuntos de dimensão 3, cuja frequência é calculada no passo seguinte, retendo-se apenas aqueles cuja frequência seja não inferior ao nível fixado previamente. Genericamente, no passo $i+1$ do algoritmo, $i \geq 1$, obtêm-se em primeiro lugar os conjuntos de tamanho $i+1$, candidatos a conjuntos frequentes C_{i+1}. Estes correspondem à reunião dos conjuntos retidos no passo anterior, L_i (o conjunto dos conjuntos frequentes de tamanho i) cuja dimensão seja igual a $i+1$. No passo seguinte identifica-se L_{i+1}, ou seja, quais os elementos de C_{i+1} cuja frequência é não inferior ao estabelecido. O cálculo das frequências associadas aos elementos de C_{i+1} faz-se numa só passagem pelo conjunto de dados, criando-se um contador para cada elemento de C_{i+1} e incrementando-se cada vez que é encontrado nos dados. Concluída a primeira parte, relativa à construção dos grupos e identificação dos grupos frequentes, passa-se à segunda parte em que se identificam as regras de associação e se calcula a acurácia correspondente. As regras de associação possíveis são todas aquelas relacionando os conjuntos A de L_i com conjuntos de B de L_{i+1} tais que A está contido em B. A acurácia estimada da regra é $fr(B)/fr(A)$.

O desafio

A prospecção de dados envolve uma grande variedade de tópicos nas áreas da computação e da estatística. Para os especialistas em ciências da computação a estatística apresenta-se quase como que impenetrável: um não acabar de vocabulário específico, pressupostos implícitos, resultados assintóticos e a dificuldade em perceber como "transformar" toda essa teoria em algoritmos. Para muitos estatísticos, a computação está demasiado cheia de discussões sobre algoritmos, pseudocódigo, eficiência computacional, onde quase não se faz referência a um modelo subjacente ou a procedimentos de inferência.

O desafio consiste em reunir as perspectivas da modelação matemática e da construção de algoritmos computacionais, para se ser bem sucedido na complexidade que a prospecção de dados envolve.

Bibliografia

Berry, M.J.A., Linoff, G.S. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. 2004. Wiley Publishing, Inc.

Hand, David J.; Mannila, Heikki; Smyth, Padhraic. *Principles of Data Mining*. 2001. MIT Press.

Swift, R. CRM. *O Revolucionário Marketing de Relacionamento com o Cliente*. 2001. Editora Campus.

Estatísticos e mineiros (de dados): inseparáveis de costas voltadas?

João A. Branco, *jbranco@math.ist.utl.pt*

Centro de Matemática e Aplicações
Instituto Superior Técnico
Universidade Técnica de Lisboa

O que faz pensar em colaboração entre estatísticos e especialistas em *Data Mining* (o método DM, mineração/prospecção de dados) é o facto de se admitir que tanto os seus objectivos como alguns dos seus métodos de trabalho são no mínimo semelhantes. E serão? Para responder é natural prosseguir comparando as duas actividades nomeadamente no que se refere às suas definições, objectivos, matéria-prima (dados) e métodos utilizados. É o que se faz a seguir nestas notas necessariamente breves, para concluir que, no nosso entendimento, essa colaboração se justifica e que deve ser promovida na prática.

1. Sobre as definições

1.1 Estatística

Apesar de certas noções e actividades rudimentares da estatística serem muito antigas a própria palavra estatística só aparece no século XVIII e a estatística que conhecemos agora no século XXI só começa a desenvolver-se de forma visível e sistemática praticamente a partir do início do século XX. Embora com princípios e métodos solidamente estabelecidos e cuja aplicação corporiza bem o método científico, os estatísticos continuam, ainda hoje, a fazer a pergunta "O que é a estatística?", como em Bartholomew (1995) e Brown e Kass (2009), e a revelar uma permanente falta de unanimidade nas suas respostas. Para o confirmar basta consultar dicionários e adequada literatura estatística (veja-se concretamente a resposta da Americam Statistical Association e da Royal Statistical Society) onde se podem encontrar definições diversas. Nestas circunstâncias qualquer pretensão de considerar que certa definição é a única definição só pode redundar em polémica. Uma visão que recebe a concordância de muitos consiste em considerar que a estatística é o estudo de métodos eficientes que permitem coleccionar, organizar, apresentar e analisar dados com o objectivo de tomar decisões. Mas qual é a razão da dificuldade inerente à definição de estatística? Será a sua natureza interdisciplinar, que deixa a definição ao sabor do interesse disciplinar de cada um? Será o seu passado que leva uns, principalmente os estatísticos do sector académico, a considerar que a estatística é um ramo da matemática e outros, principalmente os utilizadores, a pensar que a estatística é análise de dados, opondo uma interpretação mais ligada à teoria e outra mais ligada à prática da estatística? Será porque a definição de estatística é construída pensando essencialmente nos métodos que a estatística usa e não nos problemas que ela resolve?

1.2 DM

Esta é uma área muito jovem quando comparada com a idade da estatística uma vez que o seu aparecimento surge apenas recentemente quando a produção automática de dados leva à acumulação de grandes volumes de informação, situação que se estabeleceu em definitivo a partir dos meados da década iniciada em 1990. Chatfield (1997) refere que o termo aparece pela primeira vez num livro da área da econometria datado de 1978 e da autoria de E. E. Leamer. Contudo há indícios de o termo ter sido usado por estatísticos já em meados dos anos sessenta do século XX (Kish, 1998). Note-se que a designação DM é muito apelativa parecendo querer dizer que desta actividade se espera conseguir informação valiosa, à semelhança do que se passa com a actividade mineira, ou com jogos do tipo caça

ao tesouro. De acordo com a interpretação dos estatísticos dos primórdios de DM, esta actividade resumia-se a considerar um grande número de modelos para um mesmo conjunto de dados escolhendo o melhor, no sentido daquele que melhor se ajusta aos dados, como se fosse o verdadeiro modelo, ou a examinar exaustivamente o conjunto de dados com vista a identificar padrões indicadores de estrutura relevante para a tomada de decisões. Do ponto de vista estatístico nenhum destes dois casos é tranquilizante pois: i) na perspectiva da estatística mais tradicional não são os dados que devem determinar o modelo, mas sim outros factores que levam o investigador a propor o modelo antes dos dados serem conhecidos, ii) nem sempre os padrões encontrados correspondem a estruturas existentes na prática pois a pesquisa continuada em dados não uniformes acaba sempre por encontrar padrões formados apenas por acaso ou resultantes de variação aleatória. Esta potencial insatisfação terá levado a que o termo DM fosse inicialmente usado de forma pejorativa, fazendo-se equivalente a designações como magia negra, pesca de dados (*data fishing*) e dragagem de dados (*data dredging*), como quem diz, não se sabe como mas a solução há-de aparecer, ou o que vier à rede é peixe, não interessando o tamanho nem a qualidade. É natural que esta visão tenha afastado muitos estatísticos de uma actividade que naturalmente lhe diz respeito.

Ao contrário da interpretação negativa dos estatísticos, os especialistas da área da computação sentem que a nova abordagem é prometedora e vêem DM como uma nova disciplina que vai usar o forte poder computacional existente para analisar grandes conjuntos de dados. Tal como em estatística as definições de DM abundam, sendo determinadas pela interpretação, perspectiva e experiência dos seus autores, geralmente associadas com o domínio de actividade em que trabalham. Numa das primeiras definições, dada por Frawley et al. (1992), afirma-se que "*Data Mining is the nontrivial extraction of implicit, previously unknown, and potencially useful information from data*". Outra forma equivalente consiste em dizer que DM é o processo de aprender com os dados, ideia que prevalece subjacente a muitas outras definições. É por isso que DM também é conhecida por KDD (*Knowledge Discovery in Data Base*) embora este consista num processo mais geral de aquisição de conhecimentos e do qual DM se pode considerar um procedimento particular. Mas que dados esperam os analistas tratar com DM, que métodos usam e que objectivos esperam alcançar?

Os dados e métodos de análise

A facilidade com que a tecnologia actualmente disponível pode produzir e captar dados tem levado a que por todo o lado, países, instituições e empresas decidam armazenar dados, na esperança de não perder eventual informação relativa a toda a actividade em jogo. Essa informação, escondida nos dados, encontra-se em geral dispersa e desconexa esperando-se que uma análise dos dados a venha revelar. É neste contexto que surge o método DM (a necessidade é a mãe da invenção) que recorre à intervenção de várias áreas entre as quais estão a estatística, a computação, a aprendizagem automática e a manipulação e gestão de bases de dados. É a grande quantidade de dados, a sua natureza e qualidade que justifica o aparecimento do DM uma vez que os métodos de análise tradicionais não foram concebidos para outra realidade. Como é que os métodos tradicionais, concebidos principalmente para dados experimentais resultantes de algum delineamento, ou pelo menos obedecendo a algum processo de recolha previamente estipulado, podem ajudar na análise de milhões de observações cada uma das quais medida num grande número de características de natureza variada, numérica ou não numérica, que vão sendo registadas manual ou automaticamente, carregando quantidades enormes de anomalias, duplicações, erros, inconsistências, omissões e outras faltas? Imagine-se, por exemplo, o papel do simples diagrama de dispersão, que tão útil se mostra para a análise do comportamento conjunto de duas variáveis. Com aquela quantidade de observações obter-se-ia uma mancha negra naturalmente inadequada para mostrar aquilo que se esperaria observar. E se o número de variáveis é de facto muito grande e maior do que o número de observações, como acontece com os *microarrays*? O que vai ser dos métodos estatísticos baseados na matriz de correlações? E como justificar o uso de modelos assentes em hipóteses que este tipo de dados naturalmente não satisfaz? Quanto aos testes de hipóteses o efeito de um grande número de observações traduz-se na rejeição sistemática da hipótese nula perdendo-se a utilidade prática do teste. Situações em que os conjuntos de dados se afastam do que é habitual ver em estudos estatísticos, em termos da natureza dos dados e do número de observações e variáveis, abundam em todos os campos de actividade em que a moderna tecnologia está implantada, por exemplo: os dados associados ao

genoma humano, o número de transacções efectuadas por grandes cadeias de empresas comerciais e financeiras e os dados recolhidos da observação regular de fenómenos astronómicos e meteorológicos.

Quanto aos objectivos do procedimento DM podem considerar-se duas grandes categorias: i) a predição de uma variável resposta a partir de outras variáveis explicativas, que assenta na utilização de métodos de classificação, quando a resposta é categórica, e métodos de regressão, quando a resposta é contínua e ii) a descrição dos dados com base numa simples análise exploratória assente em métodos da estatística descritiva com vista à descoberta de tipos de distribuições, tendências e correlações, ou recorrendo a modelos de natureza descritiva como a segmentação e análise de clusters dos dados, ou pesquisando diferentes tipos de anomalias (detecção de *outliers*) ou procurando combinações entre os objectos que geraram os dados com o objectivo de estabelecer regras de associação entre eles. Métodos Bayesianos, análise de componentes principais e séries temporais são outro recurso do DM. Fora da estatística DM recorre a redes neuronais e variados métodos de optimização para construir os algoritmos que usa para atingir os objectivos a que se propõe. O bom funcionamento do DM está dependente do uso de sofisticado software e de potente hardware o que faz com que as empresas se sirvam do DM, como chamariz, para vender estes produtos, desenvolvendo uma actividade bastante lucrativa.

2. Estatística e DM

Com base no que foi exposto podemos concluir que estatística e DM são dois processos de resolução de problemas que apresentam semelhanças nos objectivos que perseguem e nas técnicas que usam para atingir esses objectivos. Mas as duas abordagens diferem em vários aspectos, o que vem confirmar as suas identidades únicas. Não sendo possível ser exaustivo destacam-se algumas dessas diferenças: i) a diferença mais marcante é a grande magnitude e complexidade dos dados analisados por DM, contrariamente ao que sucede em estatística que geralmente se ocupa de conjuntos pequenos de dados estáticos, limpos ou com pequena quantidade de anomalias e outras perturbações, constituindo uma amostra com vista a responder a questões concretas relativas à população de onde a amostra foi retirada, ii) a limpeza, compressão e transformação dos dados são operações essenciais do processamento inicial em DM mas não são geralmente requeridas em estatística iii) os métodos tradicionais da estatística não são geralmente aplicáveis directamente ao tratamento dos grandes conjuntos de dados que o DM consegue analisar, iv) enquanto que a estatística pode ser vista como um processo de analisar relações o DM é um processo de descobrir relações, v) a abordagem da estatística é de natureza confirmatória ao passo que DM segue uma abordagem exploratória.

Apesar do esclarecimento que acaba de fazer-se continua a subsistir alguma confusão entre os profissionais das duas áreas, sobretudo entre aqueles cujas opiniões são mais radicais, sobre a relação entre DM e estatística. Para clarificar o problema vários autores têm dado as suas contribuições: Friedman (1998), Hand (1998, 1999a, 1999b), Kuonen (2004) e Zhao e Luan (2006). Do lado dos estatísticos em geral há uma postura, possivelmente influenciada por uma interpretação inicial sobre DM, explicada anteriormente em 1.2, que encara o método depreciativamente e questiona o real valor da estatística em DM em face de uma proclamada deficiente formação em estatística dos especialistas em DM , veja-se Goodman (2001) e a sequências de respostas de estatísticos e mineiros que este artigo despoletou. Contudo alguns estatísticos de renome consideram que DM não passa de um novo desenvolvimento da estatística no qual os profissionais se devem envolver com vista a contribuir para resolver os imensos desafios que o novo paradigma vem colocando à estatística, sob pena de outros enfrentarem esses desafios com as ferramentas que souberem construir. O desinteresse dos estatísticos por áreas emergentes (*pattern recognition*) que lhes dizem respeito não é novo, e isso tem consequências para o desenvolvimento da disciplina. Do lado dos profissionais de DM sente-se também antipatia pela estatística mas orgulho por conseguirem progredir num campo onde os estatísticos ainda não chegaram.

David Hand é um dos estatísticos que desde o princípio tem fomentado a ligação entre as duas áreas e destacado os benefícios que podem resultar da interacção dos dois campos de actividade. Outros como Leo Breiman sugerem a utilização de novas ferramentas que não a tradicional modelação para resolver os novos problemas colocados pelos grandes volumes de dados. Veja-se, Breiman (2001), o seu interessante percurso, de académico para consultor e o regresso à academia e os ensinamentos que nos trouxe. Felizmente há sinais da interacção da estatística com áreas do foro computacional,

onde vão procurar guarida alguns dos problemas do foro da estatística. Veja-se, por exemplo, Christmann e Shen (2009). O que deve ser feito para incentivar esta interacção? Como devemos actualizar o ensino da estatística de forma a contemplar estas preocupações (Ganesh, 2002)? Estes são apenas dois tópicos, entre muitos outros, em que vale a pena reflectir.

3. Conclusão

O problema da análise de dados, quer se trate de um pequeno conjunto de dados, quer se trate de um conjunto extremamente volumoso é um problema de natureza estatística. A estatística clássica desenvolveu-se e preparou-se para estudar os pequenos conjuntos, sendo muitos dos seus métodos inapropriados para conjuntos volumosos. Compete à estatística moderna adaptar-se, criando métodos adequados, e prosseguir o seu desenvolvimento para enfrentar os desafios e as novas questões que os grandes conjuntos vieram colocar. Se não o fizer vai haver outros, que não os estatísticos, que o tentam fazer, e que o farão, pois a marcha do progresso não se detém perante estas dificuldades. Em DM há um grande empenho em responder às questões que os grandes volumes de dados levantam. É um facto que DM não pode funcionar sem estatística e que a estatística precisa das técnicas de DM se pretender analisar grandes conjuntos de dados. E uma vez que a estatística e DM, embora distintas, partilham objectivos e métodos, a nossa convicção é que a cooperação entre os especialistas das duas áreas proporciona um avanço mais rigoroso, mais seguro e mais rápido do processo geral da análise de dados. Mas uma cooperação sistemática, se existe, é ainda muito precária.

4. Referências bibliográficas

Bartholomew, D. J. (1995). What is statistics? *Journal of the Royal Statistical Society*, A, **158**, 1-20.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, **16**, 199-231.

Brown, E. N. e Kass, R. E. (2009). What is statistics? (With discussion). *American Statistician*, **63**, 105-110

Chatfield, C. (1997). Data mining. *Royal Statistical Society News,* **25**, 3, 1-2.

Christmann, A. e Shen, X. (2009). Editorial: On the interface of statistics and machine learning. *Statistics and Its Interface*, **2**, 255-256.

Frawley, W., Piatetsky-Shapiro. Mathews, C. (1992). Knowledge discovery in Databases: an overview. *AI Magazine*, 213-228.

Friedman, J. H. (1998). Data mining and statistics: what is the connection**.** Unpublished. www-stat.stanford.edu/~jhf/ftp/dm-stat.ps.

Ganesh, S. (2002). Data mining: should it be included in the statistics curriculum? *The 16 th International Conference on Teaching Statistics (ICOTS 6).* Cape Town, South Africa.

Godman, A. (2001). Statistics is the road from data mining to knowledge discovery. *KDnuggets*: *News*: **n24**: item1.

Hand, D. J. (1998). Data mining: statistics and more? *The American Statistician*, **52**, 112-118.

Hand, D. J. (1999a). Statistics and Data mining: intersecting disciplines. *SIGKDD Explorations*, **1**, 16-19.

Hand, D. J. (1999b). Data mining: new challenges for statisticians. *Social Science Computer Review* **18**, 442-449.

Kish, L. (1998). Royal Statistical Society News, **25**, 6, 8.

Kuonen, D. (2004). Data mining and statistics: what is the connection? *The Data Administrative Newsletter*. Switzerland.

Leamer, E. E. (1978). *Specification searches: ad hoc inference with nonexperimental data*. John Wiley & sons, New York.

Zhao, C. M. e Luan, J. L. (2006). Data mining: Going beyond traditional statistics. *New Directions for Institutional Research*, **131**, 7-16

Métodos de Análise Espectral Singular - Implementação Computacional

Miguel de Carvalho, *mbcarvalho@fct.unl.pt*
Banco de Portugal / CMA Universidade Nova de Lisboa

Paulo Canas Rodrigues, *paulocanas@fct.unl.pt*
Wageningen University / CMA Universidade Nova de Lisboa

1. Introdução

A Análise Espectral Singular (AES) consiste numa generalização das técnicas de componentes principais para o domínio das séries temporais (Golyandina & al, 2001). A ideia basilar da AES consiste na decomposição da série temporal em diversos blocos distintos que possam ser identificados como componentes referentes a tendência, movimentos sazonais, ruído, etc. São também conhecidas na literatura técnicas para articular com a AES, por forma a permitir a condução de experiências de previsão. A mais popular dessas técnicas é sem dúvida o *recurrent forecast algorithm*.

O método encontra a sua motivação original na decomposição de Karhunen-Loève (Loève, 1978), e outros resultados clássicos sobre a representação ortogonal de processos estocásticos contínuos. Ainda que exista alguma falta de consenso na literatura grande parte das raízes da AES são geralmente atribuídas aos trabalhos seminais de Basilevsky e Hum (1979) e de Broomhead e King (1986). Outras referências clássicas incluem Vautard e Ghil (1989) e Vautard et al. (1992). O domínio de aplicação do método inclui a Climatologia (Allen & Smith, 1996), Meteorologia (Paegle & al., 2000) e a Geofísica (Ghil & Vautard, 1991; Ghil et al., 2002; Kondrashov & Ghil, 2006). A AES é também aplicada na condução de exercícios de previsão. Com efeito, foi recentemente sugerido na literatura (Hassani & al., 2009) que os métodos de análise espectral singular dominam, em termos de previsão de médio e longo prazo, o célebre modelo ARIMA (*Autoregressive integrated moving average*).

A mecânica do método AES pode ser dissociada em duas fases – **decomposição** e **reconstrução**. A decomposição engloba os passos de *embutimento* e *decomposição em valores singulares*. A reconstrução, por seu turno, inclui os passos de *agrupamento* e de *diagonalização por médias*. Nesta nota pretendemos ilustrar o **package SSA (Singular Spectrum Analysis)** – um package que temos vindo a desenvolver em `R` [`R` Development Core Team]. Este package permite realizar a reconstrução de séries temporais e a condução de exercícios de previsão com base em métodos AES. Neste momento o package está a ser sujeito a uma bateria de testes de verificação. Logo que possível iremos disponibilizar o package no `CRAN` (The Comprehensive `R` Archive Network) no site `http://cran.r-project.org`.

Esta nota está estruturada do seguinte modo. A Secção 2 apresenta uma breve resenha sobre decomposição ortogonal de processos estocásticos. O *modus operandi* dos métodos de análise espectral singular é descrito na Secção 3. A Secção 4 ilustra alguns dos comandos do nosso package.

2. Breve Resenha sobre Decomposição Ortogonal de Processos Estocásticos

De modo a motivar a AES, iremos introduzir de seguida alguns resultados preliminares referentes à representação ortogonal de processos estocásticos. (A descrição do método AES e motivação teórica

subjacente aqui exposta segue de perto de Carvalho e Rodrigues, 2009.) Estes resultados teóricos são basilares para uma compreensão sólida da mecânica subjacente à AES. Uma das representações ortogonais mais célebres é dada pela decomposição de Karhunen-Loève (Loève, 1978) (vide Lema 1 em baixo). Essencialmente, esta decomposição garante que qualquer variável aleatória que seja contínua em média quadrática possa ser representada como uma combinação linear de funções ortogonais. Existe ainda um resultado análogo para as funções de autocovariância, caso em que a decomposição é conhecida como teorema de Mercer. Este resultado estabelece que, sob determinadas condições de regularidade, a função de autocovariância $\gamma(r, s)$ pode ser escrita através da seguinte soma

$$\gamma(r,s) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\omega_i(r)\omega_i(s),$$

onde os $\omega_i(r)$ designam as funções próprias (*eigenfunctions*) da função de autocovariância $\gamma(r, s)$, e os λ_i denotam os valores próprios correspondentes. Uma prova deste resultado clássico pode ser encontrada por exemplo em Hochstadt (1989), página 90. A decomposição estabelecida pelo teorema de Mercer pode posteriormente ser empregue para providenciar uma representação ortogonal do próprio processo estocástico. De facto, esta decomposição é central no estabelecimento da decomposição de Karhunen-Loève, também frequentemente designado por *proper orthogonal decomposition theorem*. Apresentamos então o resultado central desta secção

Lema 1. *(Decomposição de Karhunen-Loève) Seja* $Y(t)$ *uma função aleatória contínua em média quadrática definida no intervalo* $I = [0, t]$. *A função* $Y(t)$ *admite em I uma decomposição ortogonal da forma*

$$Y(t) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\omega_i(r)v_i,$$

para algumas quantidades estocásticas ortogonais, v_i *, sse* λ_i *e* $\omega_i(r)$ *forem respectivamente dadas pelas ortonormalizações dos valores próprios e das funções próprias da função de autocovariância* $\gamma(r, s)$.

A prova deste resultado pode ser encontrada por exemplo em Loève (1978), nas páginas 144–145. Apesar da generalidade deste resultado teórico, na prática preferimos frequentemente uma variante discreta desta decomposição em análise multivariada. Assim, nas aplicações preferimos considerar os vectores próprios da versão discreta da função de autocovariância $\gamma(r, s)$ em substituição das funções próprias. Adicionalmente, do ponto de vista das aplicações, é frequente recorrer à truncagem de um número finito de termos na decomposição descrita em cima. De facto, o nome desta decomposição está na origem da designação "transformação de Karhunen-Loève" com que frequentemente são baptizadas as técnicas baseadas em componentes principais.

3. Modus Operandi da Análise Espectral Singular

3.1 Análise Espectral Singular

Nesta secção descrevemos resumidamente o *modus operandi* da AES. Para uma introdução a esta técnica pode consultar-se por exemplo Golyandina et al. (2001). Conforme referido anteriormente, a AES pode ser dissociada em duas fases, nomeadamente a **decomposição** e a **reconstrução**. Cada uma destas fases inclui dois passos. Comecemos pela fase da decomposição. Esta inclui os passos do embutimento e da decomposição em valores singulares, os quais introduzimos de seguida.

Embutimento

Este é o passo introdutório do método. O conceito central associado a este passo é dado pela matriz de trajectórias, ou seja, uma colecção de vários desfazamentos sucessivos de uma fragmento da série original $\mathbf{y} = [y_1 \cdots y_n]'$. Formalmente, definimos a matriz de trajectórias do seguinte modo

$$\mathbf{Y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_k \\ y_2 & & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_l & y_{1+1} & \cdots & y_{1+(k-1)} \end{bmatrix},$$

onde k é escolhido por forma a incluirmos na matriz $\mathbf{Y}$ todas as observações da série original, ou seja, $k = n - l + 1$.

De modo a fixar terminologia, doravante iremos denominar cada vector da forma $\mathbf{y}_{i,l} = [y_l \cdots y_{l+(i-1)}]'$, como uma janela de comprimento l. O parâmetro l deverá ser definido pelo utilizador (Hassani et al., 2009). Note-se que $\mathbf{Y}$ é uma matriz Hankel, estando a série original contida na junção formada pela primeira coluna e a última linha. Pode também ser vantajoso, do ponto de vista conceptual, pensar na matriz de trajectórias como uma sequência de k janelas, ou seja, $\mathbf{Y} = [\ \mathbf{y}_{1,l} \cdots \mathbf{y}_{k,l}]$.

Decomposição em Valores Singulares

No segundo passo é realizada uma decomposição em valores singulares (doravante DVS). Assim, a partir de uma análise da matriz $\mathbf{YY}'$ obtêm-se os valores próprios $\lambda_1 \geq ... \geq \lambda_d$, onde $d = car(\mathbf{YY}')$, bem como os correspondentes vectores singulares esquerdos ($\mathbf{w}_i$) e direitos ($\mathbf{v}_i$). Deste modo é possível rescrever a matriz de trajectórias como

$$\mathbf{Y} = \sum_{i=1}^{d} \sqrt{\lambda_i}\, \mathbf{w}_i \mathbf{v}_i' ,$$

sendo evidente a semelhança com a decomposição de Karhunen-Loève.

De seguida iremos focar a nossa atenção na segunda fase do método – a reconstrução. Esta fase inclui os passos do agrupamento e da diagonalização por médias.

Agrupamento

No passo denominado de agrupamento, ocorre a selecção das m componentes principais. Considere-se $I = \{1, \cdots, m\}$ e $I^c = \{m+1, \cdots, d\}$. Essencialmente, a ideia central do agrupamento passa pela escolha proficiente dos m primeiros triplos próprios associados ao sinal e exclusão dos remanescentes ($d - m$) associados ao ruído. Dito de outro modo, neste passo pretende-se realizar uma selecção apropriada do conjunto I , de modo a decompor a série $\mathbf{Y}$ em

$$\mathbf{Y} = \sum_{i \in I} \sqrt{\lambda_i}\, \mathbf{w}_i \mathbf{v}_i' + \varepsilon ,$$

onde ε designa uma componente de erro, e a parte remanescente representa o sinal. Na prática, o agrupamento pode ser conduzido utilizando, por exemplo, métodos rea- justados para a seleção de um número comedido de componentes principais m.

