Boletim

SPE
Sociedade Portuguesa de Estatística

Publicação semestral *outono de 2021*

Machine Learning e Inteligência Artificial

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

SPE

Sociedade Portuguesa de Estatística

https://www.spestatistica.pt

Sociedade Portuguesa de Estatística desde 1980

Junta-te à

SPE
Sociedade Portuguesa de Estatística

"Se já és sócio da SPE, incentiva os teus colegas, colaboradores e alunos a juntarem-se à SPE"

A SPE

- Oferece descontos em congressos e outros eventos organizados pela SPE.
- Oferece distinções e reconhecimento através dos seus prémios.
- Oferece oportunidades para ampliares a tua rede de contactos através da comunidade SPE.
- Oferece aos sócios acesso a um sistema de acreditação internacional.
- Valoriza sócios, comunidade e profissão apostando na educação, formação e inovação.
- Defende a profissão e molda o seu futuro.

Junta-te à SPE:

https://www.spestatistica.pt/socios/admissao-formulario

Quota anual

- Regular: 30€
- Estudante: 15€
- Promotor*: 0€
- Recém-licenciado†: grátis!

† Até um ano após terminar a licenciatura.
* Um "sócio promotor" angaria 2 novos sócios por ano.

Editorial

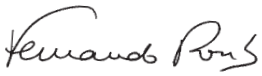
… o virtual, o presencial e o "*e-mundo*" na vida…

1. O Tema Central, como sabemos, é uma das várias secções do Boletim SPE. De algum modo, ele é o "nome próprio" de cada edição.
Para além das outras, que mantêm a sua designação, O Tema Central surge diferente em cada edição; e é à volta dele que os diversos autores constroem os textos que, gentilmente, dão corpo a esta publicação da SPE.
A Sociedade Portuguesa de Estatística, como uma associação científica, quer no seu conceito, quer no seu regulamento, tem "a língua de Camões" como instrumento de comunicação; cuja utilização deve ser incentivada.
É um estímulo desde logo para os jovens, que se iniciam no mundo científico e mais facilmente passam para outros patamares. É uma etapa de passagem para a internacionalização. E aqui surge de imediato, a escolha da língua inglesa e os desafios e todas as opções que tem de fazer-se.
A SPE, na feliz iniciativa de há já vários anos, criou a Comissão de Nomenclatura. É uma equipa liderada pelo colega Carlos Daniel Paulino. Esta em profícuo trabalho produziu o Glossário, disponível na página da SPE e que mantém muito útil e dinâmico.
Estas palavras continuam alguma reflexão que na SPE já se fez e que deve continuar a fazer-se.
Felicitada e aplaudida a iniciativa do Glossário, um instrumento que "legaliza" termos em português com o suporte da SPE; temos a vida prática onde, como nos Congressos, somos confrontados com um mundo multilingue. E surge então o momento em que é necessário compatibilizar, por exemplo, português e inglês. Um caso concreto surgiu na escolha do nome para o Tema Central da presente edição do Boletim SPE. Devíamos optar por Aprendizagem Automática (tal como mostra o glossário) ou por *Machine Learning*? Consultei colegas. O contexto fez-nos decidir pela segunda; embora sabendo que não é unânime a escolha. Também deve ser ponto de análise, a questão de saber qual o melhor sucesso para a divulgação. E este é um ponto prático muito importante. A maioria dos autores (portugueses) usa o termo inglês – como também se verifica nos textos dos nossos autores convidados para este Boletim.
Por outro lado, mesmo em plateias portuguesas e tal como diversos outros termos, por exemplo, usamos *outlier* sem prejuízo de o sabermos "valor discordante" em português. São muitos os termos que no dinamismo da língua vamos assimilando. Podemos mesmo dizer que incentivar a publicação, a comunicação, é o primeiro e principal objetivo. Cumprida essa missão, a obrigatoriedade de só usar termos em português, assim podemos concluir, ficará muito mais diluída.
Esta reflexão, que não é a primeira vez que abordamos, foi feita no momento da edição do anterior Boletim primavera 21 da SPE e onde, como se sabe e é hábito, divulgámos o título do Tema Central da Edição seguinte.
E parece que foi premonitória esta reflexão de primavera, que agora ficou continuada com a importante análise e proposta feita pela CENE – Comissão Especializada de Nomenclatura Estatística (e que mais adiante divulgamos). Com o equilíbrio e a ponderação requeridos, fundamentais para tão importantes decisões, é uma excelente proposta da CENE e que deve ser objeto da nossa maior atenção. O Boletim SPE está aberto e deseja registar esse debate. Até porque, também assim, se constrói o Memorial SPE.
2. Na mesma linha também as diversas opções que a pandemia tem exigido. E aqui obviamente surge o virtual versus o presencial.
Um expoente na SPE, é o XXV Congresso, o primeiro virtual, teve de ser adiado um ano e "decorreu em Évora", conforme relatamos nesta edição. Foi um excelente trabalho, embora com enormes desafios, a tantos níveis inovadores, para otimizar a comunicação científica, partilha de Ciência Estatística. É um mundo novo que se abre, em particular aos jovens, onde também devemos descobrir "a nossa opção" para melhor desempenhar a nossa tarefa: virtual ou presencial?
Na escrita de todos os dias, já surgem os "abraços virtuais" ou os "e-regards". Na realidade eles sempre foram virtuais, desde que não presenciais.
É o futuro e o despertar para uma "nova realidade".