Diagonalização por Médias

Neste passo pretendemos obter a reconstrução da componente determinística da série, ou seja, o sinal. Um modo natural de cumprirmos este objectivo passa por transformar a matriz $\mathbf{Y} - \varepsilon$, obtida no passo anterior, numa matriz Hankel. A ideia aqui será então reverter o processo realizado até agora, voltando a obter uma variante reconstruída da matriz de trajectórias. Formalmente, iremos considerar o espaço $M_{l,k}$, formado pelas matrizes de dimensão ($l \times k$), e denotamos a base canónica de R^n por $\{\mathbf{h}_l\}_{l=1}^{n}$. Além disso definimos $\mathbf{X} = [x_{i,j}] \in M_{l,k}$. O procedimento de diagonalização por médias é realizado pela aplicação $\overline{D} : M_{l,k} \to R^n$, a qual é definida do seguinte modo

$$\overline{D}(\mathbf{X}) = \sum_{w=2}^{k+l} \mathbf{h}_{w-1} \sum_{(i,j)\in A_w} \frac{x_{i,j}}{|A_w|},$$

A notação | . | é aqui usada para representar o operador cardinal, e

$$A_w = \{(i,j) : i + j = w\},$$

para i = 1,..., l, e j = 1,..., k. Estamos agora em condições de escrever a componente determinística da série através do procedimento de diagonalização por médias descrito acima. Deste modo, a série reconstruída é dada por

$$\overline{\mathbf{y}} = \overline{D}\left(\sum_{i\in I} \sqrt{\lambda_i}\, \mathbf{w}_i \mathbf{v}_i'\right).$$

3.2 Método de Previsão Recorrente

Esta secção é dedicada à descrição do método recorrente de previsão (*recurrent forecast algorithm*) (Golyandina et al., 2001). O método parte do pressuposto de que é possível escrever a i-ésima observação y_i como combinação linear das (l-1) observações anteriores. Formalmente, consideramos que é válida a seguinte fórmula linear recurrente

$$y_i = a_1 y_{i-1} + a_2 y_{i-2} + \cdots + a_{l-1} y_{i-(l-1)},\ i \geq l$$

mediante uma escolha adequada de coeficientes de previsão $\mathbf{a} = [a_1 \cdots a_{l-1}]'$. A escolha de coeficientes de previsão é discutida em detalhe em Golyandina et al. (2001) e Golyandina e Osipov (2007). A previsão 1 passo à frente fará então uso dos (l-1) valores reconstruídos da série, isto é

$$\hat{y}_{n+1} = \sum_{i=1}^{l-1} a_i \tilde{y}_{(n+1)-i}\ .$$

A previsão 2 passos à frente é semelhante desde que sejam realizados os devidos ajustes. Esta previsão faz uso das últimas (l-2) reconstruções e da previsão para o período anterior. Mais especificiamente temos

$$\hat{y}_{n+1} = a_1 \hat{y}_{n+1} + \sum_{i=1}^{l-2} a_i \tilde{y}_{(n+1)-i}\ .$$

Este esquema de previsão pode ser facilmente generalizado por indução.

4. Implementação Computacional de Métodos de Análise Espectral Singular

De modo a ilustrarmos o funcionamento de alguns dos comandos do **package SSA** iremos examinar um célebre conjunto de dados considerado por Brown (1963). A série temporal é representada na Figura 1.

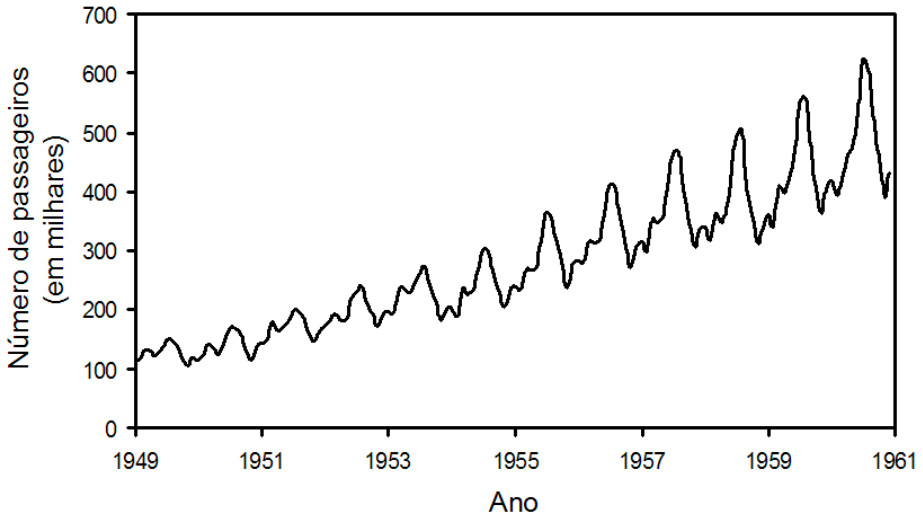


Figura 1 : Tráfego mensal de passageiros englobando diversas companhias aéreas internacionais

Esta série inclui 144 observações mensais referentes ao número total de passageiros (em milhares) num grupo de diversas companhias aéreas internacionais. Os dados são referentes ao período compreendido entre Janeiro de 1949 e Dezembro de 1960. Por motivos de completude, os dados estão contidos no package sendo a sua leitura realizada através dos comandos

```
> require(SSA)
> data(brown)
```

Conforme pode ser facilmente verificado a partir da análise da Figura 1 a série tem um forte movimento sazonal e uma tendência acentuada. Seguindo a sugestão de Golyandina e Osipov (2007) consideramos os parâmetros de decomposição e reconstrução *(l, m) = (36, 13)*. Por forma a ajustar o modelo basta considerar os seguintes comandos :

```
> l<-36
> m<-13
> mod<-ssa(brown,l,m)
> mod
SSA AdjustmentResults:
======================
ssa.default(y=passengers, l= l, m= m)
SeriesReconstruction:
[1]  112.1413 117.6243 134.1743 125.9842 122.0263 133.3661
     147.2968 146.2568 136.0309 115.4658 101.1044 117.6049
     119.8891 126.6296 144.1180 133.1830 133.7002 150.9362
[19] 167.6170 169.3653 155.6285 135.0246 120.1678 139.4706
 ...
```

Por forma a obtermos os resíduos do modelo basta considerar a linha de comandos

```
> resid(mod)
```

Os valores ajustados do modelo podem ser guardados numa variável auxiliar através do comando

```
> reconstruction<-fitted(mod)
```

Na Figura 2 representamos os valores ajustados do modelo e respectivos resíduos

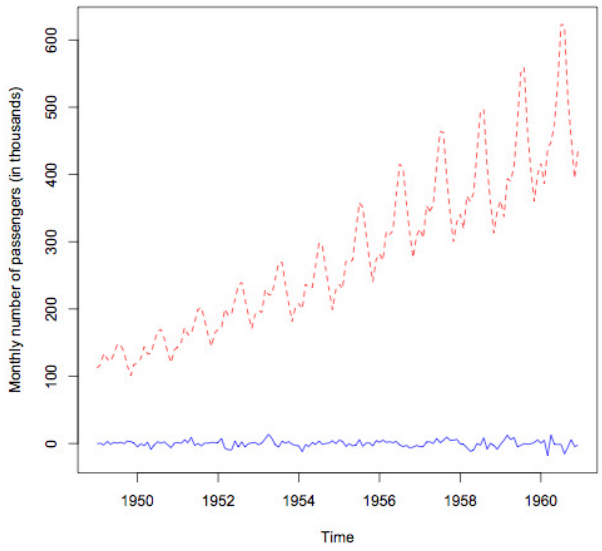


Figura 2 : Reconstrução (linha a tracejado) versus os resíduos do modelo (lina a cheio)

Outros outputs gráficos do modelo podem ainda ser obtidos através do comando `Dbar`, que permite a condução do passo de diagonalização por médias referido anteriormente. Por exemplo na Figura 2 representamos as 13 componentes principais da série temporal sob análise.

Os movimentos sazonais e a presença de tendência na série estão patentes na reconstrução que se encontra na Figura 2. Além disso estas características da série podem ainda ser destrinçadas mais detalhadamente na Figura 3. Por exemplo, a componente principal 1 (PC1) descreve claramente os movimentos mais pronunciados da tendência da série. As componentes principais com movimentos de carácter mais ciclíco (por exemplo PC2) estão claramente associadas à sazonalidade da série. Estas componentes são frequentemente interpretadas com o auxílio a periodogramas. Essencialmente um periodograma é um diagrama de dispersão em que se representa uma componente *versus* a componente seguinte. As componentes 12 e 13 parecem estar nitidamente associadas a ruído ou espúria.

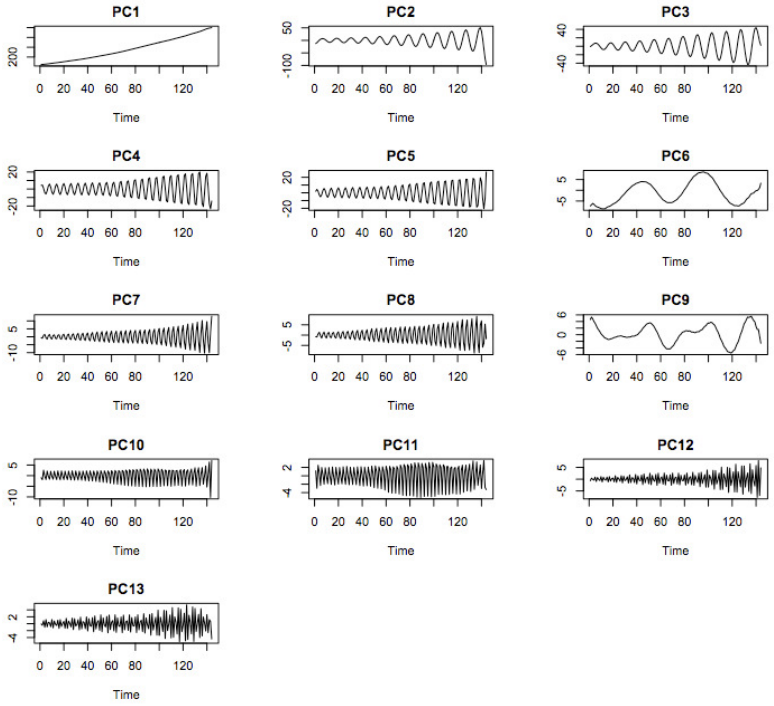


Figura 3 : As 13 primeiras components principais do modelo representadas como séries temporais

É ainda possível através do **package SSA** conduzir exercícios de previsão (*out-of-sample forecast*). Por exemplo, por forma a fazer previsão a 1 ano, através do *recurrent forecast algorithm*, basta introduzir os seguintes comandos

```
> data(brown)
> l<-36
> m<-13
> #número de passos à frente
> N<-12
> mod<-ssa(brown,l,m)
> f<-predict(mod,N)
```

No object `f` são armazenadas várias características de interesse, nomeadamente

- série original
- reconstrução
- previsões
- coeficientes associados à previsão

Por forma a obter as previsões basta invocar o comando

```
> f$forecast
[1] 460.6431 416.8169 475.3071 488.3577 521.6165 590.6237
    687.9366 694.4050 566.7915 500.4668 431.9732 480.0941
```

De modo semelhante podemos obter os coeficientes associados à previsão através do comando

```
> f$coefficients
[1] 0.205304936 0.069295346 0.179630801 0.067935233
    0.156785844 0.007025028 0.133941890 0.002672591
    ...
```

Podemos ainda obter um gráfico com as previsões e o gráficos dos coeficientes de previsão através da seguinte instrução

```
> plot(f)
```

O gráfico correspondente é representado na Figura 4. É interessante observar que a periodicidade da série é facilmente capturada na representação dos coeficientes de previsão. Este aspecto é mais pronunciado nos picos referentes às previsões 12 e 24 (com coeficientes de previsão respectivos de $a12$=0.59 e $a24$=0.48)

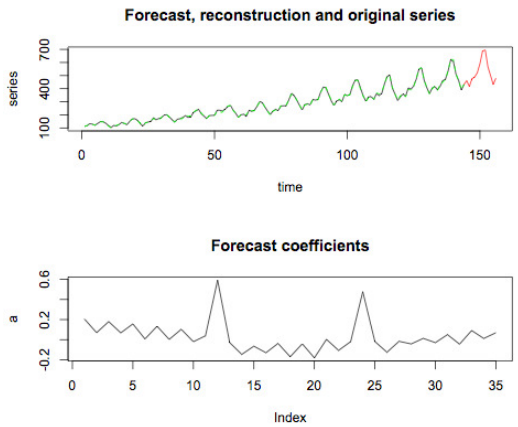


Figura 4 : Representação gráfica da série, reconstrução e das previsões (em cima) e dos coeficientes de previsão **a** (em baixo)

4. Referências

Allen, M. R. & Smith, L. A. (1996) Monte Carlo SSA: detecting irregular oscillations in the presence of colored noise. *Journal of Climate* 9, 3373–3404.

Basilevsky, A. & Hum, D. (1979) Karhunen Loève analysis of historical time series with an application to plantation births in Jamaica. *Journal of the American Statistical Association* 74, 284–290.

Broomhead, D. S. & King, G. P. (1986) Extracting qualitative dynamics from experimental data. *Physica D* 20, 217–236.

Brown, R. G. (1963) Smoothing, forecasting and prediction of discrete time series. Prentice-Hall, New Jersey.

de Carvalho, M. & Rodrigues, P. C. (2009) Método de imputação recorrente : análise espectral singular com valores omissos. XVII Congresso Nacional da Sociedade Portuguesa de Estatística.

Ghil, M. & Vautard, R. (1991) Interdecadal oscillations and the warming trend in global temperature time series. *Nature* 350, 324–327.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F. & Yiou, P. (2002) Advanced spectral methods for climatic time series. *Review of Geophysics* 40, 1–41.

Golyandina, N. & Osipov E. (2007) The "Catterpillar"-SSA method for analysis of time series with missing values. *Journal of Statistical Planning and Inference* 137, 2642–2653.

Golyandina, N., Nekrutkin, V. & Zhigljavsky, A. (2001) *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC, London.

Hassani, H., Heravi, H. & Zhigljavsky, A. (2009) Forecasting European industrial production with singular spectrum analysis. *International Journal of Forecasting* 25, 103–118.

Hochstadt, H. (1989) *Integral Equations*. Wiley, New York.

Loève, M. (1978) *Probability Theory II*. 4th Edition. Springer Verlag, New York.

Kondrashov, D. & Ghil, M. (2006) Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics* 13, 151–159.

Paegle, J. N., Byerle, L. A. & Mo, K. C. (2000) Intraseasonal modulation of south American summer precipitation. *Monthly Weather Review* 128, 837–850.

Vautard, R. & Ghil, M. (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D* 35, 395–424.

Vautard, R. Yiou, P. & Ghil, M. (1992) Singular spectrum analysis: a toolkit for short noisy chaotic signals. *Physica D* 58, 95–126.

Gráficos em R – uma breve introdução

Conceição Amado e Ana M. Pires, *{camado,apires}@math.ist.utl.pt*

Departamento de Matemática e CEMAT, IST, UTL, Lisboa

O R permite obter muitos tipos de gráficos, quer usando funções pré-definidas, tais como as funções gráficas tradicionais `plot()`, `hist()` e `barplot()`, quer criando novas funções.
Coloque-se no espaço de trabalho do R os dados `ConsumoCar.data`[1] e aplique-se a função `attach` (com esta função a informação relativa a estes dados fica disponível no espaço corrente de trabalho). Como exemplo, aplique-se a função `plot()` para criar um diagrama de dispersão (ou gráfico de pontos). Uma janela gráfica do R (R Graphics Device ou RGD) abre para se visualizar o gráfico.[2]

```
>car<-read.table("http://www.math.ist.utl.pt/~camado/SPE/R/ConsumoCar.txt",
sep ="\t",header=TRUE) #leitura dos dados usando ligação internet
>attach(car)
>plot(Desloc.,Combustível)
```

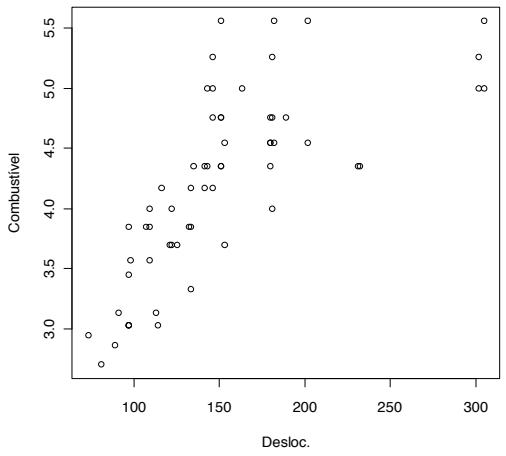


Figura 1: Exemplo de um diagrama de dispersão usando a função `plot()`.

O resultado que é apresentado na Figura 1 permite constatar que, por defeito, o tipo de gráfico é apenas de pontos. Mas como argumento da função `plot()` pode ser escolhido outro tipo, por exemplo, linhas (`type="l"`), aparência de histograma (type="h") ou em escada (`type="s"`), etc. Também os eixos, escalas, títulos, símbolos a desenhar e cores podem ser escolhidos usando os parâmetros `pch=`, `col=`, `lty=`, `ylab=`, etc., nas funções gráficas, como será ilustrado mais adiante. Para descrição detalhada de

1 Disponíveis em http://www.math.ist.utl.pt/~camado/SPE/R/.
2 Nesta exposição usou-se a versão 2.9.2 do R.

todos os parâmetros gráficos veja-se a função `par()`. As cores (especificadas usando o parâmetro `col=)` dependem do mapa de cores do dispositivo gráfico da máquina.

Para visualizar os nomes das cores disponíveis em R faça-se:

```
> colors()                          # Inglês americano ou
> colours()                         # Inglês britânico
```

Para mais detalhes sobre as cores no R consultar, por exemplo: `http://research.stowers-institute.org/efg/R/Color/Chart/.`

Diagrama de pontos ou diagrama de dispersão permite comparar dois ou mais conjuntos de dados quantitativos desenhando um número finito de pontos (observações) num espaço definido por duas escalas, os eixos x e y, respectivamente. Estes gráficos podem ser construídos de diversas formas em R, a forma mais simples é a que foi apresentada no início usando a função `plot()`. O gráfico da Figura 1 pode também ser construído usando a função `scatterplot()` da biblioteca `car` . Esta função é mais completa podendo mostrar ainda linhas de ajustamentos, diagramas marginais em caixa e outras características.

```
> library(car)
> scatterplot(Desloc.,Combustível)
```

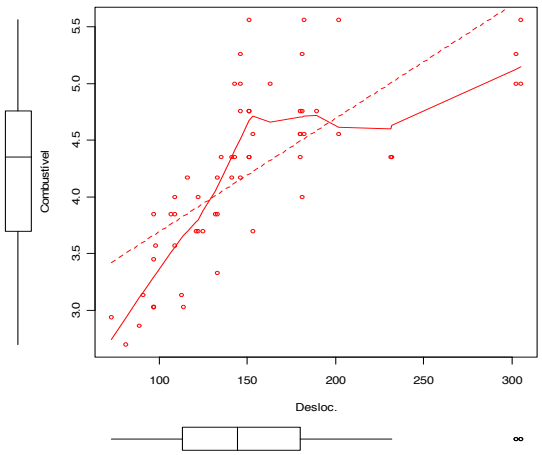


Figura 2: Exemplo de um diagrama de dispersão usando a função `scatterplot()`.

Diagramas circulares são muitas vezes usados para mostrar as proporções relativas dos diferentes valores de um conjunto de dados (mais especificamente, a frequência dos níveis de uma variável categórica). Por exemplo, a informação sobre o número de mandatos que os vários partidos políticos obtiveram nas Eleições Legislativas de 2009 em Portugal, pode mostrar-se num gráfico circular. O código R poderá ser:

```
> ele09<-read.table("http://www.math.ist.utl.pt/~camado/SPE/R/Legis09.txt", sep =
"\t",header=TRUE)
> attach(ele09)
> Mand09<-na.exclude(ele09$Mandatos[ele09[,3]>0]) #exclui partidos sem mandatos
> Part09<- na.exclude(ele09$Partido[ele09[,3]>0]) #exclui partidos sem mandatos
> labels09<-sprintf("%s (%d)", Part09, Mand09)
> pie(Mand09,
 col=c("deeppink","orange","blue","black","red"),
 labels=labels09, main="Resultados Eleições Legislativas 2009\n(Mandatos por Partido)")
```

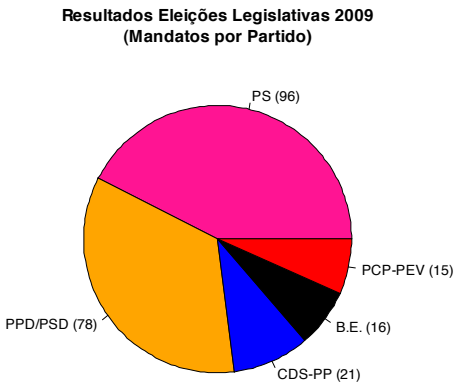


Figura 3 : Exemplo de um diagrama circular.

Construa-se agora este gráfico em 3D:

```
> library(plotrix)
>    pie3D(Mand,    col=c("deeppink","orange","blue","black","red"),labels=    labels09,
main="Resultados Eleições Legislativas 2009\n(Mandatos por Partido)",explode=0.1)
```

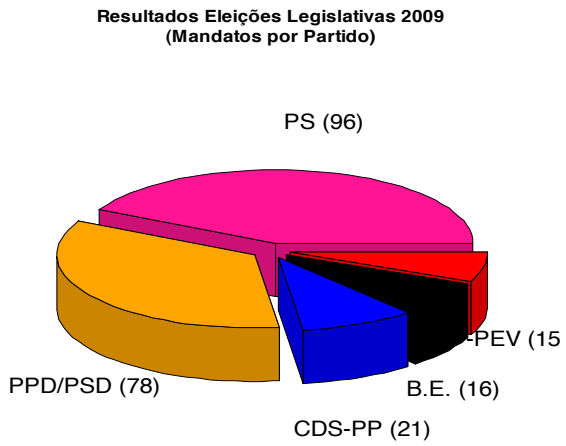


Figura 4: Diagrama circular em 3D.

Apesar de estes gráficos serem populares nos media, os diagramas circulares não são muito populares entre os estatísticos. Consultando a página de ajuda do R do comando `pie()` fazendo, por exemplo, `?pie`, existe a seguinte chamada de atenção na secção "Notes": "*Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data*."

Diagrama de barras é outra forma de representar estes dados. As barras podem estar orientadas na horizontal ou vertical. O comando `barplot()`difere ligeiramente do `pie`. A principal diferença reside no uso de `names.arg` em vez do `labels` para representar os níveis da variável. (Notar que em muitos casos é necessário converter a variável para tipo `character`, caso contrário será tratada como factor).

O principal problema com os gráficos de barras é a visualização de nomes dos níveis da variáveis, pois podem ser demasiado extensos e então o R omite-os. Existem várias soluções para resolver este problema, a mais simples é diminuir o tamanho da fonte da letra usando o comando `cex.names`, o

qual fixa o tamanho da fonte. No gráfico presente diminuiu-se para 80% (0.8) o tamanho que está por defeito.

```
>barplot(Mandatos[-c(17,18,19,20)],
 col=c("deeppink","orange","blue","black","red"),
 names.arg = Partido[-c(17,18,19,20)], main="Resultados Eleições Legislativas
2009\n(Mandatos por Partido)")

> barplot(Mandatos[-c(17,18,19,20)],
 col=c("deeppink","orange","blue","black","red"),
 names.arg   =Partido[-c(17,18,19,20)],   main="Resultados   Eleições   Legislativas
2009\n(Mandatos   por   Partido)",   cex.names=0.8,xlab="Partidos",   ylab="Número   de
Mandatos",ylim= c(0,100))
```

Esta última instrução produz o seguinte gráfico:

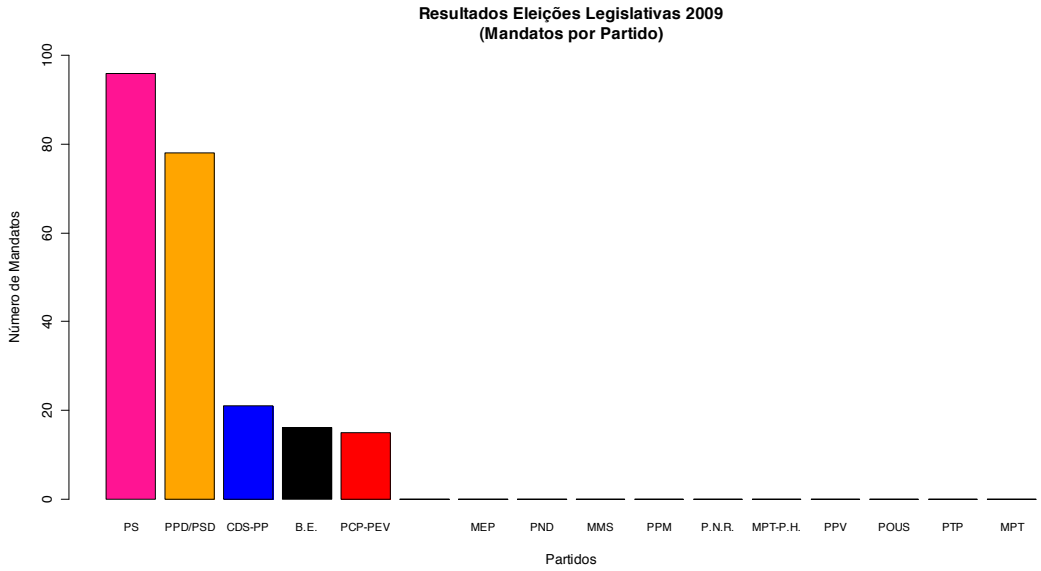


Figura 5: Exemplo de diagrama de barras.

Considere agora que se pretende comparar os resultados das eleições de 2005 com as de 2009, em termos de mandatos, para partidos que conseguiram algum mandato. O código apresenta-se de seguida

```
> ele05<-read.table("http://www.math.ist.utl.pt/~camado/SPE/R/Legis05.txt", sep =
"\t",header=TRUE)
> Mand05<-na.exclude(ele05$Mandatos[ele05[,3]>0]) #exclui partidos sem mandatos
> Mand_0509<-rbind(Mand09,Mand05)
> barplot(Mand_0509,col=c("deeppink","deeppink","orange","orange","blue",
"blue","black","black","red","red"),density=c(150,50),names.arg =Part09,
main="Resultados Eleições Legislativas 2009 e 2005\n(Mandatos por Partido)",
beside=TRUE)
> legend(12.5, 115, c("2009","2005"), density=c(150,50), fill=c("gray","gray"))
#adiciona uma legenda ao gráfico
```

e o gráfico encontra-se na Figura 6. Este código (devidamente modificado) permite construir histogramas múltiplos.

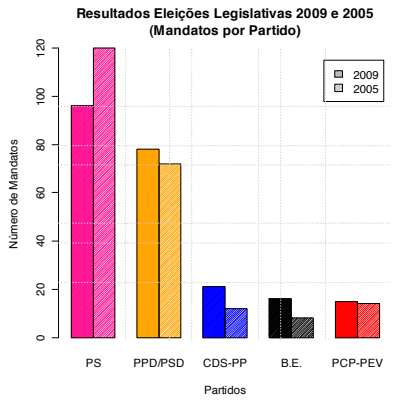


Figura 6: Diagrama de barras múltiplo.

Histogramas: o Teorema do Limite Central (TLC) é muitas vezes ilustrado usando a soma de variáveis aleatórias com distribuição assimétrica, mostrando que essa soma converge para uma distribuição simétrica (a normal) à medida que o número de variáveis da soma aumenta. Construa-se a função seguinte:

```
SimulSomaExp<-function(n,obs=10000){
 mu<-1
 sig<-1
 teste<-matrix(0,obs,n)
 x<-seq(-4,4,0.01)
 for (j in 1:n) {
  teste[,j]<-rexp(obs,1)
  res<-apply(teste,1,sum)
  resSt<-(res-j*mu)/(sig*sqrt(j))
  hist(resSt,prob=TRUE,xlim=c(-4,4),ylim=c(0,0.4),
     xlab="Soma St.",ylab="Densidade",main="",col="lightgrey")
  title(main="Simulação da soma de n va. Exp(1)",line=3)
  title(sub=paste("n =",j,"Obs. =",obs),line=-24,cex.sub=1.0,font.sub=2)
  lines(x,dnorm(x),col=2,lwd=2)
  if (interactive()) {cat("\nCarregue em  <Return> para continuar: ")
              readline()}
  }
}
> SimulSomaExp(20) #executar a função com a soma de 1 a 20 variáveis com distribuição
exponencial
```

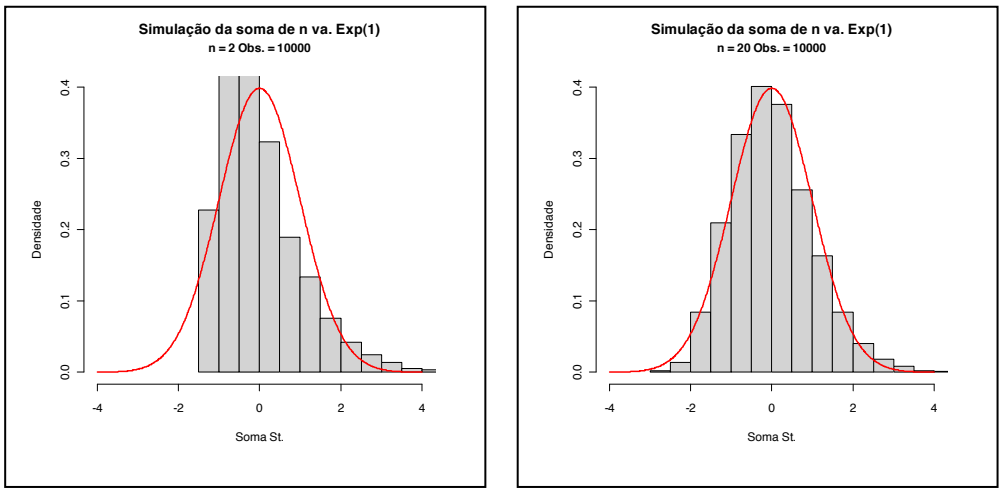


Figura 7: Histogramas relativos a 10000 observações e densidade da distribuição normal para ilustração do TLC.

Diagrama de caixa, *boxplot* ou caixa de Tukey: é um gráfico simples que é construído com apenas 5 pontos mas que possui muita informação. Mostra de uma forma clara a distribuição dos dados e as suas principais características (simetria, caudas, …). Permite comparar a distribuição de vários conjuntos de dados em simultâneo. Como exemplo construa-se o diagrama de caixas para dados simulados de uma distribuição normal de valor médio 10 e variância 10 e para uma variável exponencial de valor médio 5, ver figura 8.

```
> snorm<-rnorm(10000,10,sqrt(10))
> sexp5<-rexp(10000,1/5)
>boxplot(cbind(snorm,sexp5),names=c("Normal","Exponencial"),col=c("lightblue","blue"))
```

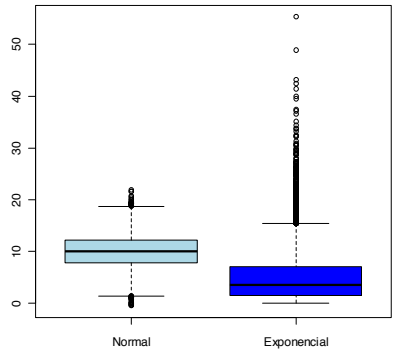


Figura 8: Exemplo de diagramas de caixa.

Acrescentar informação a um gráfico: construa-se o objecto `car.ajuste` resultado da análise de regressão linear e tendo a janela do gráfico da Figura 1 aberta faça-se:

```
> car.ajuste<-lm(Combustível~Desloc.)
> abline(car.ajuste$coef,lty=4)
> plot(Desloc., Combustível,pch=3)
> lines(Desloc., fitted(car.ajuste), col=6)
```

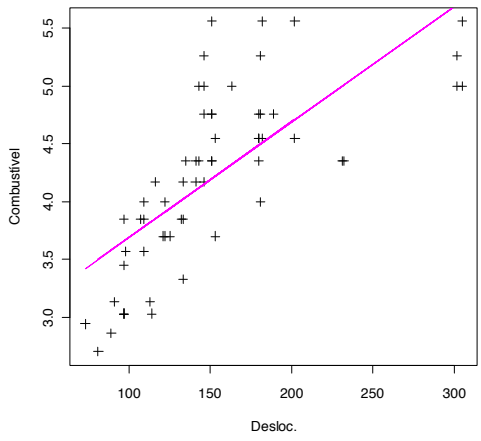


Figura 9: Exemplo da aplicação de várias funções e opções gráficas do R.

Vários exemplos podem ser dados, um outro, é a utilização da função `points()`, esta permite adicionar pontos a um gráfico em branco (onde os eixos e escala estão já presentes) ou a um gráfico já desenhado, fazendo:

```
> plot(Desloc., Combustível,type="n")
> points(Desloc., Combustível, col="green", pch=8)
```

As funções `lines()`, `points()`, `abline()`, e `text()`, como ilustrado acima, são funções que permitem usar e modificar várias características a um gráfico existente.

Para obter múltiplos gráficos numa mesma janela devem ser usadas as funções `mfrow()` ou `mfcol()`. Os seguintes exemplos ilustram a utilização de `mfrow()` e o resultado apresenta-se na figura 10.

```
> par(mfrow=c(2,2), mar=rep(4,4))
> plot(Desloc., Combustível, col=2, pch=4)
> lines(Desloc., car.ajuste $fitted, col=5, lty=4)
> qqnorm(car.ajuste $res, col=6,xlab="Quantis teóricos",ylab="Quantis empíricos")
> boxplot(car.ajuste $res, ylab="Resíduos")
> hist(car.ajuste $res, xlab="Resíduos",ylab=Frequência, main="Histograma dos Resíduos")
```

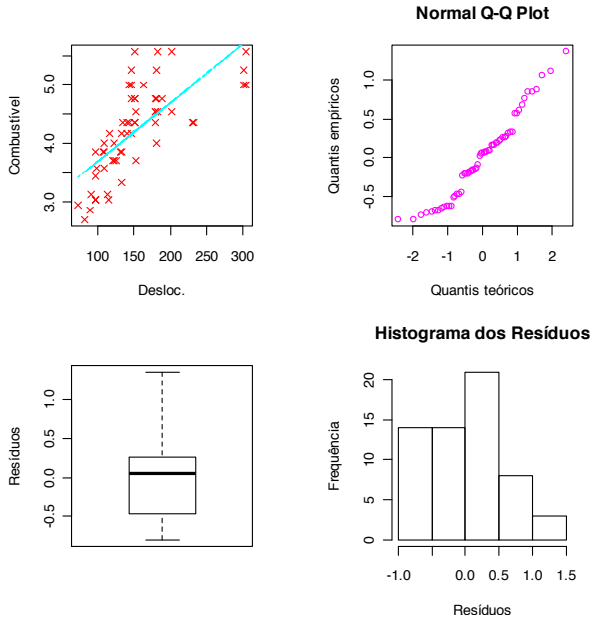


Figura 10: Exemplo de vários gráficos numa página.

Dados multivariados: para visualização de dados multivariados podem usar-se funções do R que permitem construir matrizes de gráficos (*scatterplot*, *matplots*), gráficos de estrelas (*star plots*) gráficos de faces (*Chernoff's faces*).

Para construir um *scatterplot matrix* basta fazer o comando

```
> iris.pt<-iris
> names(iris.pt)<-c("Comprim.Sépala", "Largura.Sépala" , "Comprim.Pétala",
"Largura.Pépala" , "Espécies" )
> pairs(iris.pt[1:4], main = "Dados Iris de Anderson",pch = 21, bg =
c("darkorange", "darkolivegreen2", "darkorchid1")[unclass(iris.pt$Espécies)])
```

surgindo o gráfico ilustrado à esquerda na Figura 11. Este tipo de gráfico apresenta os diagramas de dispersão entre todos os pares de variáveis num conjunto de dados multivariado. Na Figura 11, à direita, encontra-se um gráfico semelhante mas onde se visualiza o histograma associado às observações de cada variável, o código R que o gerou é o seguinte:

```
fun.hist <- function(x, ...)
{
usr <- par("usr"); on.exit(par(usr))
par(usr = c(usr[1:2], 0, 1.5) )
h <- hist(x, plot = FALSE)
breaks <- h$breaks;
nB <- length(breaks)
y <- h$counts; y <- y/max(y)
rect(breaks[-nB], 0, breaks[-1], y, col="yellow1", ...)
}

> pairs(iris.pt[1:4], main = " Matriz de diagramas de dispersão com histogramas\n Dados
Íris ",pch = 21, bg = c("darkorange", "darkolivegreen2",
"darkorchid1")[unclass(iris.pt$Espécies)],
diag.panel=fun.hist)
```

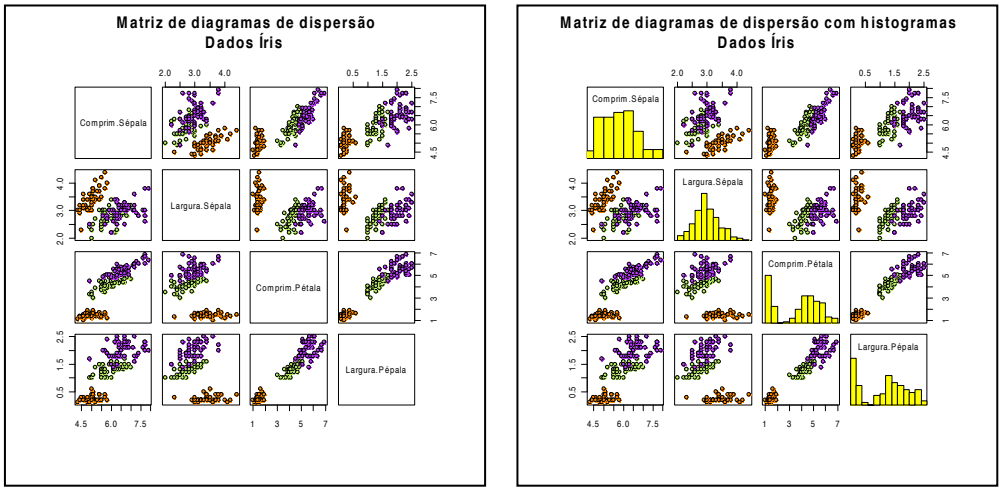


Figura 11: Exemplos de matrizes de diagramas de dispersão.

Pode ainda usar-se a função `matplot()` para representar num único gráfico as observações que correspondem a vários diagramas de dispersão. Os comandos seguintes ilustram esta aplicação e o resultado pode ser visualizado na Figura 12.

```
> table(iris$Species)
> iS <- iris$Species == "setosa"
> iV <- iris$Species == "versicolor"
> matplot(c(1, 8), c(0, 4.5), type= "n", xlab = "Comprimento", ylab = "Largura", main =
"Dimensões da Pétala e da Sépala nas flores Iris")
> matpoints(iris[iS,c(1,3)], iris[iS,c(2,4)], pch = "sS", col = c(2,4))
> matpoints(iris[iV,c(1,3)], iris[iV,c(2,4)], pch = "vV", col = c(2,4))
> legend(1, 4, c("Setosa Pétalas", "Setosa Sépalas","Versicolor Pétalas","Versicolor
Sépalas"),pch = "sSvV", col = rep(c(2,4), 2))
```

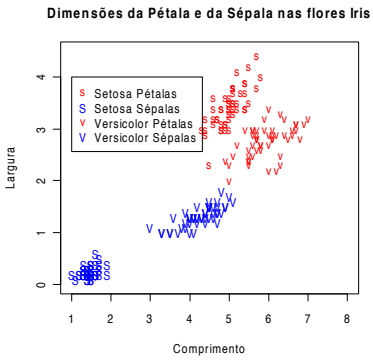


Figura 12: Exemplo de aplicação da função **`matplot`()**.

Outras representações gráficas interessantes são os gráficos de estrelas e as faces/caras de Chernoff. No primeiro, cada estrela representa um caso (uma linha da matriz de dados, um objecto) e cada ponta da estrela representa uma variável particular, ou coluna. Quer o tamanho, quer a forma de cada estrela têm significado, o tamanho reflecte a magnitude total do ponto e a forma revela as relações entre as variáveis. Comparando duas estrelas pode verificar-se de forma rápida as similaridades ou não entre dois casos (objectos) - estrelas de forma semelhante indicam casos similares. Para criar um gráfico de estrelas basta[3] executar a função do R `stars()` como exemplificado no código seguinte e ilustrado na figura 13.

```
> UE27<-read.table("Europa.txt",header=TRUE,sep="\t")
> stars(UE27,labels=UE27$País,key.labels =abbreviate(colnames(UE27[,2:6]),
minlength = 5),key.loc=c(12.5,2),col.stars=rainbow(27))
```

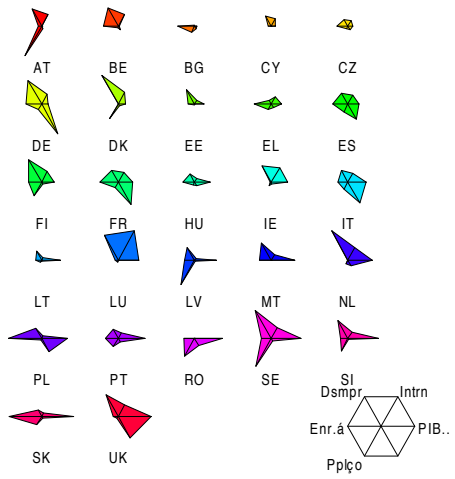


Figura 13: Exemplo de gráfico de estrelas.

As caras ou faces de Chernoff são também uma forma de representar dados multivariados. Cada variável, numa dada observação, é associada a uma característica da cara. Dois casos (objectos) podem ser comparados fazendo a comparação das diversas características. A função do R para criar as caras de Chernoff (disponível na biblioteca do R TeachingDemos) é:

```
> library(TeachingDemos)
> faces(UE27[,2:6],scale=TRUE)
```

[3] Os dados Europa.txt encontram-se disponíveis em `http://www.math.ist.utl.pt/~camado/SPE/R/`

E a visualização é a apresentada na Figura 14.

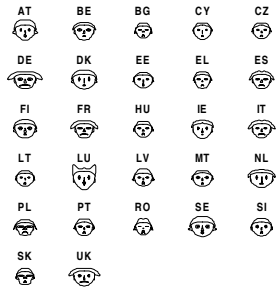


Figura 14: Exemplo de gráfico de faces de Chernoff.

Uma das características mais poderosas das funções gráficas do R é efectuarem gráficos de objectos genéricos do R. Ou seja, a função reconhece um objecto R e este é usado como argumento de uma função gráfica produzindo um conjunto de gráficos que, de algum modo, representam esse objecto, por exemplo os dendrogramas na análise de *clusters*, a análise de diagnóstico nos modelos de regressão ou simplesmente se um objecto tem uma variável-x que é um factor então a função gráfica plot() constrói diagramas em caixa da variável-y (eg. `plot(iris[,5],iris[,1])`).

Na *R Graph Galley*[4] encontra-se um leque de gráficos construídos em R. Demonstrações interessante também podem ser visualizadas fazendo, na linha de comandos de uma sessão de R, os comandos: `demo(image)`, `demo(graphics)`, `demo(persp)`, `demo(lattice)`.

Como se pode verificar, os princípios de construção de gráficos em R não são difíceis. Para uma informação mais detalhada, veja-se a secção Graphics do manual oficial de introdução ao R.

R Essencial

Tipos de funções gráficas

De nível elevado: funções tais como `plot`, `hist`, `boxplot`, ou `pairs` que produzem um gráfico ou inicializam um.

De baixo nível: funções que adicionam informação a um gráfico existente, criado com uma função de nível elevado. Os exemplos são `points`, `lines`, `text`, `axis`.

Funções *trellis*: funções tais como `xyplot`, `bwplot`, ou `histogram`, que podem produzir um conjunto de gráficos numa única chamada.

Guardar gráficos como imagens externas

O R suporta uma variedade de formatos de imagem. Para verificar as opções disponíveis, escreva-se na linha de comandos `?device`. As opções incluem o postscript, o pdf, o png e o JPEG (entre outros). É mesmo possível gerar comandos para construir um gráfico em LaTeX.

[4] http://addictedtor.free.fr/graphiques/thumbs.php

De seguida apresenta-se a forma de exportar um gráfico para JPEG. [5]

```
> UE27<-read.table("Europa.txt",header=TRUE,sep="\t")
> jpeg("estrelasUE27.jpg")
> stars(UE27,labels=UE27$País,key.labels =abbreviate(colnames(UE27[,2:6]),
minlength = 5),key.loc=c(12.5,2),col.stars=rainbow(27))
> dev.off()
> q()
```

Antes de terminar o R, é importante fechar o ficheiro com o comando `dev.off()`, assegurando que o R não continua a escrever nele em sessões futuras.

R Commander

O R não incorpora uma interface gráfica para interacção com o utilizador (GUI), mas inclui ferramentas para as construir. A biblioteca `Rcmdr` fornece um GUI (`R Commander`) que permite, ao utilizador, efectuar cálculos básicos de estatística e gráficos utilizando opções apresentadas nos menus.

[5] Notar que, nalguns sistemas operativos é possível realizar esta operação directamente a partir do menu.

PAM: Um pacote estatístico para estimar (bio)diversidade (e algo mais) através de modelos poissonianos de abundâncias

Nuno Sepúlveda, *nunosep@gmail.com*

Instituto Gulbenkian de Ciência
Centro de Estatística e Aplicações da Universidade de Lisboa

Carlos Daniel Paulino, *dpaulino@math.ist.utl.pt*
Departamento de Matemática, Instituto Superior Técnico
Centro de Estatística e Aplicações da Universidade de Lisboa

Jorge Carneiro, *jcarneir@igc.gulbenkian.pt*
Instituto Gulbenkian de Ciência

1. Introdução

A estimação de diversidade de objectos distintos numa população (ou num conjunto) é um problema transversal a diversas áreas científicas, mas que assume especial importância em Ecologia, onde a monitorização da biodiversidade – aqui vista como o número de espécies a residir numa certa comunidade ecológica – visa não só compreender os fenómenos actuantes num determinado habitat, mas também servir de suporte à tomada de decisões ao nível de politicas ambientais. Diversas abordagens têm sido propostas para responder a tal desafio inferencial, que vão desde a utilização de modelos paramétricos a estimadores não-paramétricos, até ao cálculo de índices que tentam capturar o grau de diversidade de uma determinada amostra (ou população). Recentemente, propuseram-se os chamados modelos poissonianos de abundâncias para dar resposta ao problema de estimação da diversidade de receptores dos linfócitos T (Sepúlveda, 2009; Sepúlveda *et al.*, 2008, 2010), tendo-se desenvolvido o pacote PAM - sigla do inglês *Poisson Abundance Models* – a usar no ambiente R.

Na sequência do último boletim SPE, onde se deu voz a vários "amantes" do programa R, este artigo tem como objectivo dar a conhecer as principais funcionalidades do pacote PAM a futuros utilizadores do mesmo. Para efeitos de ilustração, apresenta-se um exemplo de estimação de (bio)diversidade molecular dos receptores de linfócitos T $CD4^+$ e $CD8^+$ provenientes de nódulos linfáticos, locais onde se dão as respostas imunitárias (Sepúlveda, 2009; Sepúlveda *et al.*, 2010).

2. Inferências sobre modelos poissonianos de abundâncias

Os dados usados para estimar a diversidade de uma população de objectos referem-se aos diferentes tipos de objectos amostrados e sua respectiva abundância. Estes dados são usualmente apresentados na forma de uma distribuição de frequências $m = \{m_i, i = 1, \ldots, n\}$ onde m_i representa o número de tipos de objectos com *i* representantes na amostra. Neste cenário, $M = \sum_{i=1}^{n} m_i$ é a diversidade de objectos na amostra e $n = \sum_{i=1}^{n} i \times m_i$ é o tamanho da amostra. A distribuição amostral de $\{m_i\}$ é tipicamente descrita pela seguinte lei Multinomial

$$f(\{m_i\} \mid D, \eta) = \frac{D!}{(D-M)!m_1! \ldots m_n!} \left[p_\eta(0)\right]^{D-M} \prod_{i=1}^{n} \left[p_\eta(i)\right]^{m_i}, \qquad (1)$$

onde D é o número de diferentes tipos de objectos presentes na população (diversidade populacional), $p_\eta(i)$ é a probabilidade de um determinado tipo de objecto ter i representantes na amostra, probabilidade essa descrita por um modelo com vector de parâmetros η. O cerne da análise reside em utilizar diferentes distribuições de probabilidade para $p_\eta(i)$ de forma a estimar a diversidade populacional D com a maior precisão possível. Neste contexto, a classe dos modelos poissonianos de abundâncias tem sido uma boa aposta para modelar as probabilidades $p_\eta(i)$.

Esta classe de modelos estatísticos postula, numa primeira instância, uma distribuição de Poisson com taxa de amostragem λ para o número de representantes na amostra de um determinado tipo de objecto. Se se supuser que todos os tipos de objectos estão igualmente representados na população, está-se perante o modelo de Poisson homogéneo. Heterogeneidade na representatividade de cada objecto distinto na população é incluída nestes modelos através da imposição de uma distribuição de probabilidade para a taxa de amostragem λ. Existem várias propostas para esta distribuição geradora de vários modelos de mistura, sendo as mais populares a Exponencial (modelo Geométrico), a Gama (modelo Poisson-Gama), a distribuição Lognormal (modelo Poisson-Lognormal) e uma mistura apropriada de distribuições Exponenciais, que dá origem ao modelo de Yule, pertencendo também à classe de distribuições em função potência. Estes modelos, para além da sua simplicidade probabilística, apresentam a vantagem de poderem ser justificados por mecanismos de geração e manutenção da diversidade dos vários tipos de objectos e da sua respectiva abundância na população em estudo; veja-se, por exemplo, o trabalho de Diserud e Engen (2000) onde se justificam as distribuições Poisson-Gama e Poisson-Lognormal no âmbito do estudo da diversidade em comunidades ecológicas através de modelos dinâmicos estocásticos. Outras distribuições, tal como a Log-Cauchy (Yina et al., 2005) ou a Zeta (Sepúlveda et al., 2008), foram ainda propostas para modelar $p_\eta(i)$, por enquanto fornecendo apenas descrições estatísticas dos dados por falta de conhecimento de algum mecanismo que as possa gerar.

A estimação da diversidade D e do vector de parâmetros η é feita através da maximização da função de verosimilhança (1). Em geral, os vários modelos apresentam funções de verosimilhanças log-côncavas em função de D e, por isso, pode-se usar o seguinte método de verosimilhança por perfil: (i) começa-se com $\hat{D} = M$, (ii) estima-se η por métodos tradicionais (numéricos ou não) de máxima verosimilhança; (iii) repete-se o passo anterior incrementando uma unidade à estimativa de D até que a função de verosimilhança deixe de crescer com D. Note-se que o modelo Poisson-Lognormal não possui uma expressão matemática em forma fechada. Contudo, Bulmer (1974) propôs um método "scoring" de Fisher que inclui integração numérica e uma aproximação para a cauda direita da distribuição de probabilidade Poisson-Lognormal.

Efectuada a fase de estimação dos modelos, há que prosseguir a análise com a respectiva avaliação da qualidade de ajustamento. Uma forma simples de atingir este objectivo, mas não isenta de alguma arbitrariedade, consiste em aplicar o conhecido teste de Pearson, em que se compara as frequências observadas com as suas contrapartidas esperadas sob um modelo através de uma estatística Qui-quadrado. Contudo, este teste não deve ser aplicado directamente à distribuição Multinomial (1) que, embora tenha sido usada para efeitos de estimação, inclui a diversidade não observada na amostra $(D-M)$ que, por isso, não deve ser contabilizada para efeitos de ajustamento do modelo. Para contornar este problema, é usual recorrer-se alternativamente à distribuição condicional de $\{m_i\}$ dada a diversidade observada M

$$f(\{m_i\} \mid M, \eta) = \frac{M!}{m_1! \ldots m_n!} \prod_{i=1}^{n} \left[\frac{p_\eta(i)}{1 - p_\eta(0)} \right]^{m_i} . \qquad (2)$$

Note-se que os modelos poissonianos de abundâncias descritos acima têm um número diferente de parâmetros. Neste cenário importa também comparar os vários modelos em termos da sua complexidade. Com esse intuito, é frequente proceder-se ao cálculo da conhecida medida de Informação de Akaike para cada modelo que se ajuste bem aos dados pelo teste de Pearson (ou outro). Para esse cálculo, é recomendado que também se use a função de verosimilhança condicional (2) em

vez da sua contrapartida não condicional (1), pelo mesmo motivo apresentado acima. O "melhor" modelo é, então, aquele que apresenta o maior valor desta medida.

3. Pacote PAM

O pacote PAM foi criado no âmbito do problema de estimação de diversidade de linfócitos T (Sepúlveda 2009; Sepúlveda *et al*., 2008, 2010), estando disponível no sítio da rede http://qobweb.igc.gulbenkian.pt/ no menu software, incluindo a informação sobre a sua instalação no ambiente R. Este contempla vários comandos escritos na linguagem R que permitem estimar a diversidade por algum índice (`diversity.index`) ou através da estimação de algum modelo poissoniano de abundâncias (`pam.fit`), testar a qualidade de ajustamento desse modelo (`gof.pam`), calcular a respectiva medida de Informação de Akaike (`aic.pam`) e estimar a distribuição teórica das abundâncias de cada tipo de objecto ao nível da população (`sad`); em Imunologia, fala-se em distribuições de tamanhos clonais, uma vez que os tipos de objectos são denominados por clones. Oferece também um comando para inferir a similaridade de um par de amostras através de algum índice de diversidade partilhada (`similarity.index`). Na secção seguinte exemplificar-se-á a utilização de alguns comandos acima referidos, dando especial enfoque à análise/estimação da diversidade.

4. Exemplo de aplicação

A aplicação que se segue serve de exemplo à utilização do próprio pacote, podendo ser carregado no ambiente R através do comando `data(tcr)`, depois da execução dos comandos `require(pam)` ou `library(pam)` para colocar o pacote em memória; uma breve descrição sobre o conjunto de dados pode ser obtida através da instrução `help(tcr)` ou, mais simplesmente, `?tcr`. O conjunto de dados é constituído por quatro variáveis (`lncd4`, `thycd4`, `lncd8` e `thycd8`) em que cada linha contém o número de células T $CD4^+$ e $CD8^+$ retirada dos nódulos linfáticos (`lncd4` e `lncd8`) e do timo (`thycd4` e `thycd8`), que apresentam um determinado tipo de receptor, definido pela sua respectiva sequência de aminoácidos. O objectivo original da análise consistiu em estimar a diversidade de receptores nas diversas populações celulares e inferir a respectiva distribuição de tamanhos clonais (Sepúlveda *et al*., 2010). Por motivos de espaço, ir-se-á restringir somente à análise da diversidade das células T $CD8^+$ dos nódulos linfáticos cujos dados observados $\{(i, m_i)\}$ se indicam a seguir (na 1ª e 2ª linhas, respectivamente):

```
R> table(tcr$lncd8[tcr$lncd8>0]
1    2   3   4   5  21  52
17   8   1   2   1   1   1
```

Uma análise mais completa destes dados pode ser encontrada em Sepúlveda (2009) ou em Sepúlveda *et al*. (2010).

Como uma primeira abordagem ao problema inferencial em causa, pode-se calcular algum índice de diversidade. A instrução `diversity.index` permite determinar o índice de Simpson ou a medida de entropia de Shannon através da especificação da opção `method` (`simpson` ou `entropy`, respectivamente). Em relação ao primeiro índice, este é definido genericamente pela probabilidade de duas unidades amostrais, retiradas ao acaso da amostra, serem dois objectos distintos. Assim, este índice toma valores entre 0 e 1 (diversidade mínima e máxima, respectivamente), em que o valor 0 obtém-se quando todas as unidades amostrais referem-se ao mesmo objecto, enquanto o valor 1 ocorre quando cada unidade amostral é um objecto distinto. Calcule-se, então, o índice de diversidade de Simpson para os dados em questão:

```
R> diversity.index(tcr$lncd8,method="simpson")
```

O valor deste índice é 0.7874, o que sugere uma diversidade de receptores algo elevada para os linfócitos T $CD8^+$ dos nódulos linfáticos.

Por vezes, é útil estimar um índice de diversidade numa dimensão amostral inferior à da amostra original. Nesse caso, o comando `diversity.index` possibilita a obtenção de uma estimativa do índice através do seguinte algoritmo *bootstrap*: (i) obter uma nova amostra de dimensão amostral inferior através da reamostragem sem reposição da amostra original, (ii) calcular o valor do índice nessa amostra simulada, (iii) repetir os passos anteriores um elevado número de vezes, (iv) estimar o índice pela mediana dos valores simulados. Este procedimento é accionado quando se especifica as opções `n` e `rep` deste comando, que indicam o tamanho das amostras simuladas e o número de repetições do algoritmo (passo (iii)), respectivamente.

A título ilustrativo, imagine-se que se pretende comparar a diversidade contida nas amostras dos linfócitos T $CD4^+$ e $CD8^+$ dos nódulos linfáticos. Contudo, estas duas amostras têm tamanhos diferentes e, assim, os valores dos índices de diversidade não devem ser comparados directamente. Para contornar este problema, Venturi *et al*. (2007) propuseram o algoritmo acima em que a dimensão das amostras simuladas é dada pelo mínimo de todas as amostras originais a comparar. No presente caso, pretende-se apenas confrontar a diversidade de duas amostras, em que a dos linfócitos T $CD4^+$ tem uma dimensão ligeiramente inferior à dos linfócitos T $CD8^+$ (n=98 vs n=122, respectivamente). Assim, há que recalcular o valor de índice de Simpson para a amostra dos linfócitos T $CD8^+$ através de 10000 reamostragens segundo o algoritmo acima para um tamanho amostral de 98 células:

```
R> diversity.index(tcr$lncd8,method="simpson",n=98,rep=10000)
```

O novo valor deste índice é então 0.7879, que é bastante inferior ao valor do mesmo índice para a amostra dos linfócitos T $CD4^+$ (0.9410). Portanto, a amostra dos linfócitos T $CD4^+$ é mais diversa do que a dos linfócitos T $CD8^+$ em termos deste índice.

Este procedimento pode ser ainda usado para obter a curva de um índice de diversidade à medida que novas unidades vão sendo incluídas na amostra. Caso a diversidade observada esteja próxima da sua contrapartida populacional, espera-se que essa curva atinja um patamar na dimensão da amostra original. Como a ordem com que se incluem as unidades amostrais é muitas vezes irrelevante para efeitos de análise, a curva referida acima deve ser obtida por simulação de Monte Carlo, o que pode ser feito através do seguinte código para a medida de entropia de Shannon:

```
R> n<-seq(10,120,10)
R> entropia<-
+ sapply(n,function(n)diversity.index(x=tcr$lncd8,method="entropy",n=n,
+ rep=10000))
R> plot(n,entropia,type="l",xlab="dimensão amostral",
+  ylab="Entropia de Shannon",las=1,cex.lab=1.5,pch=19)
R> points(seq(10,120,10),entropia,pch=19,cex=1.5)
```

Como se pode constatar na Figura 1A, o valor da entropia de Shannon aumenta com a dimensão amostral, mas não parece atingir um patamar. Assim, a diversidade observada não parece ser representativa da respectiva diversidade populacional. Portanto, há necessidade de fazer a melhor estimação possível da diversidade.

A estimação da diversidade propriamente dita é feita através do ajustamento dos modelos poissonianos de abundâncias descritos na Secção 2. Com esse fim, utiliza-se o comando `pam.fit` que essencialmente recebe os dados em diferentes formatos (para mais pormenores, consulte-se a ajuda deste comando), a indicação do modelo a ajustar (opção `model` que pode tomar os valores `homog` - modelo Poisson homógeneo, `geom` – modelo geométrico, `gamma` – modelo Poisson-Gama, `lognorm` – modelo Poisson-Lognormal, e `yule` – modelo Yule) e, caso necessário, um intervalo onde se julgue encontrar a estimativa da diversidade (opção `div.ini` que passa a ser obrigatória, caso se pretenda ajustar os modelos Poisson-Gama ou Poisson-Lognormal). Um modelo muito popular em Ecologia pela sua observação recorrente em diferentes contextos ecológicos é o Poisson-Lognormal, que pode ser estimado através da instrução:

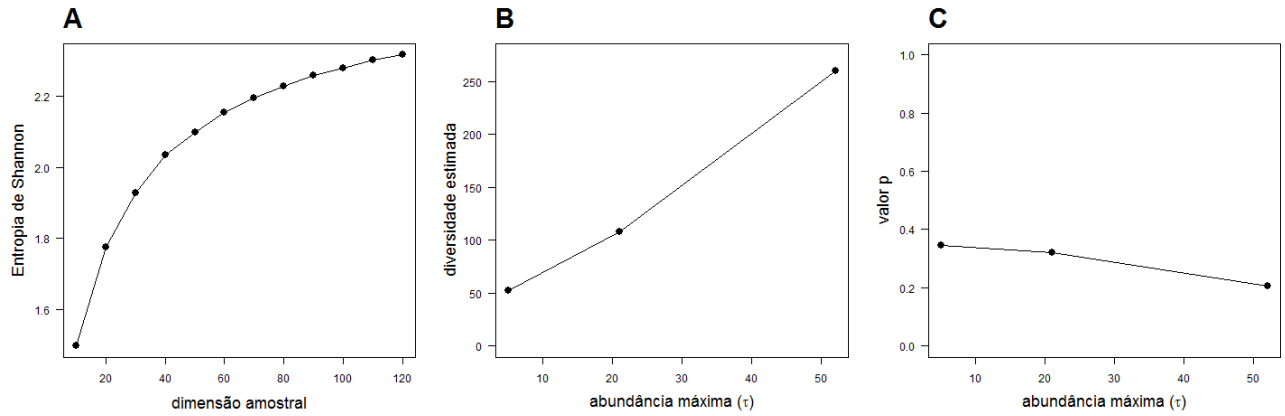


Figura 1: **A.** Medida de entropia de Shannon em função da dimensão amostral. **B.** Diversidade estimada pelo modelo Poisson-Lognormal quando ajustado aos dados referentes a todos os receptores com uma abundância menor ou igual a τ (abundância máxima). **C.** Variação do respectivo valor-P do teste de Pearson para o ajustamento daquele modelo em função da abundância máxima τ permitida na fase de estimação.

```
R> fit<-pam.fit(x=tcr$lncd8,model="lognorm",div.ini=c(250,1000))
```

tendo-se utilizado um intervalo para a diversidade entre 250 e 1000. Note-se que, durante a execução da instrução `pam.fit` para a estimação deste modelo, imprime-se na consola do R a evolução das respectivas estimativas (diversidade, média e desvio-padrão da distribuição Normal associada ao logaritmo das taxas de amostragens, e o valor correspondente da função log-verosimilhança) para efeitos de monitorização. Caso se detecte algum problema na estimação, recomenda-se que se aborte manualmente a execução dessa instrução, voltando a executá-la a partir de um novo intervalo para a diversidade. Para visualizar os resultados da estimação, basta usar o comando genérico `summary` ou então digitar apenas o nome do objecto R associado à instrução executada (neste caso, o objecto `fit`), tendo-se obtido

```
Call:
pam.fit.default(x = tcr$lncd8, model = "lognorm", div.ini = c(250,1000))

Model
Lognormal-Poisson

           Estimates
diversity 260.000000
mu         -3.850676
sigma2      6.075205
```

Os resultados obtidos para os restantes modelos podem ser encontrados em Sepúlveda *et al.* (2010).

Depois de se ter estimado o modelo, há que avaliar a sua qualidade de ajustamento. Para isso, tem-se o comando `gof.pam` que executa o teste de ajustamento de Pearson, considerando as seguintes categorias de abundância: 1, 2,…,*m* e "*>m*". O comando recebe como argumentos um objecto da classe `pam` (criado pelo comando `pam.fit`) e um valor para `trunc` que indica a categoria de abundância máxima *m* para o ajustamento do modelo. Esta instrução é agora aplicada ao objecto `fit` com *m*=4:

```
R> res<-gof.pam(fit,trunc=4)
```

Os resultados do teste de ajustamento de Pearson podem ser visualizados através do comando `summary` aplicado ao objecto `res` criado pelo comando `gof.pam`, isto é, `summary(res)`. Assim, obtém-se os seguintes resultados:

```
Call:

gof.pam(x = fit, trunc = 4)

Model
Lognormal-Poisson

           Estimates
diversity 260.000000
mu         -3.850676
sigma2      6.075205

Pearson's Goodness-of-fit Test
  Obs df p.value
 3.16  2   0.206
```

Para além da qualidade de ajustamento de um modelo, há que inferir também sobre a sua complexidade. Assim, disponibiliza-se o comando `aic.pam` que calcula a medida de Informação de Akaike, recebendo como argumento um objecto R criado pela instrução `pam.fit`, isto é,

```
R> aic.pam(fit)
```

que conduz ao valor 40.27 para o modelo Poisson-Lognormal ajustados aos dados sob análise. Este valor deve ser então comparado com os obtidos para os restantes modelos, como foi feito em Sepúlveda *et al*., 2010.

A distribuição empírica de $\{m_i\}$ mostra uma cauda gorda à direita devido à presença de alguns receptores extremamente abundantes; por exemplo, o mesmo receptor foi observado em 52 células. Nesta situação recomenda-se que se avalie a robustez das estimativas da diversidade na presença ou ausência dessas sequências na análise. Com esta preocupação em mente, retorne-se ao comando `pam.fit` que possibilita o controlo da abundância máxima de um determinado tipo de objecto permitida na fase da estimação de um modelo, através da opção `max.abund`. Por exemplo, se se pretender estimar o modelo Poisson-Lognormal com base na subamostra de todos os receptores dos linfócitos T $CD8^+$ com abundância menor ou igual a 21 células, basta escrever o seguinte código:

```
R>fit.21<-
+ pam.fit(tcr$lncd8,model="lognorm",div.ini=c(100,250),max.abund=21)
```

A qualidade de ajustamento destes dados truncados é feita tal como demonstrado anteriormente, mas agora sobre o objecto `fit.21`, isto é.,

```
R> res.21<-gof.pam(fit.21,trunc=4)
```

As Figuras 1B e C mostram a robustez das estimativas da diversidade sem os receptores mais abundantes e os respectivos valores-P do teste de ajustamento de Pearson. Por um lado, a Figura 1B revela que a estimativa da diversidade de receptores diminui quase três vezes quando se elimina dos dados os dois receptores mais abundantes (com 21 e 52 células, respectivamente). Por outro lado, a Figura 1C indica que o modelo Poisson-Lognormal se ajusta bem aos dados truncados ao nível de significância usual de 5%.

Em jeito de conclusão, este exemplo mostrou as funcionalidades básicas do pacote PAM para análise de diversidade contida numa amostra. Mostrou-se que o modelo Poisson-Lognormal se ajusta bem aos dados completos e truncados dos receptores dos linfócitos T $CD8^+$, mas a respectiva estimação da diversidade não é robusta em relação à presença dos receptores extremamente abundantes na amostra.

5. Trabalho futuro

No exemplo apresentado neste artigo e na maior parte dos dados encontrados na literatura, os tamanhos amostrais, ou a diversidade a inferir, não são "suficientemente" altos para levantar problemas computacionais no que diz respeito ao processamento das respectivas amostras pelo pacote PAM. Contudo, recentes desenvolvimentos tecnológicos irão permitir obter num futuro próximo, em tempo real e a relativo baixo custo, amostras de dimensões até agora inimagináveis, nomeadamente em estudos sobre a diversidade molecular de espécies biológicas. Estas novas técnicas experimentais já foram utilizadas num estudo ecológico com uma amostragem de cerca de 280 mil sequências de ADN para estimar a diversidade molecular de microorganismos num lago austríaco (Medinger *et al.*, 2010). Na área da Imunologia, semelhantes técnicas foram aplicadas para inferir a diversidade de receptores das células T em humanos (Freeman *et al.*, 2009). O tamanho da amostra analisada ronda as 120 mil sequências de aminoácidos que codificam os respectivos receptores apresentados pelas células T. Neste cenário interessa perguntar se o pacote estará à altura de lidar com a análise de tais amostras.

Acerca deste assunto, Sepúlveda *et al.* (2010) utilizaram com êxito o pacote PAM para analisar amostras simuladas até dez mil indivíduos, mesmo no modelo Poisson-Lognormal, o mais exigente em termos computacionais. Contudo, importa notar que as amostras foram geradas a partir dos modelos disponíveis no próprio pacote, o que pode ter contribuído para o bom comportamento dos algoritmos de estimação até agora implementados, que convergiram eficientemente para a solução correcta. Assim, uma investigação mais profunda à qualidade de processamento de grandes volumes de dados necessita ainda de ser realizada, o que se espera fazer futuramente com a análise de dados reais da magnitude mencionada acima.

O pacote PAM irá ser ainda mais desenvolvido de forma a disponibilizar um maior leque de inferências a realizar sobre os dados, nomeadamente, calcular intervalos de confiança para a diversidade e para os parâmetros dos modelos a analisar. Metodologias relacionadas com a análise da diversidade partilhada entre várias populações são outras das funcionalidades que se pretende para o pacote. Até agora o utilizador tem apenas à sua disposição o comando `similarity.index` (não ilustrado neste artigo), que calcula a similaridade entre duas amostras em termos dos conhecidos índices de Jacard e de Morisita-Horn, segundo as recomendações de Venturi *et al*. (2008) para a sua análise. Num futuro próximo, deseja-se estender o pacote com comandos que elevem este tipo de análise para um cenário paramétrico, tal como realizado em Sepúlveda (2009) para o modelo bivariado Poisson-Lognormal.

Em suma, o pacote PAM oferece um conjunto de metodologias básicas para a análise estatística da diversidade. Os comandos são de sintaxe simples e qualquer utilizador com o mínimo de experiência na linguagem R não terá qualquer dificuldade em executá-los e adaptá-los a outros propósitos inferenciais. Assim, o pacote parece ter todos os ingredientes para se tornar numa ferramenta útil a aplicar em estudos de diversidade. Espera-se que a sua futura utilização possa demonstrar não só a sua utilidade inferencial, mas também algumas das suas limitações e possíveis deficiências. Portanto, termina-se este artigo apelando aos futuros utilizadores que nos façam chegar comentários sobre a utilização do mesmo, ou sugestões que possam melhorar o seu desempenho.

Agradecimentos

Os autores agradecem a Christian Schlotterer pela sua disponibilização em discutir com os autores o seu artigo (Medinger *et al*., 2010). Um especial obrigado a Eurico de Sepúlveda e Frederico Poleto pela sua paciência em ler e criticar este artigo.

Referências

Bulmer, M.G. (1974). On fitting the Poisson Lognormal distribution to species abundance data. *Biometrics* 30:101-110.

Diserud, O. H. and Engen, S. (2000). A General and Dynamic Species Abundance Model, Embracing the Lognormal and the Gamma Models. *The American Naturalist*, 155:497-511.

Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H. and Holt, R. A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research* 19:1817-1824.

Medinger, R., Nolte, V., Pandey, R. V., Jost, S., Ottenwalder, B., Schlotterer, C. and Boenigk, J. (2010). Diversity in a hidden world: potential and limitation of next generation sequencing for surveys of molecular diversity of Eukaryotic microorganisms. *Molecular Ecology* (no prelo).

Sepúlveda, N. (2009). How is the T-cell repertoire shaped? Tese de doutoramento, Instituto de Ciências Biomédicas Abel Salazar, Universidade de Porto, Porto.

Sepúlveda, N., Paulino, C. D. e Carneiro, J. (2008). Diversidade de linfócitos T efectores e reguladores. Em *Estatística: da Teoria à Prática* (Hill, M. M., Ferreira, J. G., Dias, M. F., Salgueiro, H., Carvalho, Vicente, P. e Braumann, C., eds.), p. 513-524, Edições SPE, Lisboa.

Sepúlveda, N., Paulino, C. D. and Carneiro, J. (2010). Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *Journal of Immunological Methods*, vol.353, p. 124-37.

Venturi, V., Kedzierska, K., Tanaka, M. M., Turner, S. J., Doherty, P. C. and Davenport, M. P. (2008). Method for assessing the similarity between subsets of the T cell receptor repertoire. *Journal of Immunological Methods*, 326:67-80.

Venturi, V., Kedzierska, K., Turner, S. J., Doherty, P. C. and Davenport, M. P. (2007). Methods for comparing the diversity of samples of the T cell receptor repertoire. *Journal of Immunological Methods*, 321:182-195.

Yina, Z.-Y., Penga, S.-L., Rena, H., Guoa, Q. e Chene, Z.-H. (2005). LogCauchy, log sech and lognormal distributions of species abundances in forest communities. *Ecological Modelling*, 184:329-340.

Integração do *R* nos menus do *PASW Statistics*[1]:
Um exemplo de aplicação com o *package* 'polycor' do R

João Maroco, *jpmaroco@ispa.pt*

Unidade de Investigação em Psicologia e Saúde; Departamento de Estatística.
ISPA - Instituto Universitário

1. Introdução

As variáveis de natureza ordinal são relativamente comuns nas ciências sociais, ciências biomédicas e ciências da engenharia. Estas variáveis, de acordo com a definição original de Stevens (1946) reflectem apreciações qualitativas, com propriedades de grandeza ordenável, mas cuja, eventual, codificação numérica é desprovida de significado quantitativo. Exemplos típicos destas variáveis incluem as variáveis de tipo Likert (e.g., Concordância com politicas sociais ['1-Discordo completamente', '2-Discordo', '3-Nem discordo nem concordo', '4-Concordo' e '5 - Concordo completamente']; Severidade de dor ['3-Dói muito; 2-Dói; 1-Não Dói]) e as variáveis ordinais, designadas por *rating scales*, que reflectem características de intensidade crescente (e.g. Fissuras em betão [Sem fissuras, algumas fissuras, muitas fissuras] e Dureza de minerais [1 – Talco; 2 – Gesso, 3 – Calcite; … ; 10 – Diamante]). A estimação da associação entre variáveis ordinais é, tradicionalmente, feita por recurso ao coeficiente de correlação de Spearman. Esta medida, desenvolvida pelo psicometrista inglês Charles Spearman avalia quão bem uma função monótona arbitrária (ordenação) é capaz de descrever a associação entre duas variáveis cujas realizações são ordenáveis, sem contudo, apresentar algum tipo de assunção sobre a natureza dessa associação. O coeficiente de correlação de Spearman é frequentemente usado em estudos de natureza descritiva e inferencial. Contudo, a sua utilização em métodos correlacionais multivariados é pouco frequente (por exemplo, na Análise Factorial de escalas psicométricas, uma técnica correlacional multivariada proposta pelo próprio C. Spearman). O coeficiente de correlação de Pearson é também frequentemente usado com variáveis ordinais quando: (*i*) as classes destas variáveis são codificadas numericamente, (*ii*) se assume uma ordem implícita entre as codificações numéricas e (*iii*) se assume uma relação linear entre estas ordens. Contudo, a utilização do coeficiente de correlação de Pearson com variáveis ordinais é controversa. Em primeiro lugar o cálculo deste coeficiente requer o cálculo das médias das variáveis o que exige uma medida de natureza pelo menos intervalar. O cálculo de médias e desvios-padrão com variáveis ordinais é, porém, uma prática frequente em algumas áreas de estudo, e.g. na Psicometria, ainda que não esteja isenta de criticismo. Como refere Stevens: "As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than the relative rank order of data" (Stevens, 1946, p. 679). Vários estudos de simulação (para uma revisão ver, e.g., Bollen, 1989, p. 434-435) têm demonstrado que as estimativas da associação entre variáveis ordinais, obtidas com a correlação de Pearson, são geralmente inferiores às verdadeiras associações entre as variáveis quantitativas de cuja discretização resultaram as variáveis ordinais simuladas. Esta atenuação é tanto maior quanto menor for o número de classes (menos de 5) e quanto maior for a oposição das assimetrias das variáveis. À medida que o número de classes aumenta (7, 10,…), a medida aproxima-se mais de uma métrica quantitativa, e as correlações de Pearson calculadas para variáveis ordinais aproximam-se das

[1] O software *R* é uma marca registada de 'The R Foundation'; O software *PASW Statistics* é uma marca registada de 'SPSS, An IBM Company'.

correlações obtidas para as variáveis quantitativas correspondentes (Bollen, 1989, p. 435; Finney e Distefano, 2006, p. 276). Mais recentemente, alguns autores têm proposto que as análises correlacionais (Análise factorial, modelos de equações estruturais, etc.…) com variáveis ordinais sejam baseadas não nos usuais coeficientes de correlação de Pearson e ou Spearman, mas sim no coeficiente de correlação policórica. Num estudo de simulação de larga escala, Babakus, Ferguson, & Jöreskog (1987) observaram que o estimador da correlação policórica produziu as melhores estimativas das correlações bivariadas entre variáveis ordinais, comparativamente às estimativas obtidas com os coeficientes de Pearson, Spearman ou τ de Kendall. Contudo, a correlação policórica não está disponível nos principais softwares *user-friendly* de análise estatística, nomeadamente no *PASW Statistics* (até à v. 18, inclusive). Neste artigo será apresentado o coeficiente de correlação policórica, a sua implementação no *package* 'Polycor' do sistema *R* e a sua integração nos menus do *PASW Statistics* por recurso ao 'PASW Custom Dialog Builder'. Este aplicativo permite disponibilizar as bibliotecas do *R* nos menus do *PASW*, libertando assim o utilizador final da necessidade de dominar a programação em *R*.

2. O coeficiente de correlação policórica

De uma forma geral, as escalas psicométricas produzem resultados que se podem assumir como estimativas quantitativas de habilidades cognitivas. Assim, os itens ordinais da maioria das escalas psicométricas, podem assumir-se como 'operacionalizações' de variáveis latentes (as habilidades) contínuas de natureza pelo menos intervalar, que só podem ser estimadas por intermédio de avaliações ordinais. De forma a aceder à verdadeira associação entre variáveis latentes, de cujos itens ordinais são manifestações, é necessário estimar a correlação não entre os itens, mas sim entre as variáveis latentes. É esta associação que o coeficiente de correlação policórica estima. O estimador do coeficiente de correlação policórica foi desenvolvido a partir dos trabalhos seminais de Karl Pearson que reconheceu a facilidade de operacionalização de variáveis latentes por intermédio de itens ordinais. Conceptualmente, uma variável ordinal X com 5 categorias pode ser interpretada como o resultado da divisão de uma variável, subjacente ou latente, ξ em 5 categorias. Por exemplo, um item de uma escala de satisfação com 5 classes de '1–Nada satisfeito', a '5– Muito satisfeito', pode ser conceptualizado como

$$X = \begin{cases} 1 - \textit{Nada Satisfeito} \text{ se } \xi \le \xi_1 \\ 2 - \textit{Pouco Satisfeito} \text{ se } \xi_1 < \xi \le \xi_2 \\ 3 - \textit{Indiferente} \text{ se } \xi_2 < \xi \le \xi_3 \\ 4 - \textit{Satisfeito} \text{ se } \xi_3 < \xi \le \xi_4 \\ 5 - \textit{Muito Satisfeito} \text{ se } \xi > \xi_4 \end{cases} \quad (1)$$

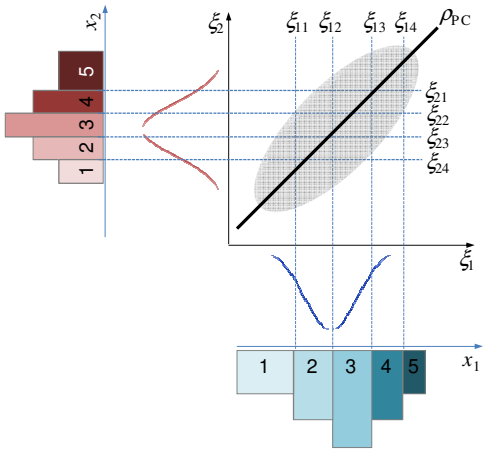


Figura 1 - Ilustração da correlação policórica entre duas variáveis contínuas latentes (ξ_1 e ξ_2) operacionalizadas por duas variáveis ordinais manifestas (X_1 e X_2) cada uma com 5 classes definidas pelos pontos-de-corte $\xi_{11},..,\xi_{14}$ e $\xi_{21},..,\xi_{24}$. O coeficiente de correlação policórica é ilustrado pela linha ρ_{PC}.

Assim, as 5 categorias de resposta podem ser interpretadas como uma aproximação aos 5 intervalos em que a variável latente ξ, contínua, foi dividida por recurso a 4 pontos de corte ou *thresholds*: ξ_1, ξ_2, ξ_3 e ξ_4 sendo, por convenção, $\xi_0=-\infty$ e $\xi_5=+\infty$. A correlação policórica estima a associação entre duas variáveis latentes, que se assumem com distribuição normal bivariada, subjacentes a duas variáveis ordinais manifestas. A figura 1 ilustra este conceito.

O coeficiente de correlação policórica pode estimar-se pelo método de máxima verosimilhança (ver, e.g., Drasgow, 2006). Neste método, a probabilidade conjunta (P_{ij}) de se observar o valor x_{1i} para a variável ξ_1 e o valor x_{2j} para a variável ξ_2 é estimada por

$$P_{ij} = \int_{\xi_{1i-1}}^{\xi_{1i}} \int_{\xi_{2i-1}}^{\xi_{2i}} \phi(\xi_1, \xi_2; \rho) d\xi_2 d\xi_1 \tag{2}$$

Onde

$$\phi(\xi_1, \xi_2; \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \times e^{\left(\frac{-1}{2(1-\rho^2)}(\xi_1^2 - 2\rho\xi_1\xi_2 + \xi_2^2\right)} \tag{3}$$

é a função de densidade normal bivariada de ξ_1 e ξ_2 com, sem perda de generalidade, μ=0, σ=1 e correlação de Pearson ρ. Se n_{ij} for o número de observações x_{1i} da variável X_1 e de x_{2j} na variável X_2, a verosimilhança das observações amostrais é

$$L = k \prod_{i=1}^{r} \prod_{j=1}^{s} P_{ij}^{n_{ij}} \tag{4}$$

onde k é uma constante e r e s são o número de classes de x_1 e x_2, respectivamente. A estimativa de máxima verosimilhança do coeficiente de correlação policórica (ρ_{PC}) obtém-se derivando o $Ln(L)$ em ordem a ρ

$$\frac{\partial Ln(L)}{\partial \rho} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}}{P_{ij}} \left[\phi(\xi_{1i}, \xi_{2j}; \rho) - \phi(\xi_{1i-1}, \xi_{2j}; \rho) \right] - \phi(\xi_{1i}, \xi_{2j-1}; \rho) + \phi(\xi_{1i-1}, \xi_{2j}; \rho) \tag{5}$$

Igualando a derivada parcial (5) a 0 e resolvendo a equação obtida em ordem a ρ obtém-se a estimativa de ρ_{PC}. Contudo, e de uma forma geral, a solução do estimador de máxima verosimilhança é obtida de forma iterativa já que resolução da derivada parcial em ordem a ρ exige o conhecimento dos restantes parâmetros da função de verosimilhança. A estimação destes parâmetros, por sua vez, exige a derivada parcial de $Ln(L)$ em ordem a cada um dos coeficientes de (4) (ρ, $\xi_{11},..,\xi_{1r}$, $\xi_{21},..,\xi_{2s}$) e a resolução do sistema com todas as derivadas parciais iguais a 0. A correlação policórica pode também estimar-se com outro algoritmo, computacionalmente mais rápido, designado por algoritmo *two-step* (Martinson & Hamdan, 1975). Neste algoritmo, começa-se por ajustar distribuições normais univariadas às distribuições marginais de x_1 e x_2, estimando os *thresholds* ($\xi_{11},..,\xi_{1r}$, $\xi_{21},...,\xi_{2s}$) para cada distribuição. No passo final, a estimativa de ρ_{PC} é obtida derivando a função $Ln(L)$ em ordem a ρ, igualando a derivada 0 e resolvendo a equação em ordem a ρ (Drasgow, 2006; Martinson & Hamdan, 1975). Se uma das variáveis for quantitativa e a outra for ordinal, a correlação entre as duas obtêm-se estimando a variável latente subjacente à variável ordinal, como anteriormente, calculando-se, em seguida, a sua associação com a variável quantitativa. Esta associação é conhecida por 'correlação poliserial'.

3. O package *POLYCOR* do sistema *R*

O coeficiente de correlação policórica pode obter-se quer em *software* comercial [PRELIS/ LISREL, EQS, MPLUS, STATA 8 (programa de Stas Kolenikov), SAS (Polychoric Macro)] quer no sistema *R* com o *package* 'Polycor' da autoria de John Fox (jfox@mcmaster.ca) e disponível nos repositórios do CRAN. O *package* 'Polycor' permite calcular correlações policóricas, poliseriais (correlação entre uma variável ordinal e uma variável quantitativa) com estimativas opcionais dos erros-padrão e testes à distribuição normal multivariada. O *package* é composto por 3 programas. O programa 'HetCor' produz uma matriz de correlações constituída por correlações de Pearson (para variáveis numéricas), correlações policóricas para (variáveis ordinais – *factors*) e correlações poliseriais (para variáveis

numéricas vs. variáveis ordinais). O programa 'Polychor' calcula as correlações policóricas e respectivos erros-padrão entre duas variáveis ordinais a partir dos valores originais ou de uma tabela de contingência. Finalmente, o programa 'Polyserial' calcula a correlação poliserial e o erro-padrão respectivo. Em ambos os cenários, o *package* só calcula os coeficientes policóricos e poliseriais se as distribuições das variáveis latentes não estiverem muito afastadas da distribuição normal.
Para ilustrar a utilização do package 'Polycor' consideremos um ficheiro de dados 'DadosAF.sav' em formato SPSS (.sav)[2] constituído por 5 variáveis ordinais com 5 pontos (X_1 a X_5), uma variável intervalar (*Age*) e uma variável nominal (*Sex*). Para integrar o *package* 'Polycor' no *R* recorremos ao *Mirror* do CRAN mais próximo e carregamos a biblioteca 'Polycor':

```
> library(polycor)
```

Para importar o ficheiro de dados 'DadosAF.sav':

```
> library (foreign)
> data<-read.spss("D:/DataAF.sav", to.data.frame = TRUE)
```

Para que o programa HetCor calcule o coeficiente de correlação apropriado à métrica das variáveis é necessário fazer a conversão da 'measure' do *PASW* para a métrica do *R*:

```
> Data$X1<-ordered(Data[,1])          # Converte métrica para ordinal
> Data$X2<-ordered(Data[,2])          # Converte métrica para ordinal
    (...)
> Data$X5<-ordered (Data[,5])
> Data$Age<-as.numeric(Data[,6])      # Converte métrica para numérico
> Data$Sex<-factor(Data[,7])          # converte métrica para nominal
```

Finalmente, para obter os coeficientes de correlação ('ML=TRUE' para usar o método de máxima verosimilhança ou 'ML=FALSE' para usar o método *two-step*):

```
> R<-hetcor(Data, ML=FALSE, std.err=TRUE)
> print (R)
```

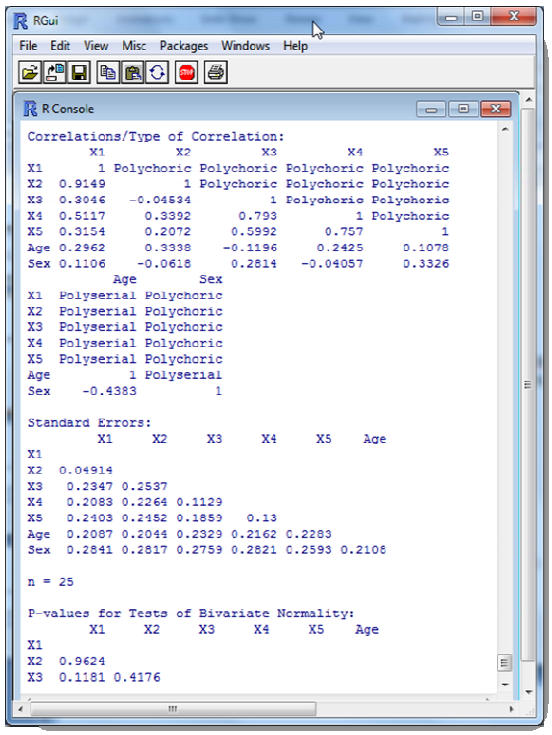


Figura 2 – *Output* do programa HetCor no sistema *R*

[2] Disponível se solicitado, por e-mail, ao autor.

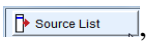
4. Integração do R nos menus do *PASW Statistics* com o 'CUSTOM DIALOG BUILDER'

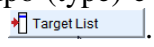
A partir da versão 16 do *PASW Statistics* é possível integrar o código do *R* (e também do Phyton) na sintaxe do *PASW*. Na versão 17 do PASW, foi desenvolvido um novo aplicativo – Custom Dialog Builder (CDB) – que permite incorporar as funções do *R* nos menus do *PASW* facilitando a interacção do utilizador do *PASW*, não familiarizado com a linguagem do *R*, com as potencialidades do *R*. Na verdade, o utilizador final dos menus do *PASW* não necessita sequer compreender o funcionamento do *R* já que todo processo corre na memória do computador sem evocar o GUI do *R*. Para permitir a comunicação entre o *PASW* e o *R* é necessário instalar o 'R integration plugin' do *package* 'R essentials' disponível gratuitamente no endereço http://www.spss.com/devcentral/. O 'R essentials' disponibiliza a instalação automática do 'PASW: R integration Plug-in', do 'R2.8.1' bem como vários exemplos de bibliotecas de *R* transpostas para os menus do *PASW*. O 'R integration Plug-in' disponibiliza as funções necessárias à exportação das variáveis do *PASW* para um *dataset* de *R*; à leitura da informação das variáveis do *PASW* para o formato *R*; à transferência dos resultados de análise (via OMS) do *PASW* para o *R*; à escrita dos resultados das análises efectuadas no *R* para o *output* do *PASW* e à transposição dos gráficos gerados em *R* para o *output* do *PASW*.
Depois de instalado o 'R integration Plug-in' o código do *R* poderá ser incorporado na sintaxe do *PASW*, num bloco específico de sintaxe:

```
BEGIN PROGRAM R.        */ Inicio Bloco de Sintaxe do PASW
(...)                   # Código de R
END PROGRAM.            */ Fim bloco Sintaxe de PASW. Não esquecer os "."
```

Dentro deste bloco, todo e qualquer código do *R* pode ser utilizado, bem como as funções do 'R integration Plug-in' para leitura de dados, de variáveis e conversão dos resultados para os *outputs* do *PASW*. Este código pode depois ser utilizado no 'Custom Dialog Builder' com modificação mínimas apenas ao nível da leitura das variáveis e das opções de análise.
No *PASW*, para iniciar o 'Custom Dialog Builder' recorremos ao menu

▸ Utilities
▸ Custom Dialogs
▸ Custom Dialog Builder

A janela do 'Custom Dialog Builder' e a barra de ferramentas de desenho ilustra-se na figura 3.
Para definir as propriedades do novo menu, preenche-se a caixa 'Dialog Properties' com o nome do menu (Dialog name); o local onde o novo menu deverá aparecer (Menu Location); o titulo do menu (Title) e a localização do ficheiro de ajuda (em formato .html). A opção 'Modeless' indica se o menu é independente de outra interacção com os menus do *PASW* (True) ou se enquanto este estiver activo não é possível usar outras funcionalidades do *PASW* (False). As especificações do nosso exemplo apresentam-se na figura seguinte. Note que o novo menu designado 'Poly' irá aparecer no menu Analyze ▸ Correlate (v. figura 4). Para desenhar o menu, recorremos às ferramentas de desenho da Fig. 3. Para definir a caixa de selecção de variáveis para análise, usa-se a ferramenta 'Source List' 

, desenha-se um rectângulo na área de desenho do CDB e definem-se as propriedades das variáveis que podem entrar na análise, nomeadamente o tipo (type) e a métrica (measurement level). Para seleccionar as variáveis a analisar usa-se a ferramenta 

. A Fig. 5 ilustra o menu com as caixas de selecção de variáveis e variáveis a analisar.

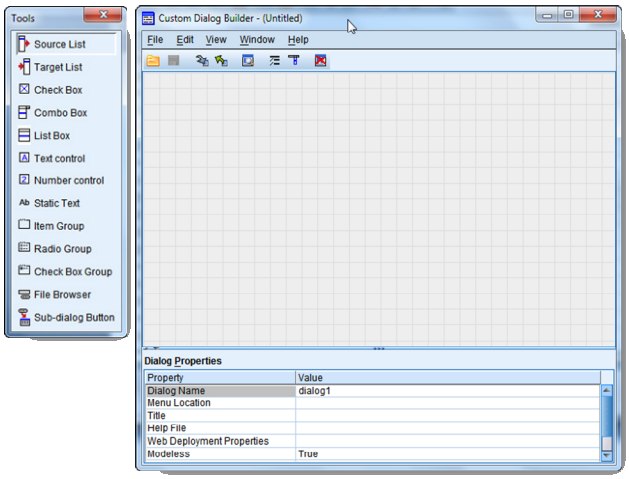


Figura 3 – 'Custom Dialog Builder' (CDB) com a barra de ferramentas

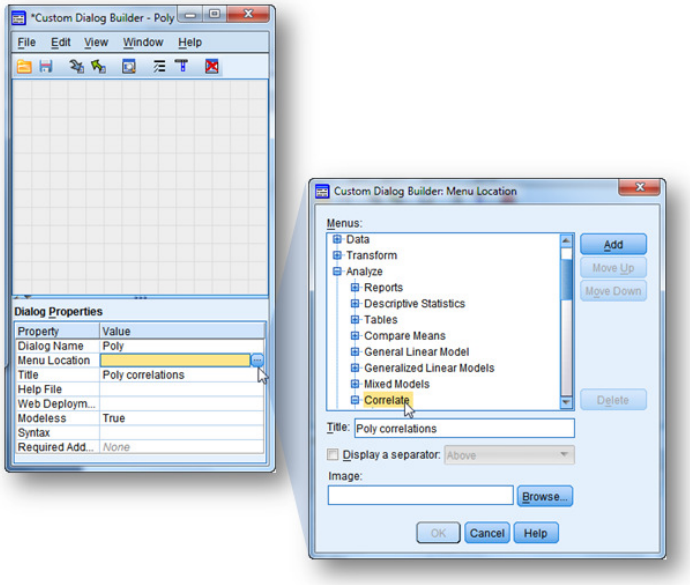


Figura 4 – Atribuição da localização do novo menu 'Poly'. Para abrir a janela de atributos da localização clica-se no botão .

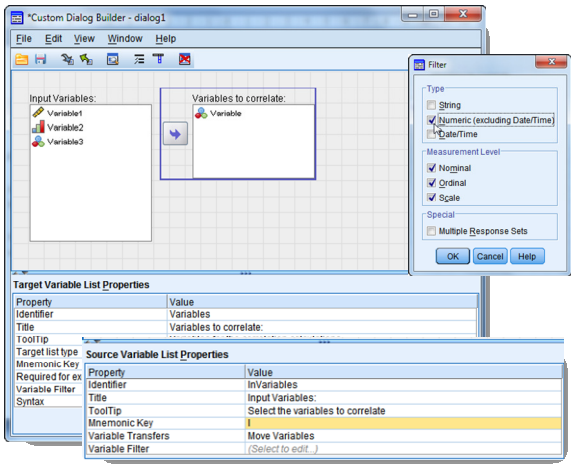


Figura 5 – Definição das variáveis a usar no menu

Para definir opções de análise, por exemplo, o método de estimação recorremos à ferramenta Check Box Group. Na janela dos 'Radio buttons' definimos as duas opções de análise como ilustra a Fig. 6.

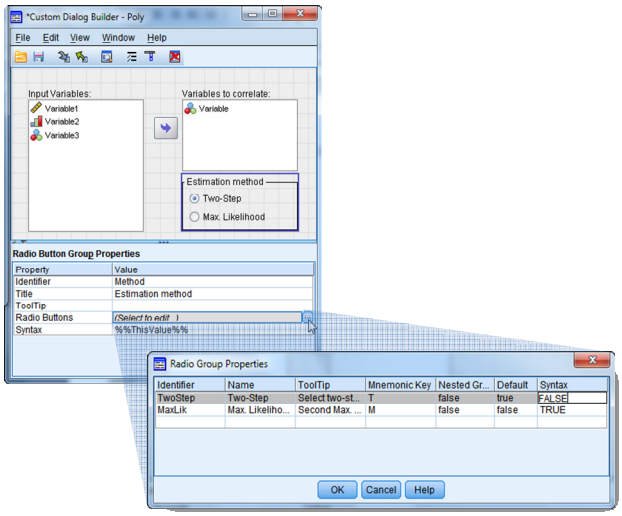


Figura 6 – Definição do método de estimação dos coeficientes. *Two-Step* (keyword FALSE na opção 'ML=' no programa HetCor) ou *Maximum likelihood* (keyword TRUE na opção 'ML=' no programa HetCor). Para abrir a janela de atributos dos botões clica-se no botão .

Finalmente, introduzimos a sintaxe do *PASW* com código do *R* para evocar o programa HetCor. No 'CDB' clica-se em 'Syntax' e digita-se a sintaxe apropriada: (v. figura 7)

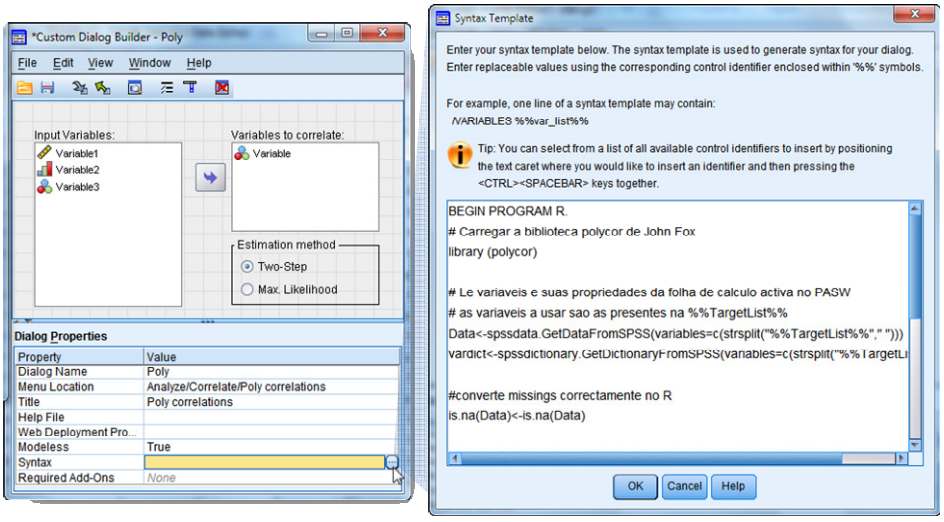


Figura 7 – Evocação da Janela para digitação da sintaxe com o código R adequado à importação das variáveis do *PASW* para o *R*, à realização dos cálculos no *R* e ao retorno dos resultados para o *output* do *PASW*.

A sintaxe, anotada, adequada ao exemplo apresenta-se a seguir:

```
BEGIN PROGRAM R.
# Carregar a biblioteca polycor de John Fox
library (polycor)

# Lê variáveis e suas propriedades da folha de calculo activa no PASW
# as variáveis a usar são as presentes na %%TargetList%%
  Data<-spssdata.GetDataFromSPSS(variables=c(strsplit("%%TargetList%%"," ")))
vardict<-spssdictionary.GetDictionaryFromSPSS (variables = c(strsplit( "%%TargetList%%","
")))

#Converte missings correctamente no R
 is.na(Data)<-is.na(Data)

# Converte tipo de variáveis nos tipos do R: numeric, non-ordered ou ordered R # factors
  Mat<-Data
```

```
 for (i in 1:length(Data)) {
  if  (vardict["varMeasurementLevel",i]=="scale")  Mat[,i]<-Data[,i]
 else
  if  (vardict["varMeasurementLevel",i]=="nominal") Mat[,i]<- factor(Data[,i]) else
  if  (vardict["varMeasurementLevel",i]=="ordinal") Mat[,i]<-ordered(Data[,i])
}

# Calcula a matriz de correlação como R List; usar str(r) para ver a
# estrutura da lista. O método de estimação e dado pela keyword %%Estimation%%
  r <- hetcor(Mat,  ML = %%Estimation%%)

#Cria matriz triangular com estimativas e tipos de correlação
  R <- signif(r$correlations, digits=3)
  R[upper.tri(R)] <- r$type[upper.tri(R)]
  R <- as.data.frame(R)

# Formata as tabelas para output no PASW
  spsspivottable.Display(R, title="Correlation types and Correlation
  Coeficients")

# Legenda para as colunas e linhas
  lab=c(strsplit("%%TargetList%%", " "))
# Dimensão da amostra
  spsspivottable.Display(round(r$n), title="N", rowlabels=lab, collabels="n")
# Erros-padrão dos coeficientes
  spsspivottable.Display(r[["std.errors"]], title="std. errors")
# Teste à Distribuição normal bivariada
  spsspivottable.Display(r[["tests"]], title="p-values for Tests of Bivariate
Normality ")

# Limpa memória de trabalho
  rm(list = ls())
END PROGRAM.
```

Finalmente, o menu 'Poly' é inserido no PASW, clicando no botão  como ilustra a figura seguinte:

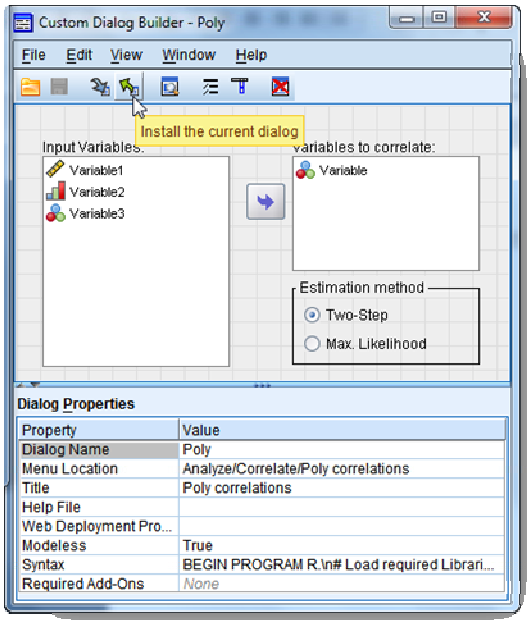


Figura 8 – Inserção do menu 'Poly' no PASW Statistics.

Naturalmente, pode também gravar-se o ficheiro (extensão .spd) com o código para o novo menu e cálculos associados. Este ficheiro[3] poderá agora ser distribuído em conjunto com os *packages* do *R*, e instalado no *PASW* pelo utilizador final com um simples duplo-click.

Depois de reiniciar o *PASW*, o menu 'Poly' acede-se por

- Analyze
  - Correlate
    - Poly

como ilustra a Fig. 9.

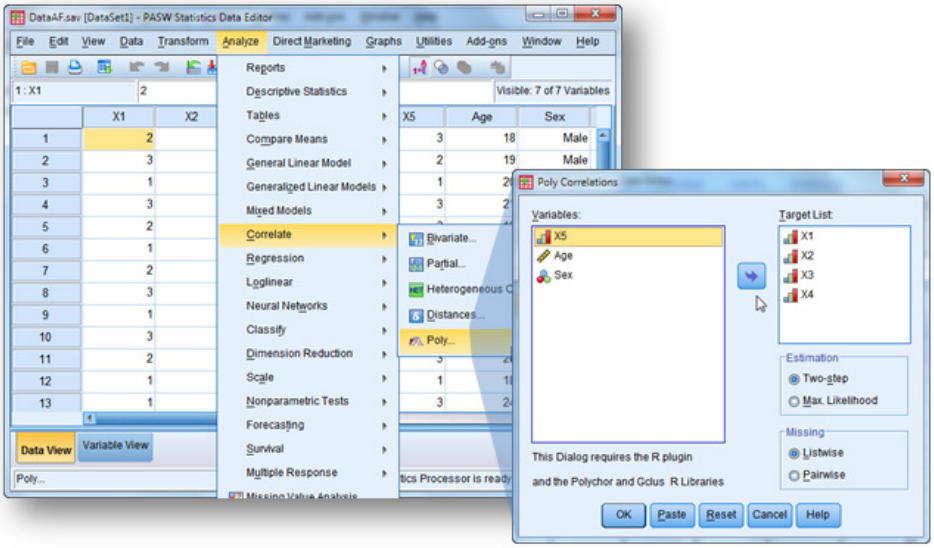
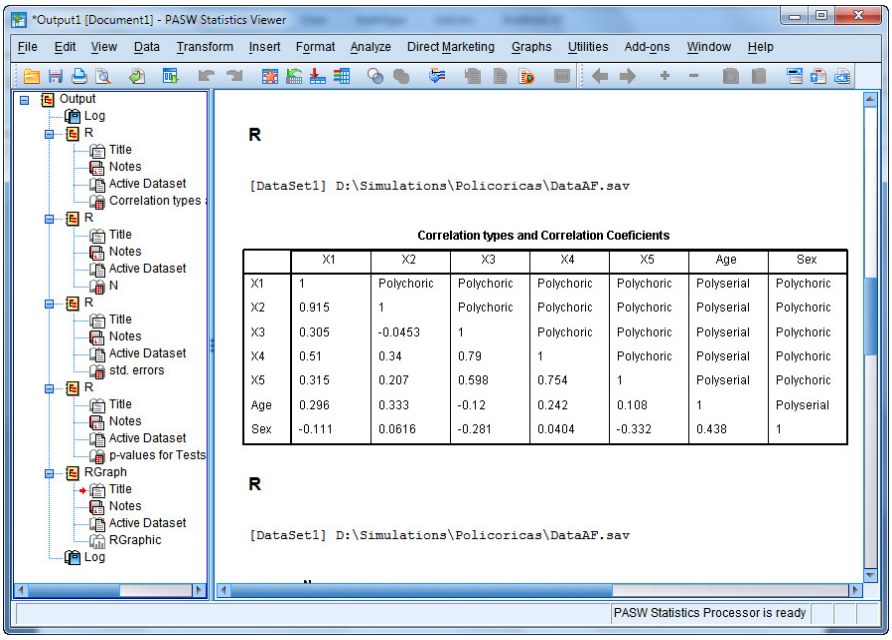


Figura 9 – *PASW Statistics* com o novo menu para cálculo das correlações policóricas

Depois de seleccionadas as variáveis para as quais se pretende calcular os coeficientes de correlação, o resultado final aparecerá na janela de *output* do *PASW Statistics*:



As tabelas (e os gráficos) do *output* do *PASW* podem agora ser exportadas e ou copiadas para outros aplicativos.

[3] O ficheiro deste exemplo pode ser solicitado, via email, ao autor, ou poder-se-á descarregar do sitio da SPSS DevCentral.

Referências Bibliográficas

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The Sensitivity of Confirmatory Maximum Likelihood Factor Analysis to Violations of Measurement Scale and Distributional Assumption. *Journal of Marketing Research*, 24(2), 222-228.

Finney, S. J., & DiStefano, C. (2006). Non-normal and Categorical Data in Strutctural Equation Modelling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: a second course* (pp. 269-314). Greenwich, Conn.: IAP.

Martinson, E. O., & Hamdan, M. A. (1975). Algorithm AS 87: Calculation of the Polychoric Estimate of Correlation in Contingency Tables. *Applied Statistics*, 24(2), 272-278.

Drasgow, F. (2006). Polyserial and Polychoric correlations. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences* (2nd ed., Vol. 9, pp. 6244-6248). New York: John Wiley & Sons.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.

A função estatística do Gabinete de Estatística e Planeamento da Educação do Ministério da Educação

Nuno Rodrigues, Céline Ambrósio
nuno.rodrigues@gepe.min-edu.pt, celine.ambrosio@gepe.min-edu.pt

Direcção de Serviços de Estatística
Gabinete de Estatística e Planeamento da Educação do Ministério da Educação

1. ENQUADRAMENTO E ORGANIZAÇÃO

A necessidade de dispor de informação estatística no domínio da Educação, com as essenciais características de qualidade e actualidade, foi progressivamente compreendida no nosso País, a um ritmo comparável ao do próprio reconhecimento do papel fundamental que cabe ao sector educativo na estruturação da sociedade. Desta forma, em 1992, o Instituto Nacional de Estatística (INE)[1] passou a delegar as funções de recolha e divulgação de informação estatística no domínio da educação, ao Ministério da Educação (ME) que, pela sua natural sensibilidade aos problemas do sistema educativo e pela necessidade de dispor de informação que lhe permitisse agir racionalmente sobre esse sistema, foi considerada a entidade mais apta para definir as principais necessidades do País em termos de estatísticas da educação. Deste modo, o Gabinete de Estudos e Planeamento do Ministério da Educação (GEP/ME), actual Gabinete de Estatística e Planeamento da Educação do Ministério da Educação (GEPE/ME)[2], tornou-se o órgão delegado do INE[3] para a produção de estatísticas oficiais da educação.

Ao GEPE foi atribuída a missão de garantir a produção e análise estatística da educação, tendo em vista o apoio técnico à formulação de políticas, ao planeamento estratégico e operacional, e uma adequada articulação com a programação financeira, bem como a observação e avaliação global de resultados obtidos pelo sistema educativo, cabendo-lhe ainda assegurar o apoio às relações internacionais e à cooperação nos sectores de actuação do Ministério da Educação. É orientado pelos princípios do Sistema Estatístico Nacional[4] de autoridade estatística, independência técnica, segredo estatístico, qualidade e acessibilidade estatística.

Uma das suas unidades orgânicas nucleares é a Direcção de Serviços de Estatística (DSE/GEPE)[5]. A esta direcção de serviços compete garantir a produção de informação adequada, no quadro do sistema estatístico nacional, nas áreas de intervenção do ME, bem como prestar apoio técnico em matéria de definição e estruturação das políticas, prioridades e objectivos do ME. Tem assim a responsabilidade de produzir, organizar e manter actualizada, com respeito pelas normas legais relativas à análise e produção estatística, uma base de dados de informação estatística relativa ao sistema educativo e assegurar, no quadro do sistema estatístico nacional, a articulação com os departamentos e organismos congéneres, a nível nacional e internacional, tendo em vista a harmonização estatística e a partilha de informação não classificada. Como base para o exercício destas funções tem a seu cargo a promoção do aperfeiçoamento dos instrumentos e processos inerentes à recolha, produção e análise da informação estatística, contribuindo para a modernização e racionalização da organização e dos

[1] Criado pela Lei n.º 1911 de 1935
[2] Criado pelo Decreto-lei n.º 213/2006, de 27 de Outubro
[3] Estatuto delegado através do Despacho Conjunto MPAT/ME/92, de 20 de Novembro
[4] Lei n.º22/2008, de 13 de Maio
[5] Portaria n.º356/2007 de 30 de Março

procedimentos de gestão do sistema educativo, bem como a gestão do sistema integrado de informação e gestão da oferta educativa e formativa.

Para além da utilização interna do Ministério da Educação, servindo de base a análises científicas e de suporte a tomada de decisões no âmbito das suas funções, e do reporte a organismos internacionais como a Organização das Nações Unidas para a Educação, Ciência e Cultura (UNESCO), a Organização para a Cooperação e Desenvolvimento Económico (OCDE) e o Eurostat (Gabinete de Estatísticas da União Europeia), as estatísticas produzidas pela DSE/GEPE são disponibilizadas ao público através do sítio do GEPE na Internet[6]. Os documentos/publicações elaborados e disponibilizados têm a intenção de se constituírem não só como instrumentos privilegiados de apoio à tomada de decisão política, aos processos de monitorização e avaliação das medidas de política, mas também ao trabalho de investigação científica. São, pois, simultaneamente, um motivo e um convite para cientistas sociais, investigadores e analistas, bem como decisores políticos, agentes do mercado direccionados para a educação e profissionais da comunicação social, laborarem sobre o seu significado e perspectivarem as suas implicações. Também neste sítio os utilizadores podem aceder a uma plataforma para efectuarem pedidos de apuramento personalizado[7] e esclarecer dúvidas.

2. DOMÍNIOS ESTATÍSTICOS

A DSE/GEPE recolhe e elabora estatísticas em vários domínios, tendo ainda a incumbência de conceber e/ou aplicar questionários e desenvolver indicadores sobre especificidades em matéria de educação e formação, recolhendo e acompanhando a recolha de informação e/ou os resultados de estudos no âmbito de projectos nacionais e internacionais.

Sem se ter a pretensão de enumerar e descrever de uma forma exaustiva as diversas actividades desenvolvidas pela DSE/GEPE, serão focadas apenas duas grandes áreas de trabalho:

- Principais operações estatísticas
  Operações estatísticas de periodicidade anual desenvolvidas, de uma forma geral, no âmbito do Sistema Estatístico Nacional e que dão origem à produção e divulgação de estatísticas e publicações com a chancela de estatísticas oficiais.
- Colaboração com entidades/projectos
  Projectos e actividades nas quais a DSE/GEPE tem a responsabilidade de participar ou coordenar.

2.1. Principais operações estatísticas

Neste âmbito, destacam-se as seguintes operações estatísticas:

- Recenseamento Escolar Anual;
- Modernização Tecnológica das Escolas;
- Actividades de Enriquecimento Curricular;
- Promoção e Educação para a Saúde nas Escolas.

2.1.1. Recenseamento Escolar Anual

O Recenseamento Escolar Anual trata-se de uma operação estatística que tem por âmbito a recolha de dados junto de todos os estabelecimentos de educação e ensino, que ministram desde a educação pré-escolar até ao ensino secundário, de natureza pública ou privada, de Portugal Continental. Visa obter informação sobre os alunos matriculados e respectivo aproveitamento, e sobre os recursos humanos – pessoal docente e pessoal não docente, em exercício no estabelecimento. Actualmente, os dados dos

[6] http://www.gepe.min-edu.pt/
[7] http://w3.gepe.min-edu.pt/ApuramentoPersonalizado/

estabelecimentos de educação e ensino públicos, bem como alguns estabelecimentos privados, são reportados por via administrativa. Em relação aos restantes estabelecimentos, a informação referente a alunos e recursos humanos é recolhida num só questionário e procede-se à inquirição dos dados pessoais de cada indivíduo. Dá-se, desta forma, a obtenção de informação detalhada e um controlo sobre a qualidade dos dados recolhidos que possibilitam o acompanhamento e a avaliação da dinâmica do sistema educativo, bem como contribuem para a tomada de decisão política.

O resultado desta operação permite assim a produção e a actualização de indicadores, nacionais e internacionais, sobre o sistema educativo, a elaboração de estudos prospectivos e de planeamento estratégico, a resposta a um leque de solicitações que chegam ao GEPE e a elaboração de publicações.

Das publicações salientam-se as *Estatísticas da Educação* anuais que integram elementos estatísticos sobre alunos matriculados, aproveitamento, pessoal docente, pessoal não docente e estabelecimentos de educação e ensino. Os dados de alunos e respectivo aproveitamento são sistematizados por níveis de ensino e apresentam desagregações segundo a natureza institucional**,** o sexo, a idade e o ano de escolaridade. Os indicadores apresentados, bem como os valores absolutos de docentes são sistematizados também por níveis de ensino e apresentam desagregações segundo a natureza institucional**,** o sexo, a idade, a situação profissional, o grau académico e o grupo de recrutamento. Os dados relativos ao pessoal não docente são apresentados segundo a natureza institucional**,** o sexo, a idade, a situação profissional, o grau académico e a categoria/função. Por seu lado, os estabelecimentos de educação/ensino apresentam-se organizados segundo a natureza institucional, a tipologia e o nível de ensino ministrado.

Alunos, docentes, pessoal não docente e estabelecimentos de educação /ensino, em Portugal Continental (1997/98 - 2007/08)

		1997/98	1998/99	1999/00	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08
Alunos	***Total***	**1 818 754**	**1 788 288**	**1 776 251**	**1 762 375**	**1 724 039**	**1 700 598**	**1 694 241**	**1 683 008**	**1 648 558**	**1 670 763**	**1 701 482**
	Pré-Escolar	201 913	207 315	214 857	221 407	226 892	232 555	238 364	243 921	246 090	247 826	250 629
	1º Ciclo	497 857	502 483	504 885	501 221	487 197	475 892	473 156	472 863	465 238	469 831	469 829
	2º Ciclo	266 612	263 113	259 030	254 979	254 606	257 782	257 274	251 285	240 227	240 199	248 326
	3º Ciclo	430 887	415 081	400 542	391 470	378 440	369 088	363 635	358 747	370 821	375 978	402 705
	Secundário	421 485	400 296	396 937	393 298	376 904	365 281	361 812	356 192	326 182	336 929	329 993
Docentes	***Pré-Escolar***	12 172	13 054	14 152	14 704	14 777	15 414	15 394	16 267	16 602	16 707	15 972
	1º Ciclo	34 239	35 182	36 625	36 722	37 918	37 214	37 251	37 506	36 244	31 371	32 286
	2º Ciclo	31 398	32 560	33 056	33 222	34 616	34 095	34 754	35 059	32 645	30 597	31 886
	3º Ciclo e Secundário	80 619	80 031	81 063	81 724	82 867	81 626	82 099	84 404	84 102	82 415	83 794
Pessoal não docente		68 874	74 709	80 348	85 204	85 540	84 116	83 509	85 273	81 186	75 966	76 009
Estabelecimentos		16 556	16 598	16 730	16 467	16 357	15 783	15 105	14 313	14 074	12 510	11 837

As *Estatísticas da Educação – Jovens* constituem uma publicação que disponibiliza informação estatística referente às diferentes modalidades de educação e formação destinadas a jovens. Encontra-se organizada em dois capítulos: o primeiro permite obter uma visão global do sistema educativo; o segundo é reservado à difusão de dados estatísticos e encontra-se organizado em duas grandes áreas – uma referente a alunos matriculados, ordenados segundo os níveis de ensino e uma outra onde é apresentada uma evolução do número de alunos.

Alunos matriculados – Jovens, por nível e modalidade de ensino, em Portugal Continental (2000/01-2007/08)

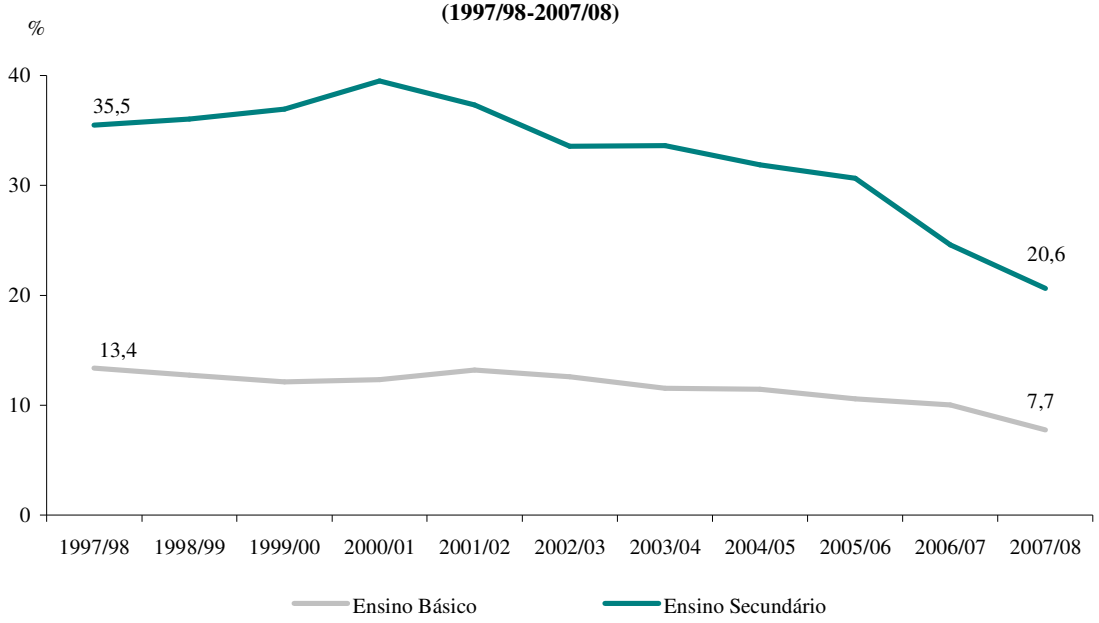
Nível e modalidade ou tipo de ensino \ Ano lectivo	2000/01	2001/02	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08
Total (Jovens)	**1636236**	**1599572**	**1579611**	**1579186**	**1579314**	**1572062**	**1597032**	**1614182**
Educação Pré-Escolar	**221407**	**226892**	**232555**	**238364**	**243921**	**246090**	**247826**	**250629**
Ensino Básico	**1093752**	**1072457**	**1059447**	**1054499**	**1047438**	**1062053**	**1072815**	**1078793**
1º Ciclo	**485517**	**473401**	**462685**	**459832**	**460502**	**465238**	**469402**	**468101**
Regular	485517	473401	462685	459470	460132	464814	469153	467851
Artístico Especializado (regime integrado)	-	-	-	362	370	424	249	250
2º Ciclo	**246336**	**248523**	**251866**	**250552**	**245028**	**238955**	**238431**	**242854**
Regular	246336	248523	251502	249550	244505	238345	237546	241639
Artístico Especializado (regime integrado)	-	-	-	273	247	225	254	259
Profissional	-	-	-	61	65	73	-	-
Cursos CEF	-	-	364	668	211	312	631	956
3º Ciclo	**361899**	**350533**	**344896**	**344115**	**341908**	**357860**	**364982**	**367838**
Regular	357900	347303	341952	338808	333765	342612	339724	322922
Artístico Especializado (regime integrado)	-	-	-	349	258	253	253	263
Profissional	971	681	504	1241	1749	1769	587	669
Cursos CEF (1)	3028	2549	2440	3717	6136	13226	24418	43984
Ensino Secundário	**321077**	**300223**	**287609**	**286323**	**287955**	**263919**	**276391**	**284760**
Regular	290984	267219	252998	250717	250081	226015	225189	208630
Cursos Gerais / Científico-Humanísticos	228179	210984	201158	199880	193085	176215	184854	185555
Cursos Tecnológicos	62805	56235	51840	50837	56996	49800	40335	23075
Artístico Especializado (regime integrado)	1629	1586	1513	1566	1685	1460	1838	1809
Profissional	28464	31418	30792	31346	33620	33341	44466	66494
Cursos CEF (2)	-	-	2306	2694	2569	3103	4898	7827

(1) Nos anos lectivos 2000/01 e 2001/02, corresponde aos CEFPI

(2) Nos anos lectivos 2002/03 e 2004/04 corresponde ao 10º ano – Via Profissionalizante

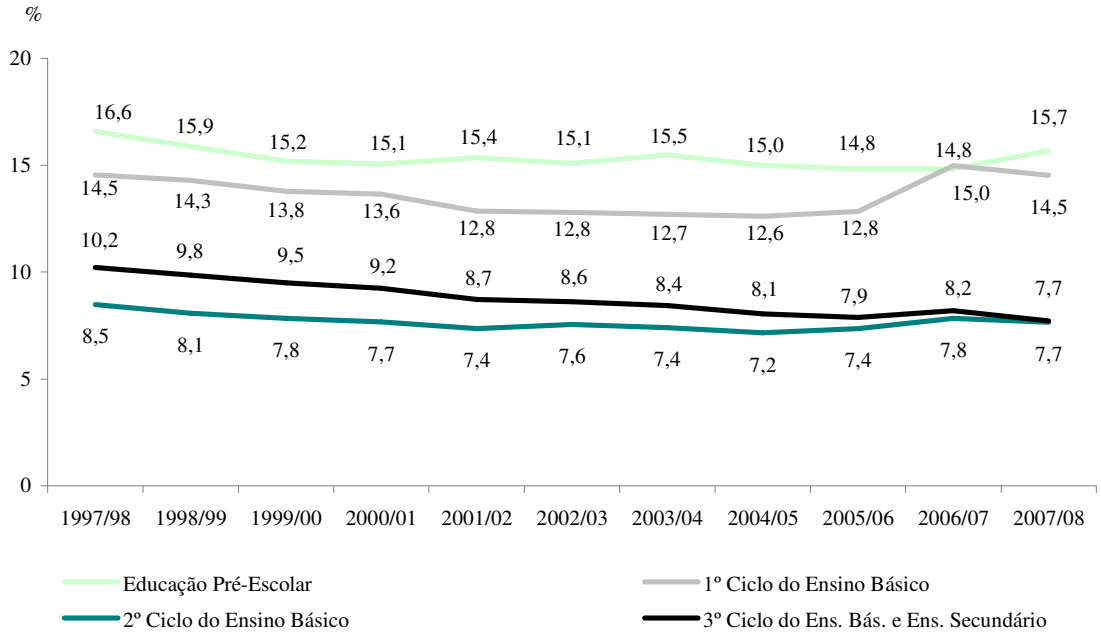
O *Perfil do Aluno* integra dados sobre a idade, sexo, nível de ensino, modalidade de ensino e resultados escolares.

Taxa de retenção e desistência, por nível de ensino, em Portugal Continental (1997/98-2007/08)



A publicação *Perfil do Docente* assenta num conjunto de indicadores que fornecem informação sobre a distribuição dos docentes, sobre as suas características individuais – idade, sexo e habilitações académicas – e acerca do exercício da profissão – grupo de docência, funções, vínculo e componente lectiva. A relação aluno/docente também é compilada. Os dados são sistematizados por nível de educação/ensino.

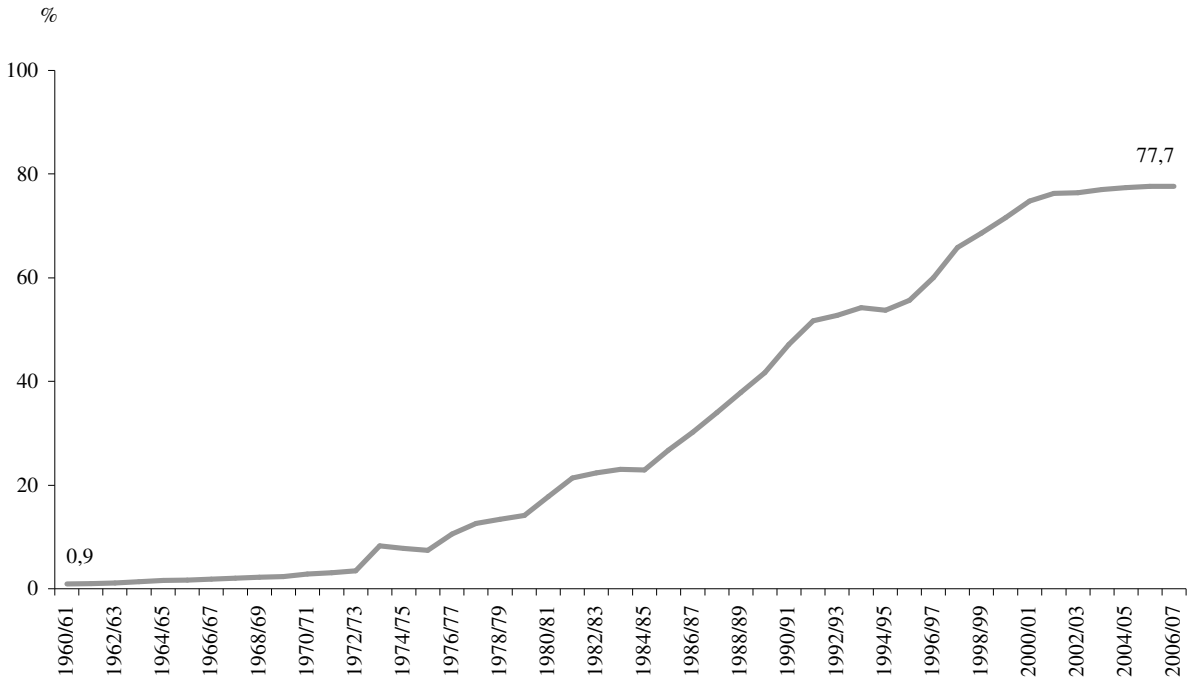
Relação aluno/docente, por nível de educação/ensino (1997/98-2007/08)



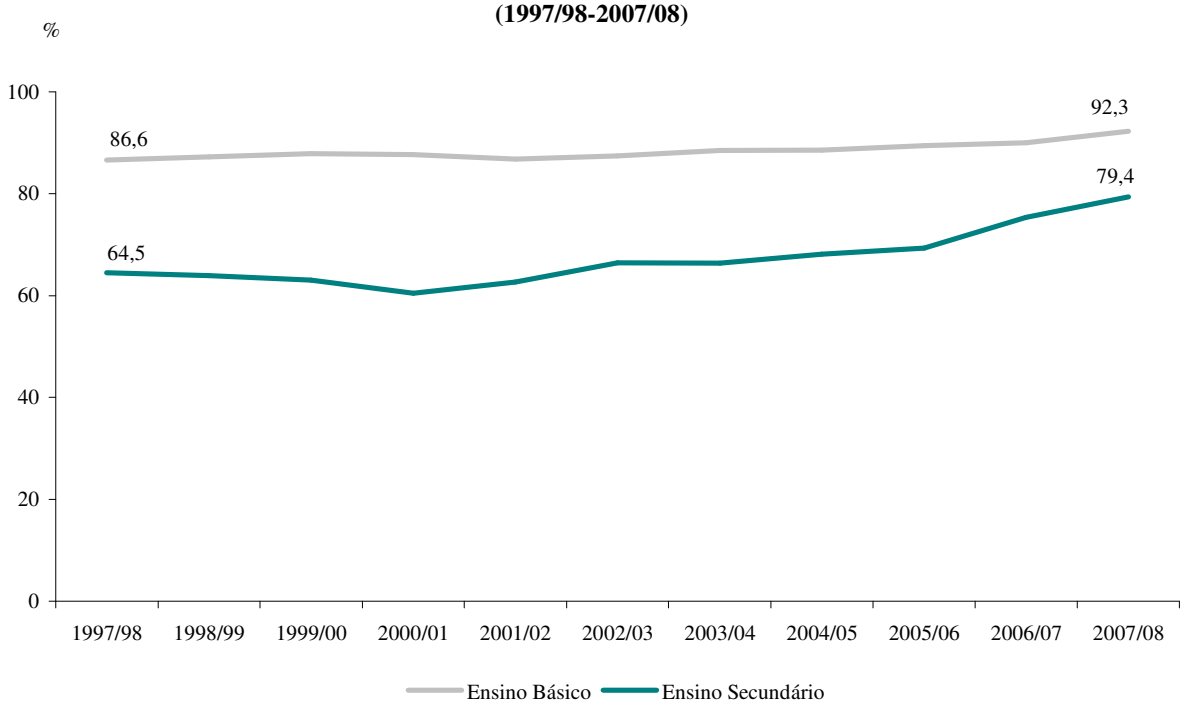
Através dos *Perfis Sectoriais do Docente* traça-se um perfil da população docente dos grupos de recrutamento do 3.º ciclo do ensino básico e do ensino secundário dominantes no sistema de ensino. Salientam-se os grupos de recrutamento de Matemática, Português, Física e Química, Biologia e Geologia e Educação Física.

A *Educação em Números* e a publicação *50 Anos de Estatísticas da Educação* apresentam uma análise histórica sobre a produção estatística relativa à educação e descrevem o sistema de educação em Portugal. Apresentam os valores absolutos de frequência escolar, de docentes e estabelecimentos. Englobam também indicadores de escolarização e de aproveitamento escolar.

Taxa real de pré-escolarização, em Portugal (1960/61-2006/07)



Taxa de transição/conclusão, por nível de ensino, em Portugal Continental (1997/98-2007/08)



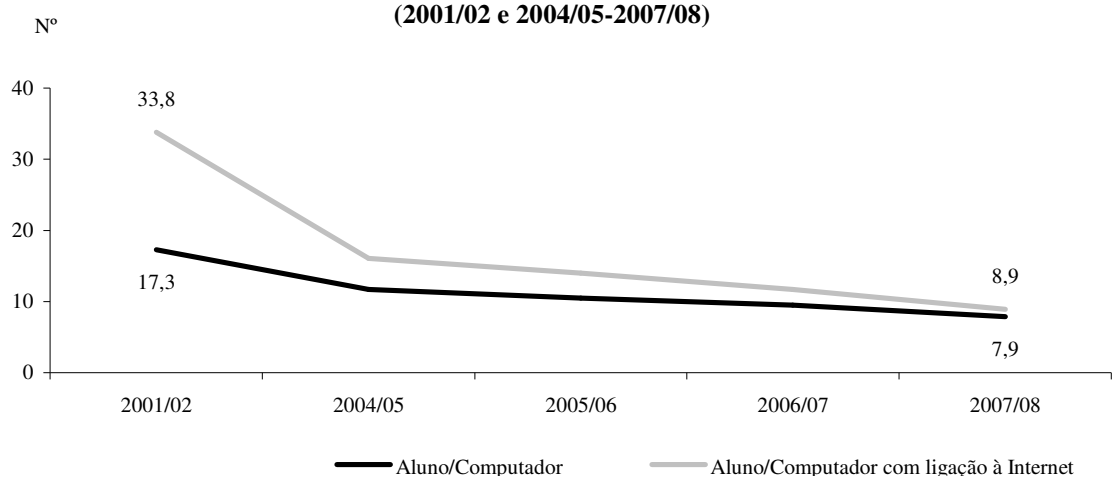
2.1.2. Modernização Tecnológica das Escolas

A Modernização Tecnológica das Escolas tem como base um inquérito anual do tipo recenseamento, através do qual se procede à recolha de informação estatística acerca da modernização tecnológica junto de todos os estabelecimentos de educação e ensino de naturezas pública e privada de Portugal Continental, que ministram os diferentes níveis de ensino, desde a educação pré-escolar até ao ensino secundário. Este inquérito integra questões sobre Internet, parque informático, equipamentos multimédia e cartão electrónico do aluno.

Através da recolha, tratamento e divulgação de informação estatística sobre a modernização tecnológica dos estabelecimentos de educação e ensino, o GEPE visa contribuir para o enriquecimento

da discussão sobre o papel das Tecnologias da informação e da comunicação (TIC) nas escolas e no sistema educativo português.

Relação aluno/computador e aluno/computador com ligação à Internet em escolas dos Ensinos Básico e Secundário Regular, em Portugal Continental (2001/02 e 2004/05-2007/08)



2.1.3. Actividades de Enriquecimento Curricular

As actividades de enriquecimento curricular têm como objectivo central adaptar os tempos de permanência das crianças nos estabelecimentos de ensino às necessidades das famílias, garantindo que os tempos sejam pedagogicamente ricos e complementares das aprendizagens associadas à aquisição das competências básicas.

Compete à DSE/GEPE recolher, tratar e divulgar informação estatística sobre as actividades de enriquecimento curricular desenvolvidas nos estabelecimentos de educação e ensino de natureza pública, que ministram o 1.º ciclo do ensino básico. Dá-se particular destaque ao inglês, às actividades físicas e desportivas, à música e ao apoio ao estudo[8]. Procede-se ao registo da evolução da oferta de actividades de enriquecimento curricular pelas escolas – tipo de actividade, número de alunos e de professores envolvidos, entidades parceiras e promotoras envolvidas.

Evolução do número de alunos abrangidos pelas actividades de enriquecimento curricular, por actividade, em Portugal Continental (2006/07-2008/09)

Actividade	Alunos abrangidos					
	2006/2007		2007/2008		2008/2009	
	N	%	N	%	N	%
Ensino do inglês nos 1.º e 2.º anos	63 988	30,5	75 622	37,1	166 135	85,3
Ensino do inglês nos 3.º e 4.º anos	185 247	88,8	184 282	88,2	185 642	88,8
Ensino da Música	273 178	65,3	264 678	64,1	274 764	68,0
Actividade física e desportiva	316 127	75,6	327 273	79,3	333 009	82,5
Apoio ao estudo	339 044	81,1	342 629	83,0	345 219	85,5
Outras*	x	x	215 567	52,2	172 705	42,8

* Inclui as expressões artísticas

x - valor não disponível

[8] Despacho n.º12 590/2006 de 16 de Junho e Despacho n.º 14460/2008 de 26 de Maio.

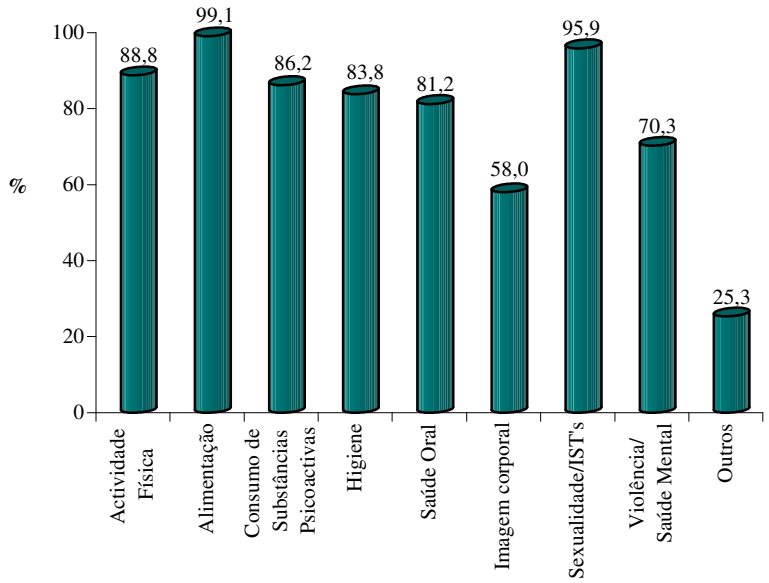
2.1.4. Promoção e Educação para a Saúde nas Escolas

Em contexto escolar, "Educar para a Saúde" consiste em dotar as crianças e os jovens de conhecimentos, atitudes e valores que os ajudem a fazer opções e a tomar decisões adequadas à sua saúde e ao bem-estar físico, social e mental. A adopção de medidas atinentes à promoção da saúde em meio escolar trata-se de uma acção que envolve um permanente desenvolvimento, visando contribuir para a aquisição de competências por parte da comunidade escolar, que lhe permitam confrontar-se confiada e positivamente consigo própria e, bem assim, fazer escolhas individuais conscientes e responsáveis, estimulado um espírito crítico e construtivo, verdadeiro pressuposto do exercício de uma cidadania activa.

Cabe à DSE/GEPE a recolha, preparação e produção da informação para o relatório sobre a avaliação da "Promoção e Educação para a Saúde nas Escolas". O inquérito associado incide sobre as áreas consideradas prioritárias, da alimentação e actividade física, consumo de substâncias, educação sexual/prevenção de infecções sexualmente transmissíveis (IST) e violência em meio escolar/saúde mental, e é aplicado nas escolas sede de agrupamento e nas escolas não agrupadas que ministram os ensinos básico e/ou o ensino secundário, de natureza pública.

Pretende-se desta forma assegurar o acompanhamento, monitorização e desenvolvimento das actividades da saúde em meio escolar, bem como analisar o impacto da implementação das medidas que visam avaliar e optimizar a promoção e educação para a saúde nas escolas.

Conteúdos de "Promoção e Educação para a Saúde abordados na escola
(% de respondentes, Portugal Continental 2009)



Nota: Pergunta de resposta múltipla. Num total de 1216 escolas verificou-se a participação de 1170.

2.2. Colaboração com entidades/projectos

A DSE/GEPE tem a responsabilidade de coordenar/participar/colaborar em diversos projectos, entre os quais:

- OECD Teaching and Learning International Survey (TALIS);
- OECD Learning to Learn;
- Observatório da Segurança Escolar;
- Observatório de Trajectos dos Estudantes do Ensino Secundário;

- Observatório do Plano Tecnológico da Educação;
- Observatório das Políticas Locais de Educação.

Nesta subsecção abordar-se-ão apenas alguns desses projectos, nomeadamente, o Estudo da OCDE sobre Ensino e Aprendizagem (TALIS), o Observatório da Segurança Escolar (OSE) e o Observatório de Trajectos dos Estudantes do Ensino Secundário (OTES).

2.2.1. OECD Teaching and Learning International Survey (TALIS)

O TALIS[9] constitui um estudo comparativo da OCDE sobre o ambiente de aprendizagem e as condições de trabalho que as escolas oferecem aos professores. Em Portugal, o trabalho de campo foi realizado pela DSE/GEPE nos primeiros meses de 2008, em que 200 escolas públicas e privadas com 3.º ciclo do ensino básico foram seleccionadas para participar, a partir de uma amostra representativa. Este estudo ofereceu aos docentes e directores/presidentes de conselho executivo a oportunidade de contribuírem para a análise da educação e das linhas de conduta da política educativa.

A análise cruzada dos dados a nível internacional permitiu aos países participantes identificar desafios semelhantes e aprenderem a partir de políticas públicas adoptadas por outros países. O inquérito foi realizado em 23 países, localizados em quatro continentes e o relatório internacional foi divulgado em 16 de Junho de 2009.

2.2.2. Observatório da Segurança Escolar (OSE)

A existência de condições de segurança na escola é fundamental para o sucesso educativo de todos os alunos, em especial daqueles que se encontram em meios particularmente desfavorecidos e em situação de risco de exclusão social e escolar.

O OSE foi criado com o objectivo de desenvolver métodos para a monitorização e avaliação da segurança nas escolas e produzir indicadores adequados ao conhecimento das situações de insegurança e violência nas escolas.

Neste sentido, desde 2006, que a DSE/GEPE, tem vindo a recolher e constituir uma base de dados relativa à participação de ocorrências nas escolas, como por exemplo situações de agressões, assaltos e roubos, uso ou venda de drogas, intimidações, incivilidades, indisciplina e destruição de bens.

Os dados recolhidos junto dos estabelecimentos de ensino, visam conhecer a realidade presenciada nas escolas no sentido de serem desenvolvidos planos de acção para reduzir a violência no meio escolar e envolvente, bem como para dinamizar iniciativas promotoras dos valores de cidadania e de civismo.

2.2.3. Observatório de Trajectos dos Estudantes do Ensino Secundário (OTES)

O OTES foi criado em 2006 pelo GEPE como uma estrutura de acompanhamento dos trajectos dos jovens no ensino secundário, do prosseguimento de estudos e/ou da sua inserção no mercado de trabalho. Integrado na Direcção de Serviços Estatística tem como meta fornecer ferramentas de diagnóstico, de monitorização e de avaliação que apoiem a tomada de decisão local e central no subsistema de ensino em causa. O OTES/GEPE prossegue desta forma com os objectivos principais de produzir e divulgar informação sobre os trajectos escolares e profissionais dos estudantes do ensino secundário e apoiar a tomada de decisão no âmbito da educação.

[9] Informação sobre o estudo em http://www.oecd.org/document/0/0,3343,en_2649_39263231_38052160_1_1_1_1,00.html

No sentido de alcançar as metas a que se propõe aplica questionários a alunos do ensino secundário em momentos distintos do seu percurso, que cobrem temáticas como as origens socioeconómicas dos estudantes, as suas escolhas e desempenho escolares, os seus trajectos profissionais e as suas práticas de cidadania.

Os dados recolhidos nos questionários são devolvidos às escolas por via electrónica, o que possibilita um melhor conhecimento sobre os trajectos da sua população estudantil, contribuindo para dar resposta às crescentes necessidades de diagnóstico, de acompanhamento e avaliação. Por seu lado, destes questionários resultaram até ao momento as publicações *Estudantes à Entrada do Ensino Secundário e Estudantes à Saída do Ensino Secundário*.

3. DESAFIOS PARA O FUTURO

As estatísticas da educação são uma actividade nuclear do GEPE e, como tal, são sempre um alvo privilegiado das acções de melhoria desta organização. Os últimos anos têm vindo a revelar uma enorme evolução nas operações estatísticas na área da educação. A utilização da Internet, de ferramentas automáticas de reporte da informação e sobretudo a utilização de dados administrativos significaram um considerável avanço nesta área, melhorando a qualidade dos dados e diminuindo o peso sobre as escolas. O futuro, mais ou menos próximo, obriga a que se continuem a aperfeiçoar estes mecanismos de transmissão de informação. A existência de uma base de dados administrativa única com toda a informação que actualmente ainda é recolhida através de inquéritos é um dos caminhos a seguir.

Enquanto tal cenário não é colocado em prática, pretende-se continuar a consolidar, assegurando a qualidade e a celeridade da produção e divulgação de estatísticas, os procedimentos de recolha, apuramento e divulgação da informação e abranger mais domínios. Pretende-se ainda apostar na produção de estudos estatísticos que proporcionem informação sobre temas específicos da educação ou retratos estatísticos ao longo de períodos alargados da história da educação em Portugal.

No âmbito da cooperação, o GEPE continuará a apoiar e acompanhar programas, projectos e iniciativas em vários domínios nacionais e internacionais, como por exemplo, o Conselho Superior de Estatística, a OCDE, a UNESCO, a CPLP, a União Europeia, entre outros.

Por fim, o maior desafio que terá de ser enfrentado em 2010, prende-se com a alteração da visão tradicional do sistema educativo. O surgimento de novas modalidade de educação e formação e de novas entidades que as ministram[10] obrigam a uma readaptação metodológica dos processos de recolha e apresentação da informação. Esse trabalho deverá ser desenvolvido com o envolvimento não só do GEPE, mas também do INE e de outras entidades com responsabilidades na área da educação e formação.

[10] Os processos de reconhecimento, validação e certificação de competências (RVCC) que decorrem em Centros Novas Oportunidades (CNO) são o exemplo mais emblemático.

Pós - Doc

Tese de doutoramento: Some Problems on Bayesian Hierarchical Modeling of non-Gaussian Spatio-temporal Data (Boletim SPE Outono de 2006, p. 26)

Jorge Mendes, *jorge.mendes@ine.pt*

Serviço de Métodos Estatísticos - DMSI
INE - Instituto Nacional de Estatística

Caros colegas e amigos,

Realizei as minhas provas públicas de doutoramento na Reitoria da Universidade de Lisboa, no dia 12 de Julho de 2006. É uma data que não se esquece! É como a data do nosso aniversário… fica cunhada na nossa memória, com um valor facial proporcional ao esforço dispendido, aos avanços e revezes sofridos e aos momentos de desânimo sentidos nos anos imediatamente anteriores. Mas calma, nem tudo é mau, pois nesse dia, após a comunicação da decisão do Júri, qual balança, o prato mais pesado sofre um repentino esvaziamento e o fiel inclina-se para o lado oposto. A partir desta altura só coisas "boas" e "menos boas" são recordadas. É, de facto é isso! Tudo o que foi sentido até ali como "mau", é agora recordado como menos bom. A tudo conseguimos colocar uma roupagem agradável que nos faz relembrar que os acontecimentos ou sentimentos desagradáveis tiveram a sua utilidade, pelo menos como provação que ultrapassámos, absolutamente. Não se inquietem os que ainda não viveram esta experiência pensado que estou a dizer umas coisas porque acho que ficam bem a quem ler. É mesmo verdade! E à medida que os anos vão passando, os pequenos acontecimentos desagradáveis ou são mesmo apagados pela nossa memória selectiva, ou são associados a ganhos, ainda que marginais, que temos vindo a ter.

Bem, não vos quero maçar com lembranças de há três ou quatro anos atrás, nem sequer com interpretações contemporâneas de realidades passadas. O facto é que o doutoramento muda as nossas vidas: pessoal e profissional. Ficamos mais seguros, afinal já chegámos à maioridade! Deixámos de ser recrutas e passámos a ser soldados rasos (pelo menos é essa agora a nossa realidade), com juramento de bandeira realizado. Isto é o que sentimos no nosso dia-a-dia.

Mas a nossa vida profissional deu ainda mais voltas! De repente somos tratados por outro título, com toda a carga reverencial que ele acarreta. Os alunos olham-nos de outro modo. Os funcionários já dobram a língua mais vezes quando pronunciam o nosso nome, pois afinal já tem que substituir o dr. pelo prof. antes do nome! Os outros, a quem tínhamos temor reverencial, são agora nossos pares (salvaguardadas as distâncias, sobretudo para o caso dos catedráticos). O assento nos conselhos científicos é agora uma realidade, com a fonte de tarefas que trás atrás de si: afinal, no seio dos conselhos científicos, somos apenas soldados rasos a quem são distribuídas tarefas que antes apenas nos pareciam meritórias e agora temos a plena noção que são nobres e muito consumidoras de tempo! Mas são necessárias! E a nossa opinião, mal ou bem, é agora considerada! A nossa visibilidade para o exterior aumenta igualmente. As instituições para quem trabalhamos reconhecem agora a legitimidade para exercer funções que nos eram vedadas, como a direcção de cursos de 1º ou 2º ciclo. Ou mesmo a coordenação de projectos de investigação ou projectos de desenvolvimento. E eles começam a aparecer nas nossas mãos. Trabalhamos neles tão afincadamente como anteriormente, com a grande

diferença de que somos agora os interlocutores directos. A própria confiança da restante sociedade civil em nós aumenta exponencialmente em cerca de três dias! Afinal não somos mortais comuns. Tivemos que passar provações que a generalidade dos mortais não passa! Com consequências, pois, felizmente, o sistema ainda conserva elementos de meritocracia! Surgem convites que nos enchem de orgulho, porque de repente provámos que somos capazes!

Comigo foi isto exactamente que aconteceu. Após cerca de dois anos em que a gratificante actividade de investigação científica era conduzida nos intervalos das funções docentes ou de gestão, surgiu um convite para dirigir a Metodologia do Instituto Nacional de Estatística. Trata-se de um lugar tradicionalmente ocupado por alguém com ligação à Academia e, por isso mesmo, deixou-me manifestamente contente. Ultrapassadas as dificuldades burocráticas, aceitei e aqui estou a dirigir uma equipa de cerca de 20 metodólogos que se dedicam a tarefas que, na Academia nem sonhamos que são essenciais à produção de estatísticas oficiais: o desenvolvimento e aperfeiçoamento de questionários para que possam medir exactamente o que se pretende recolher, o desenho de planos de amostragem complexos, em quase tudo diferentes dos planos teóricos, o dimensionamento de amostras com restrições que nos passam completamente ao lado nas disciplinas de amostragem, a resolução de problemas associados à recolha de informação, a procura de soluções metodológicas que sirvam a estrutura de recolha no campo e os objectivos para que as operações estatísticas são concebidas, a automatização de processos de apuramento rotineiros e extremamente consumidores de tempo, a garantia da observação das regras do segredo estatístico legalmente consagradas, a defesa permanente dos princípios de qualidade das estatísticas oficiais, etc.

Institucionalmente, "o INE, I. P., tem por missão a produção e divulgação da informação estatística oficial, promovendo a coordenação, o desenvolvimento e a divulgação da actividade estatística nacional". Os métodos estatísticos estão integrados no Departamento de Metodologia e Sistema de informação. Trata-se de uma unidade orgânica de 2º nível que não decorre directamente da Portaria que publicou os Estatutos do INE, I.P., mas de uma decisão gestionária do seu Conselho Directivo. Globalmente, compete a este departamento:
a) Apoiar científica e metodologicamente a produção estatística do Sistema Estatístico Nacional (SEN) e gerir o respectivo sistema de metainformação;
b) Criar um sistema geral de amostragem e desenvolver metodologias para controlo da carga estatística sobre os respondentes;
c) Certificar tecnicamente as operações estatísticas do SEN e outras que sejam submetidas ao INE, I. P., por outras entidades públicas;
d) Assegurar a gestão das classificações/nomenclaturas para uso no SEN;
e) Realizar o registo prévio dos instrumentos de notação, a utilizar na produção das estatísticas oficiais;
f) Assegurar a gestão, manutenção e coordenação do Sistema de Informação Geográfica do INE, I. P.;
g) Desenvolver um sistema integrado para processamento e utilização partilhada de dados estatísticos;
h) Desenvolver as soluções informáticas necessárias às actividades do INE, I. P.;
i) Coordenar e garantir a segurança informática, em particular a confidencialidade, integridade, disponibilidade e autenticidade;
j) Assegurar a gestão das infra-estruturas informática e de comunicações.

Ora é precisamente no "apoio científico e metodológico à produção estatística do Sistema Estatístico Nacional" e na "criação de um sistema geral de amostragem e desenvolvimento de metodologias para controlo da carga estatística sobre os respondentes" que se insere a competência dos Métodos Estatísticos. Materializando, aos Métodos Estatísticos compete (a) a definição de Universos e constituição de Bases de Amostragem das diversas operações estatísticas, quer junto das famílias, quer junto de instituições (e.g. empresas), (b) a selecção de amostras em planos de amostragem complexos, respeitando os princípios de redução da carga estatística e acompanhamento dos problemas levantados durante a recolha, providenciando, em conjunto com o departamento de matéria responsável pelo inquérito soluções cabais, (c) o registo das características e opções metodológicas no documento metodológico que descreve cada operação e o respectivo apoio ao departamento de matéria, e (d) o tratamento dos dados recolhidos, no que diz respeito à ponderação, realização de alguns apuramentos e

cálculo de medidas de erro. Como é natural abordar-se na disciplinas de amostragem, a maior parte dos problemas advêm, quer das desactualizações das bases de amostragem, quer da falta de cooperação dos respondentes. No INE estes problemas não são uma excepção uma vez que as bases de amostragem não são perfeitas e os dados recolhidos padecem de várias "doenças". Esta realidade obriga à avaliação constante de várias opções metodológicas e à tomada de decisões vitais, à luz da manutenção de altos padrões de qualidade da informação estatística produzida pelo INE, no quadro do SEN. Tomemos consciência que ouvimos diariamente o nome do INE ser pronunciado nos órgãos de comunicação social como fonte ou da taxa de desemprego, ou do indicador de clima económico, ou do indicador da produção industrial, etc. e que essas informações são vitais na tomada de decisões dos agentes económicos. Afinal, o INE tem por atribuição "produzir informação estatística oficial, com o objectivo de apoiar a tomada de decisão pública, privada, individual e colectiva, bem como a investigação científica"

É um mundo tão diferente daquele que experimentamos na Academia e que nem sequer imaginávamos que existia. Acresce ao trabalho eminentemente técnico, a necessidade de coordenar uma equipa por onde passa a metodologia de quase todas as operações estatísticas por amostragem do INE. E esta é uma responsabilidade que na Academia não tem par. Mesmo a coordenação de uma equipa de investigação jamais se assemelha à coordenação de uma equipa deste género que serve de charneira às tarefas desenvolvidas pelos restantes 600 ou 700 técnicos do INE que, de Norte a Sul e Ilhas (nos serviços Regionais de estatística), lidam diariamente com a necessidade de produzir estatísticas oficiais nos mais variados domínios. Esta liderança está suportada numa estrutura orgânica e hierárquica da Administração Pública, em geral, e dos Institutos Públicos, em particular, que diverge significativamente da estrutura da Universidade.

Não sei o que pensarei desta experiência daqui a 20 anos, por exemplo, mas tenho que admitir que o enriquecimento pessoal e profissional é valioso e pesará, com certeza, no meu CV, embora tenha que reconhecer que, para quem cresce na Universidade, trata-se de uma realidade diferente e estranha. Desde que a área de metodologia do INE foi criada que quem a chefia, tradicionalmente ligada à Universidade, não fica mais do que 3 ou 6 anos. Comigo não será diferente, mas guardarei boas memórias e uma experiência que enriquecerá seguramente a actividade de investigação científica, que conto retomar logo que me seja possível, na minha Instituição, o Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa.

Este texto é o retrato subjectivo dos meus últimos anos após as provas de doutoramento. Outros colegas haverá que têm visões diferentes. Respeito-os e desejo a todos votos das maiores felicidades e conquistas, em qualquer dos domínios em que exerçam actividade.

Saudações académicas,

Jorge M. Mendes

Tese de doutoramento: Modelação Estatística da Qualidade das Águas de Superfície da Bacia Hidrográfica do Rio Ave (Boletim SPE Outono de 2006, p. 26)

Arminda Manuela Gonçalves, *mneves@mct.uminho.pt*

Departamento de Matemática e Aplicações
Universidade do Minho

1. INTRODUÇÃO

A água é um bem precioso e potencialmente indutor de riqueza. Numa região como a do Vale do Ave, cuja base económica está fortemente ligada à indústria (onde predomina a indústria Têxtil), a água é sem dúvida um factor determinante na localização industrial neste vale.

A Bacia Hidrográfica do Ave, situada no Noroeste de Portugal (Fig. 1), tem uma superfície aproximada de 1390 Km^2, e o seu curso principal, o rio Ave, corre ao longo de 101 Km desde a sua nascente na Serra da Cabreira, até à foz em Vila do Conde. Como principais afluentes que constituem a rede hidrográfica do rio Ave destacam-se os rios Este, na margem direita, e os rios Selho e Vizela, na margem esquerda.

Fig. 1 - Enquadramento geográfico da Bacia Hidrográfica do rio Ave

A região do Vale do Ave caracteriza-se por um povoamento muito disperso, mas com elevada densidade populacional. Na Bacia Hidrográfica do rio Ave, os cursos de água apresentam, de um modo geral, uma situação de clara degradação ambiental, com graves perturbações tanto a nível físico-químico como biológico, com a excepção para os sectores próximos das nascentes. O desfasamento entre o desenvolvimento socio-económico registado ao longo dos anos e a construção de infra-estruturas, traduz-se num estado de extrema poluição dos cursos de água.

Foi em meados da década de 70 que se começou a encarar com preocupação o aumento da poluição das águas na bacia hidrográfica do Ave, começando a organizar-se grupos de trabalho em que esta bacia foi escolhida como piloto para o ensaio de um modelo de gestão integrada de recursos hídricos em Portugal. Só em 1990 foi aprovada uma solução regional para a "zona mais poluída" (concelhos de Guimarães, de Vila Nova de Famalicão e de Santo Tirso), traduzida num "Estudo Prévio de Drenagem, Tratamento e Rejeição das Águas Residuais do Vale do Ave", com recursos aos fundos comunitários.

Cria-se assim o Sistema Integrado de Despoluição do Vale do Ave (SIDVA), que compreende três Estações de Tratamento de Águas Residuais (ETAR's), em cada um dos três concelhos cuja área foi considerada "zona mais poluída". A 1ª Fase da construção das ETAR's inicia-se em Dezembro de 1993, tendo sido concluída no início de 1997. Só em finais de 1998, início de 1999, a 2ª Fase da obra foi concluída.

Assim, na minha tese são desenvolvidas novas metodologias no sentido de permitir a modelação estatística da concentração de poluentes nas águas de superfície da Bacia Hidrográfica do rio Ave, bem como a avaliação do desempenho das Estações de Tratamento de Águas Residuais (ETAR's) instaladas, nesta bacia.

2. METODOLOGIAS

2.1 A análise dos dados

O conjunto de dados utilizados neste estudo diz respeito a valores mensais de 11 variáveis qualidade da água observados entre 1988 e 2003, em 20 estações de amostragem de Qualidade (Fig.2). A Figura 2 também apresenta a localização das três ETAR's na bacia hidrográfica. As variáveis de qualidade em estudo são: a Condutividade, os Sólidos Suspensos Totais, o pH, a Temperatura, o Oxigénio Dissolvido, a Oxidabilidade, a Carência Química de Oxigénio, a Carência Bioquímica de Oxigénio (5 dias), o Azoto Amoniacal, os Nitratos e as Coliformes Fecais. Estas variáveis são consideradas, de acordo com a Direcção de Serviços de Controlo de Poluição (DSCP), relevantes para a avaliação e previsão da qualidade das águas superficiais de um rio sujeito a descargas de efluentes. Os seus valores traduzem as concentrações de poluentes de fontes poluidoras de origem doméstica e/ou industrial.

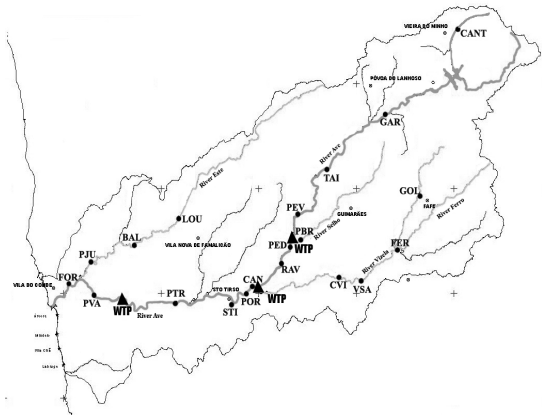


Fig.2 – Distribuição espacial das 20 estações de amostragem de qualidade e das 3 ETAR's.

Utilizando métodos da Estatística Descritiva efectua-se uma análise preliminar espaço-temporal das variáveis de Qualidade da água de forma a identificar os trechos do rio Ave e dos seus principais afluentes especialmente críticos, identificar as variáveis de Qualidade com os piores e os melhores comportamentos e detectar padrões de comportamento ao longo do tempo observado. Para a generalidade das variáveis de Qualidade verifica-se que a variabilidade cresce a partir da estação de amostragem de Qualidade das Taipas (TAI), atingindo o máximo na zona entre a confluência com o rio Selho e a estação de amostragem de Qualidade de Ponte da Trofa (PTR). A partir daí decresce sensivelmente até à foz.

No seu percurso em direcção ao mar, o rio Ave é engrossado por uma série de afluentes que para além de contribuírem para o aumento do seu caudal, fazem aumentar a sua carga poluente.

2.2 A análise de *clusters*

Aplicou-se uma análise de *clusters*, numa variante apropriada, para a classificação do conjunto das 20 estações de amostragem de Qualidade da bacia em grupos homogéneos (no espaço e no tempo), com base nas variáveis de Qualidade que foram seleccionadas. Utilizámos métodos hierárquicos com o procedimento aglomerativo.

Na análise foram operados vários métodos hierárquicos aglomerativos. Os resultados obtidos pelos métodos da ligação completa, da ligação média e de *Ward* mostraram os grupos (*clusters*) com melhores definições. Os *clusters* obtidos por estes três métodos foram praticamente os mesmos, foram discutidos os resultados obtidos pelo método da ligação completa.

Teve-se que construir matrizes de dissemelhança baseadas numa "nova medida de dissemelhança", para ser aplicada a duas quaisquer estações de amostragem. Essa "nova distância" goza das mesmas propriedades da distância euclidiana mas apresenta vantagens quando há dados omissos, como no nosso caso.

Como as variáveis não se apresentam definidas na mesma unidade de medição e apresentam variabilidades muito distintas, a análise de *clusters* foi efectuada para as 11 variáveis em separado e para as 8 variáveis que podem ser medidas em mg/l.

A análise de *clusters* hierárquica permitiu assim destacar, nesta bacia hidrográfica, a existência de 4 grupos de estações de amostragem de Qualidade classificados quanto à qualidade da água (para as 8 variáveis medidas em mg/l) como: praticamente Sem Poluição, Fracamente Poluída, Poluída e Muito Poluída.

As estações de amostragem de Qualidade aparecem, em geral, agrupadas de acordo com a proximidade geográfica, que está associada à proximidade de áreas (regiões) com características semelhantes. Características de diferentes naturezas como de as do meio bio-físico, as sócio-económicas resultantes das diferentes concentrações urbanas e/ou industriais, mas que contribuem para que em termos da qualidade da água haja homogeneidade.

2.3 O *Kriging* na análise dos recursos hídricos

O comportamento espaço-temporal de uma variável de qualidade está associada à variação do caudal (um aumento de caudal representa, usualmente, uma diminuição da concentração de poluente por diluição) e esta acompanha, geralmente, a variação sazonal da precipitação.

O número e localização das estações com medições de caudal na bacia hidrográfica do rio Ave não é o mais apropriado, existindo zonas significativas da bacia que não se encontram monitorizadas. No nosso estudo temos a necessidade de dispor de medições mensais médias em área, dependentes da quantidade de precipitação que cai sobre uma certa região geográfica e que influenciam um determinado local da bacia hidrográfica do rio Ave. Para isso, aplicámos os métodos de Polígonos de *Thiessen* (abordagem determinística) e de *Kriging* Ordinário (abordagem estocástica) às séries das observações de Precipitação nas 19 estações de amostragem Metereológicas existentes na bacia, com o objectivo de estimar a intensidade média, em área, da precipitação mensal nos vinte locais das estações de amostragem de Qualidade que vão representar o factor hidro-metereológico na modelação (espaço-temporal) da qualidade da água.

Com o primeiro método definiram-se vizinhanças (áreas de influência) para cada uma das estações de amostragem de Qualidade e, com o segundo processo, foram obtidas estimativas, em área, da quantidade média de precipitação que cai sobre essas vizinhanças e que influenciam o factor hidro-metereológico das respectivas estações, num determinado mês.

É claro que todo este processo de estimação teve em consideração os vários estudos existentes sobre as disponibilidades hídricas duma bacia (ou sub-bacia) pela caracterização indirecta via precipitação, bem como foi pedida a opinião de especialistas do Instituto Nacional da Água acerca da metodologia usada, que a consideraram uma metodologia "realista" e bem conseguida para a resolução do nosso problema.

2.4 A aplicação dos modelos lineares

Após a análise de *clusters* efectuada ajustam-se Modelos Lineares às 11 variáveis de qualidade observadas nos grupos homogéneos classificados de Poluído ou Muito Poluído, seleccionando-se aqueles que as melhor descrevem e explicam, ao longo do tempo. Pretende-se também estudar tendências e padrões na evolução das séries no período antes e depois de Setembro 1999 para que, desta forma, se possa avaliar o desempenho das três ETAR's, desde então activadas. Para isso, os valores das séries são divididos, no período antes e depois de Setembro de 1999, e o estudo de cada uma das partes é efectuado não independentemente uma da outra, mas sim na sua totalidade.

Assim, o modelo inicial que aplicamos às séries associadas a um *cluster* s é $Z_s(t) = T_s(t) + S_s(t) + C_s(t) + \varepsilon_s(t)$ (1), $t = 1,...,N$. O modelo decompõe-se numa componente determinística de tendência $T(t) = a_1 + b_1 t I_{1,t} + a_2 I_{2,t} + b_2 t I_{2,t}$ (que considera os dois períodos antes e depois de Setembro de 1999), numa componente determinística sazonal $S(t)$, numa componente $C(t)$, o factor hidro-metereológico, que representa a influência das covariáveis $P(t-2)$, $P(t-1)$ e $P(t)$: a quantidade média de precipitação, em área, em tempo t, $t-1$ e $t-2$. Finalmente, a componente estocástica $\varepsilon(t)$ que pretende-se que seja um mero ruído branco.

Aplicou-se o método dos mínimos quadrados como processo de estimação dos parâmetros, pois possui um suporte teórico consistente, e o reconhecimento de propriedades óptimas dos estimadores obtidos por este método, nomeadamente BLUE e UMVUE.

A sazonalidade bem como a componente $C(t)$ são sempre calculadas considerando toda a série, isto é, admitimos que os seus coeficientes como variáveis explicativas independentes se mantêm os

mesmos, no primeiro e no segundo período observado. Como a tendência $T(t)$ é linear incluindo um termo constante, foi necessário reparametrizar o modelo.

Às séries das variáveis de qualidade aplica-se o modelo completo (1) com todas as componentes e retira-se, posteriormente, uma a uma, as variáveis associadas aos mais altos valores-p para a estatística de teste t, até obtermos todas as variáveis regressoras com um efeito significativo (consideramos $p<0,05$). O modelo reduzido final também foi testado contra o modelo completo com ajuda de um teste-F.

As séries apresentam diferentes tendências (no primeiro e no segundo período), e, por vezes, essa tendência muda de sinal, verificando-se um decréscimo (ou acréscimo) ao longo do tempo da variável de qualidade após 1999 (após a entrada em funcionamento das ETAR's). São estatisticamente significativas a maioria das componentes periódicas de 12 meses da sazonalidade para a explicação do comportamento das variáveis de qualidade em que, os meses de Inverno, em geral, contribuem para uma diminuição do nível da variável e os meses de Verão, contrariamente, contribuem para um aumento dos seus níveis. Assim como, para algumas variáveis a componente $C(t)$ é também estatisticamente significativa.

3. CONCLUSÕES

Pudemos concluir que a situação da qualidade da água desta bacia até Setembro de 2003, na maioria dos trechos fluviais (*clusters*), não evoluiu tão favoravelmente como o esperado após a entrada em funcionamento das ETAR's. Como não dispusemos de dados analíticos a partir desta data, que nos permitissem documentar diferenças na avaliação depois desse período, pudemos chegar à conclusão que o período considerado após activação do SIDVA pode ter sido demasiado pequeno para que fosse visível uma melhoria efectiva da qualidade da água. De facto, sabemos que as ETAR's não começaram a funcionar com a sua capacidade total, tendo esta aumentado gradualmente.

Do nosso estudo parece claro que os significativos esforços que têm vindo a ser desenvolvidos nos últimos anos, associados à construção do SIDVA, não se traduzem ainda em resultados analíticos sobre a qualidade da água na rede hidrográfica.

Espera-se que, com este trabalho na área da Estatística Aplicada, se tenha contribuído para a discussão e compreensão de um problema ambiental, tão importante para a comunidade envolvente, que é o problema do controlo da qualidade das águas de superfície da Bacia Hidrográfica do rio Ave.

Agradecimentos

A autora agradece ao Sr. Eng. Pimenta Machado da Direcção Regional do Ambiente e Recursos Naturais/Norte (DRARN), à Sr.ª Eng.ª Cláudia Brandão do Instituto Nacional da Água (INAG) e à Sr.ª Professora Teresa Amorim do Departamento de Engenharia Têxtil da Universidade do Minho, por me ter disponibilizado os dados das várias redes de medição da bacia hidrográfica em estudo, bem como outras informações que foram preciosas para o desenvolvimento deste trabalho.

Referências

El-Sharaawi, A. (1995). Trend detection and estimation with environmental applications. *Mathematics and Computers in Simulation* **39**: 441-447.

Esterby, S. (1993). Trend analysis methods for environmental data. *Environmetrics* **4**: 459-481.

DRARN, INAG (1999). "Plano de Bacia Hidrográfica do Rio Ave." Projecto co-financiado pela Comunidade Europeia - Fundo de Coesão. Porto.

METCALF e EDDY (1991). "Wastewater Engineering, Treatment, Disposal and Reuse". *McGraw-Hill*.

Smith, R.A., Alexander, R.B., Wolman, M.G. (1987). Water-Quality trends in the Nation's Rivers. *Science* **235**: 1607-1615.

Vega, M., Pardo, R., Barrado, E., Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research* **32**: 3581-3592.

Tese de doutoramento: Processos Auto-Regressivos de Coeficientes Aleatórios na Modelação de Dados de Contagem (Boletim SPE Outono de 2006, p. 26)

Dulce Gomes, *dmog@uevora.pt*

Departamento de Matemática e Centro de Investigação em Matemática e Aplicações
Universidade de Évora

Vou abordar o estudo de modelos de séries temporais de valores inteiros não-negativos, também designadas por séries de contagem. Mais concretamente, vou centrar-me no estudo de uma classe de modelos auto-regressivos de ordem 1, de coeficientes aleatórios e baseados numa generalização da operação binomial *thinning*, proposta por Steutel e Van Harn em 1979. Este operador, representado usualmente por $*$, foi, na sua forma original, definido do seguinte modo:

Definição 1. *Seja* $\{U_i\}_{i\in N}$ *uma sequência de variáveis aleatórias independentes e identicamente distribuídas com distribuição de Bernoulli de valor esperado* α*, independente da variável aleatória Y inteira e não-negativa. Define-se a operação* $*$ *entre o parâmetro* α *e a variável Y como*

$$\alpha * Y = \begin{cases} \sum_{i=1}^{Y} U_i, & Y > 0 \\ 0, & Y = 0. \end{cases}$$

McKenzie (1985) definiu um modelo para séries temporais de valores inteiros não-negativos baseado no operador binomial *thinning* com o objectivo de modelar séries temporais com determinadas distribuições marginais. Com esse objectivo, McKenzie tomou um modelo do tipo AR(1) ($Y_t = \phi Y_{t-1} + \varepsilon_t$, $|\phi| \le 1$), uma sequência de variáveis aleatórias i.i.d., $\{\varepsilon_i\}$, e substituiu no modelo AR(1) a operação multiplicação pelo operador binomial *thinning* obtendo o denominado modelo INAR(1)

$$Y_t = \alpha * Y_{t-1} + \varepsilon_t, \ \alpha \in [0,1].$$

Ao substituir no modelo AR(1) a operação multiplicação pelo operador binomial *thinning*, para além de garantir que as variáveis apenas tomam valores nos inteiros não-negativos, McKenzie considera (dada a história passada do processo) os indivíduos que permanecem no sistema como uma variável aleatória. Ou seja, em cada instante de tempo *t*, cada indivíduo tem uma probabilidade α de permanecer no sistema e uma probabilidade $1-\alpha$ de o abandonar. Cada indivíduo permanece ou sai do sistema de um modo independente. Deste modo, o número de indivíduos no sistema no instante *t* depende dos indivíduos que permanecem no mesmo (com probabilidade α) mais os "novos" elementos que entram no sistema. O parâmetro α neste modelo de valores inteiros é, portanto, análogo ao parâmetro ϕ no modelo AR(1). Ou seja, enquanto que no modelo AR(1) um dos parâmetros a estimar é a proporção ϕ de indivíduos que permanecem no sistema, no modelo proposto por McKenzie esse parâmetro corresponde à probabilidade de um indivíduo, ao acaso, permanecer no sistema.

Tendo como ideia base este tipo de modelos, foi proposta uma nova classe de modelos. No essencial, tinha-se como finalidade construir um modelo que permitisse incorporar variáveis explicativas e que contemplasse a sobredispersão existente nos dados.

Nesta nova classe de modelos, a operação binomial *thinning* foi substituída pela operação *thinning* generalizada — representada por $\circ^G$, com *G* uma dada distribuição do tipo discreto, e associada com a operação — e o coeficiente do modelo, α, (inicialmente uma constante) substituído por uma variável aleatória.

A operação *thinning* generalizada entre duas variáveis aleatórias α e *Y* foi definida do seguinte modo:

Definição 2. *Dadas duas variáveis aleatórias,* α *e Y, e uma família de funções de distribuição* $G(\mu,\sigma)$ *parametrizada em* μ *(valor médio) e* σ *(desvio padrão) define-se a operação* $\circ^G$ *mediante a seguinte condição para a variável aleatória* $\alpha \circ^G Y$

$$\alpha \circ^G Y | \alpha, Y \cap G(\alpha Y, \sigma).$$

Ou seja, $\alpha \circ^G Y$ condicional ao conhecimento dos valores de α e de *Y* tem função de distribuição *G* de valor médio $\mu = \alpha Y$ e desvio padrão σ que também poderá depender de α e de *Y*.

De modo a construir o modelo, começa-se por tomar uma sucessão de variáveis aleatórias i.i.d. de valores inteiros não negativos $\{\varepsilon_t\}_{t\in Z}$ e uma outra sucessão, independente da anterior, de variáveis aleatórias não negativas $\{\alpha_t\}_{t\in Z}$. Define-se então o processo $\{Y_t\}_{t\in Z}$ através da equação às diferenças, estocástica,

$$Y_t = \alpha_t \circ^G Y_{t-1} + \varepsilon_t, t \in Z,$$

onde

i. para cada *t*, a variável $\alpha_t \circ^G Y_{t-1} | \alpha_t, Y_{t-1}$ é independente de Y_{t-1-k}, α_{t-k} e ε_{t-k}, $\forall k \in Z_+$;

ii. a variável Y_t, dado α_t, é independente de α_{t+k} e ε_{t+k}, $\forall k \in Z_+$.

Dentro desta nova classe de modelos foram considerados dois casos distintos: no primeiro caso, os coeficientes do modelo são eles próprios também um processo estocástico com uma dada estrutura de dependência e, no segundo caso, considerou-se que os coeficientes constituem uma sucessão de variáveis aleatórias i.i.d., sendo o modelo assim definido um caso particular do primeiro.

Estes modelos foram designados, respectivamente, por modelos generalizados auto-regressivos de ordem 1 duplamente estocásticos, abreviadamente DSINAR(1) generalizado (acrónimo de *Doubly Stochastic INteger AutoRegressive*) e modelos generalizados auto-regressivos de ordem 1 de coeficientes aleatórios, abreviadamente RCINAR(1) generalizado (*Random Coefficient INteger AutoRegressive*).

Relativamente a este tipo de modelos, foram deduzidas as condições necessárias para a estacionariedade fraca de ambos os processos (DSINAR(1) e RCINAR(1) generalizados). Foram obtidos os momentos e as funções de autocovariância. Derivou-se os estimadores dos parâmetros dos modelos e foram estabelecidas as propriedades assimptóticas destes. Foram igualmente efectuados estudos por simulação de modo a avaliar o desempenho dos vários estimadores (Gomes e Canto e Castro (2009)). No que diz respeito ao estudo do ajustamento do modelo, foi também abordada a metodologia *bootstrap* para a obtenção de intervalos de confiança para os parâmetros dos modelos e para as funções de autocorrelação e autocorrelação parcial dos resíduos (Gomes (2005)).

Alguns dos principais resultados alcançados:

- No processo RCINAR(1) Generalizado não existe a necessidade de se impor (para que estes sejam estacionários) que α_t tenha suporte em (0,1). Pelo que se torna possível utilizar este tipo de modelos em situações em que as séries apresentem ocasionalmente grandes picos.
- O processo RCINAR(1) generalizado é um processo AR(1) de coeficiente $E[\alpha_t]$ e com inovações ruído branco.

- Os processos DSINAR(1) e RCINAR(1) generalizados são apropriados para modelar séries temporais não-lineares de valores inteiros não-negativos, permitem incorporar variáveis explicativas e permitem contemplar a sobredispersão dos dados.
- Não são necessárias amostras de grande dimensão para se obterem boas estimativas dos parâmetros de ambos os modelos. Contudo, para amostras de dimensão moderada, o estimador de máxima verosimilhança condicional é mais eficiente e produz estimativas menos enviesadas.
- O processo DSINAR(1) generalizado revelou-se adequado para modelar o número de óbitos registados diariamente em Évora, usando como variáveis explicativas as temperaturas máximas e mínimas diárias. O modelo postulado conseguiu captar grande parte da variabilidade dos dados.

Alguns avanços dentro desta linha de investigação:

- Encontra-se em fase de estudo a dedução das condições necessárias para a estacionariedade forte dos processos DSINAR(1) e RCINAR(1) generalizados.
- Análise do comportamento extremal do processo RCINAR(1) generalizado. Nomeadamente, foram identificadas as condições para que o máximo, convenientemente normalizado, convirja para um limite não degenerado e caracterizou-se esse limite (Gomes, Temido e Canto e Castro (2009)).

Outras linhas de investigação:

- Estudos de metodologias de *clustering* espácio-temporais na caracterização de algumas doenças, em particular, da tuberculose e das anomalias congénitas em Portugal (Nunes *et al*. (2008)).
- Identificação de alguns dos condicionantes do método de *clustering* espácio-temporal, baseado na *Spatial Scan Statistics*, e propostas de possíveis soluções (Nunes e Gomes (2008) e Gomes e Nunes (2009)).

Referências

Gomes, D. (2005). Processos Auto-Regressivos de Coeficientes Aleatórios na Modelação de Dados de Contagem. *Tese de Doutoramento em Matemática, Universidade de Évora*.

Gomes, D. e Canto e Castro, L. (2009). Generalized integer-valued random coefficient for a first order structure autoregressive (RCINAR) process. *Journal of Statistical Planning and Inference, 139, pp. 4088-4097*.

Gomes, D., Temido, M. G. e Canto e Castro, L. (2009). Comportamento extremal de processos de valores inteiros RCINAR(1) generalizados. *XVII Congresso Anual da Sociedade Portuguesa de Estatística (SPE), Sesimbra. Programa e Livro de Resumos, p.276*.

Gomes, D. e Nunes, C. (2009). Spatial Scan Statistics: novos desenvolvimentos. *XVII Congresso Anual da Sociedade Portuguesa de Estatística (SPE), 30 de Setembro a 3 de Outubro de 2009, Sesimbra. Programa e Livro de Resumos, p.147*.

McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bulletin, 21, 645-50*.

Nunes, C. e Gomes, D. (2009). Processo de detecção de aglomerações espácio-temporais: alguns condicionantes. *Estatística. Arte de Explicar o Acaso. Actas do XVI*

Congresso da Sociedade Portuguesa de Estatística. (I. Oliveira, E. Correia, F. Ferreira, S. Dias, C. Braumann, eds.), Edições SPE, Lisboa, pp. 477-488.

Nunes, C., Briz, T., Gomes, D. e Matias Dias, C. (2008). A dimensão espácio-temporal em saúde pública: da descrição clássica à análise de clustering. *Revista Portuguesa da Saúde Pública. Vol. 26, Nº1, Lisboa, pp.5-14*.

Steutel, F. and Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Prob. 7, 893-99*.

Ciência Estatística

• Artigos Científicos Publicados

Aars, J., Marques, T.A., Buckland, S.T., Andersen, M., Belikov, S., Boltunov, A. & Wiig, O. (2009). Estimating the Barents Sea polar bear subpopulation size. *Marine Mammal Science*. 25: 35-52.

Abrantes, D., Pontes, L., Pinheiro, M. F., Andrade, M. and Ferreira, M. A. M. (2008). Towards a Systematic Probabilistic Evaluation of Parentage Casework in Forensic Genetics: a Modest Attempt to Define a General Standardized Approach to Simple and Complex Cases**.** *Forensic Science International: Genetics Supplement Series* 1, 635-637.

Andrade, M. and Ferreira, M. A. M. (2009). A note on Dawnie Wolfe Steadman, Bradley J. Adams, and Lyle W. Konigsberg, "Statistical Basis for Positive Identification in Forensic Anthropology. American Journal of Physical Anthropology 131:15-26 (2006). *International Journal of Academic Research (IJAR)*, 1 (2), 23-26.

Gomes, M. I., Henriques-Rodrigues, L., Pereira, H., Pestana, D. (2009). Tail index and second order parameters' semi-parametric estimation based on the log-excesses. *J. Statistical Computation and Simulation* (acceptd, 2009) DOI 10.1080/00949650902755178).

Gomes, M. I., Caeiro, F., and Pestana, D. (2009). A note on the asymptotic variance at optimal levels of a bias-corrected Hill estimator. *Statistics and Probability Letters* 79, 295–-303.

Gomes, M. I., and Pestana, D. (2009). A simple second order reduced bias tail index estimator. *J. Statistical Computation and Simulation*, 77, 487-504, 2009.

Knoth, S., M.C. Morais, A. Pacheco and W. Schmid (2009). Misleading signals in simultaneous residual schemes for the mean and variance of a stationary process. *Communications in Statistics - Theory and Methods.* Special Issue 'Celebrating 50 Years in Statistics Honoring Professor Shelley Zacks', Vol. 38, pp. 2923-2943.

Marques, T. A. (2009). Distance sampling: estimating animal density. *Significance*. 6: 136-137.

Marques, T. A., Thomas, L., Ward, J., Dimarzio, N. & P. L. Tyack (2009). Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales. *The Journal of the Acoustical Society of America*. 125: 1982-1994.

Mendes, S., Fernández-Gómez, M.J., Galindo-Villardón, M.P., Morgado, F., Maranhão, P., Azeiteiro, U. and Bacelar-Nicolau, P. (2009) Bacterioplankton dynamics in the Berlengas Archipelago (West coast of Portugal) using the HJ-biplot method. Arquipélago. *Life and Marine Sciences* 26, 25-35.

Mendes, S., Fernández-Gómez, M.J., Resende, P., Pereira, M.J., Galindo-Villardón, M.P. and Azeiteiro, U.M. (2009). Spatio-temporal structure of diatom assemblages in a temperate estuary. A STATICO analysis. *Estuarine, Coastal and Shelf Science* 84, 637-664.

Pereira, C., Bernardo, M., Pestana, D., Costa Santos, J. and Mendonça, M.C. (2010). Contribution of teeth in human forensic identification – Discriminant function sexing odontometrical techniques in Portuguese population *Journal of Forensic and Legal Medicine* Volume 17, Issue 2, 105-110.

Ramos, M. R., Carolino, E., Oliveira, T., Silva, A. P., Carvalho, R. and M. Bicho. (2009). Haptoglobin, acid phosphatase and demografic factors: obesity risk. *Biometrical Letters*, Vol. 46, nº1, 43-54.

Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L:, Strindberg, S., Hedley, S.L., Bishop, J.R.B. Marques, T.A. and Burnham, K.P. (2010). Distance software: design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* 47:5-14.

Valente, V., Oliveira, T.A.. (2009). Hierarquical Linear Models in Education Sciences: an application. *Biometrical Letters*, Vol. 46, nº1, 71-86.

• Capítulos de Livros

Gomes, M. I., D. Pestana, F. Sequeira, S. Mendonça, and S. Velosa (2009). Uniformity of Offsprings from Uniform and Non-Uniform Parents, in Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2009*, 31th International Conference on Information Technology Interfaces, 243-248.

Aleixo, S. M., J. Leonel Rocha and D.D. Pestana (2009). Dynamical Behaviour on the Parameter Space: New Populational Growth Models Proportional to Beta Densities, in Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2009,* 31th International Conference on Information Technology Interfaces, 213-218.

Pestana D., S. M. Aleixo and J. Leonel Rocha (2009). Hausdorff Dimension of the Random Middle Third Cantor Set, in Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2009*, 31th International Conference on Information Technology Interfaces, 279-284.

• Teses de Mestrado

Título: *A Estatística e as Probabilidades no Ensino Secundário: Análise dos Programas de Matemática A e B na perspectiva do Professor e dos Alunos*
Autora: Sara Cristina Baião Caldeira, *caldeira_sara@hotmail.com*
Orientadora: Maria Helena Mouriño Silva Nunes

Título: *Aplicação dos Modelos Lineraes Generalizados às Telecomunicações Móveis: caracterização dos clientes que desactivam os seus serviços*
Autora: Paula Figueiredo Mestre, *paula.figueiredo@vodafone.com*
Orientadora: Teresa Alpuim

Título: *O Ensino da Estatística e Probabilidade ao nível pré-graduado com recurso à folha de cálculo EXCEL*
Autora: Adelaide Proença, *adelaide.joao@sapo.pt*
Orientadora: Maria Eugénia Graça Martins

Título: *Cartões de Crédito*
Autor: João Curado Silva, *joao.curadosilva@millenniumbcp.pt*
Orientadores: Dinis Pestana e Ana Cristina Moita

Título: *Avaliação de metodologias de pre-processamento de dados de microarrays*
Autora: Ana Luísa Romão de São Marcos, *anasãomarcos@ua.pt*
Orientadoras: Adelaide Valente de Freitas e Gladys Castillo

Título: *Métodos de biclustering no problema de selecção de genes*
Autor: André Alexandre de Sebastião Marques,
Orientadoras: Gladys Castillo e Adelaide Valente de Freitas

Título: *Modelo Preditivo da Criminalidade - Georeferenciação ao Concelho de Lisboa*
Autor: Paulo Abel de Almeida João, *jolriao@gmail.com*
Orientadores: Victor Lobo e Fernando Bação

Título: *Modelos de Planos em Blocos Incompletos: Revisão e Perspectivas*
Autora: Paula Cecília dos Santos Leitão Caetano Alves, *paula.c.alves@netcabo.pt*
Orientadora: Teresa Oliveira

• Teses de Doutoramento

Título: Estimação em Pequenos Domínios com Modelos Espaciotemporais de Nível Área
Autor: Luís Nobre Pereira, *Lmper@ualg.pt*
Orientadores: Pedro Simões Coelho e Rui Sousa Nunes

Na minha tese foram apresentados desenvolvimentos metodológicos ao nível da estimação em pequenos domínios no âmbito das sondagens, quando os dados amostrais têm uma natureza espacial e cronológica. Foi proposto um estimador EBLUP (*Empirical Best Linear Unbiased Predictor*) assistido por um modelo linear misto de nível área, que permite especificar explicitamente a ligação entre os parâmetros de interesse e a informação auxiliar disponível. Foi também proposta uma metodologia que permite a introdução de restrições na estimação, garantindo a calibração das estimativas produzidas para diferentes níveis de agregação. Foram ainda propostos estimadores do Erro Quadrático Médio de Predição (EQMP) dos estimadores EBLUP, derivados pela metodologia delta e por metodologias por reamostragem. Todos estes desenvolvimentos metodológicos foram aplicados na estimação em pequenos domínios do preço médio de transacção da habitação em Portugal.
Na componente empírica da minha tese foram realizados dois estudos empíricos por simulação de Monte Carlo. No primeiro estudo, por simulação *design-based*, foi avaliada a qualidade dos estimadores propostos relativamente a outros estimadores habitualmente utilizados na estimação em pequenos domínios. Os resultados deste estudo permitiram concluir que os estimadores EBLUP são os que apresentam propriedades estatísticas de melhor qualidade. Em particular, o estimador EBLUP espaciotemporal proposto foi o que apresentou melhores propriedades *design-based*. Os resultados também permitiram constatar que a garantia da calibração das estimativas conduz ao aumento do enviesamento e a perdas de eficiência dos estimadores, comparativamente a um estimador equivalente que não garante essa calibração. Contudo, se essa calibração for garantida a um nível pouco agregado, então o efeito sobre a variância do estimador é quase insignificante.
No outro estudo, por simulação *model-based*, foi avaliado o desempenho dos estimadores do EQMP dos EBLUP temporal e espaciotemporal. Os resultados alcançados em ambos os casos permitiram concluir que os estimadores baseados em métodos por reamostragem apresentam um desempenho muito bom, quando comparado com o do respectivo estimador analítico.

Luís Nobre Pereira

PRÉMIO ESTATÍSTICO JÚNIOR 2010

Candidaturas até
28 DE MAIO DE 2010

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

SOCIEDADE PORTUGUESA
DE ESTATÍSTICA

PRÉMIOS "ESTATÍSTICO JÚNIOR 2010"

Está aberto, até 28 de Maio de 2010, o concurso para atribuição de prémios **"Estatístico Júnior 2010"**, de acordo com o seguinte regulamento:

1. A atribuição de prémios **"Estatístico Júnior 2010"** 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 2010"** devem ser alunos do 3.º Ciclo do Ensino Básico, do Ensino Secundário, ou dos Cursos de Educação e Formação de Adultos (EFA) no ano lectivo 2009/2010.

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 deverá ser enviado impresso em papel para efeitos da avaliação.

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2010"** 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 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 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 XVIII 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/BoletimCandidaturaPEJ10.pdf
http://www.spestatistica.pt/static/docs/RegulamentoPEJ10.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 2010"** 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 2010"** poderão não ser atribuídos.

Apoio da Porto Editora

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO SPE 2010

Está aberto, até **15** de **Junho** de **2010**, o concurso para atribuição do **Prémio SPE 2010**, de acordo com o seguinte regulamento:

1. Pretendendo dar destaque ao XVIII Congresso Anual da **SPE,** a principal reunião científica organizada pela Sociedade Portuguesa de Estatística, é instituído o **Prémio SPE 2010**.
2. Este prémio destina-se a estimular a actividade de estudo e investigação científica em Probabilidade e Estatística entre os jovens que trabalham nestas áreas.
3. O **Prémio SPE 2010** é constituído por uma quantia de 1000 euros.
4. Ao **Prémio SPE 2010** podem concorrer trabalhos originais sobre temas de Probabilidade e Estatística, desde que não tenham sido objecto de qualquer prémio atribuído por outra instituição.
5. Os autores dos trabalhos candidatos ao **Prémio SPE 2010** devem ser estudantes ou investigadores em alguma instituição portuguesa ou bolseiros portugueses, devem ser sócios da **SPE** e não devem ter completado os 35 anos de idade até 15 de Junho de 2010. Os autores não devem ter recebido o Prémio SPE nas quatro edições anteriores.
6. O trabalho deve ser escrito em português e não poderá exceder 25 páginas A4.
7. As candidaturas deverão vir acompanhadas do trabalho concorrente e do *curriculum vitae* dos autores e ser dirigidas ao Presidente da **SPE**, em carta registada, para a morada abaixo indicada. O carimbo do correio validará a data de entrega.
8. A admissibilidade e a apreciação dos trabalhos submetidos a concurso é da competência de um júri, cuja constituição será da responsabilidade da Direcção da **SPE**.
9. O júri é soberano nas suas decisões, não havendo lugar a recurso.
10. O trabalho galardoado com o **Prémio SPE 2010** será apresentado em sessão plenária pelo seu autor ou autores no XVIII Congresso Anual da **SPE** e será publicado nas respectivas Actas.
11. A atribuição do **Prémio SPE 2010** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será feita no XVIII Congresso Anual da **SPE** na sessão plenária da sua apresentação.
12. O **Prémio SPE 2010** poderá não ser atribuído.

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