Como muito bem diz o testemunho, por dentro e na prática, vivido pela COL do XXV Congresso da SPE e que felicitamos pelo excelente desempenho:
"O contacto humano e as interações sociais entre os participantes nunca poderão ser substituídos por esta realidade virtual. Temos muitas saudades dos encontros presenciais e esperamos que voltem em breve".
O fundamental é descobrir a excelência para transmitir e fazer progredir a Ciência Estatística.
3. Neste ano tão especial, a SPE desenvolveu intensa atividade e muitas ações concretas, desde a publicação do anterior Boletim. De entre elas, nesta edição, damos especial realce ao XXV Congresso da SPE, a entrega virtual do Prémio Estatístico Júnior que foi feita neste Congresso, a celebração do Dia Europeu da Estatística e a atribuição do Prémio Carreira SPE à Prof. Manuela Neves.

O júri do Prémio de Carreira da Sociedade Portuguesa de Estatística em 2021, por unanimidade, decidiu galardoar a Professora Manuela Neves, pelo seu currículo de excelência e pelo desempenho nas mais diversas ações a favor da SPE e da Ciência Estatística.
Amiga de longa data e companheira em muitas lides científicas, a Manuela, como todos muito bem sabemos, tem uma vasta e profícua atividade em prol da Estatística. É habitual e dedicada colaboradora do Boletim. Tenho o enorme prazer da sua amizade. Nas lides práticas da vida profissional, para além dos mais diversos júris que a atividade da Estatística e da SPE exige, é a organização de Congressos um dos seus "pontos extremos"; onde a sua experiência é longa e a classifica como "grande conselheira". No meu caso e acima de tudo isto, tive a felicidade de poder contar com a sua preciosa ajuda na equipa diretiva que constituímos para representar a SPE, nos primeiros anos deste milénio. Renovo a minha gratidão. Muitos parabéns e votos de boa continuação no desempenho da senioridade estatística que todos lhe reconhecemos.
Adiante damos notícia, com mais desenvolvimento e na secção Prémios, testemunhos, no final desta edição.

Embora com a falta do abraço presencial, a cerimónia de atribuição do prémio decorreu com os ingredientes costumados – desde o testemunho à gratidão, do reconhecimento à emoção, enfim, com tudo aquilo de que, na realidade, também se compõe a nossa passagem pela mãe natureza.
Assim (também) se Vive!



O Tema Central do próximo Boletim SPE será:

Liderança Estatística

Mensagem do Presidente

Caros Sócios da Sociedade Portuguesa de Estatística,

Gostaria em primeiro lugar de agradecer ao Professor Fernando Rosado por organizar um volume especial dedicado aos métodos de Aprendizagem Automática e de Inteligência Artificial. Conforme referido no Plano de Atividades de 2021, estas são áreas estratégicas e com as quais ambicionamos explorar novas ligações e oportunidades. Neste âmbito, contribuí a título pessoal com uma palestra *Mundus Mathematica* no ISEL que explorou a interface entre essas áreas e a Estatística,

https://dmtalks.isel.pt/webinars/inteligencia-estatistica/

Recordo ainda que a nossa Sociedade patrocinou já em 2021 o evento *Interfaces Between Statistics, Machine Learning and AI (*https://edin.ac/3eEJT7n) coorganizado pelo *Bayes Centre*, que é o Centro de Inovação em Inteligência Artificial da Universidade de Edimburgo.

Como já devem ter reparado, **a Estatística chegou ao Nobel**! De facto, o Prémio Nobel da Economia de 2021 incide sobre métodos estatísticos para a análise de causas e efeitos.

Vou agora fazer um apanhado das novidades. Muita coisa se passou desde a última Mensagem do Presidente...

§1. Sócios institucionais

Temos neste momento novos **sócios institucionais** que gostaria de vos apresentar e a quem gostaria de dar as boas-vindas! Os sócios institucionais da SPE são neste momento os seguintes:

Os sócios institucionais são fundamentais para que possamos continuar a ampliar as ligações entre a nossa Sociedade e a Sociedade Civil - quer em termos de investigação, quer no estabelecimento de novas alianças estratégicas. Vou partilhar mais acerca deste ponto numa próxima oportunidade.

§2. WikiSPE

Está em marcha o projeto **WikiSPE**, o qual ambiciona elevar o perfil e a visibilidade da "Nossa Estatística" no mundo Wiki. Ainda que o projeto não tenha sido concluído, gostaria de partilhar convosco duas páginas Wikipédia de anteriores Presidentes da SPE que já foram, entretanto, produzidas no âmbito deste projeto:

https://pt.wikipedia.org/wiki/Maria_Eduarda_Silva https://pt.wikipedia.org/wiki/Carlos_Braumann

Qualquer um de vós pode agora contribuir para estas páginas ou até para a redação de versões em inglês das mesmas. Voltarei a este projeto em breve até porque vários colegas estão envolvidos e gostaria assim de lhes agradecer; por agora os meus agradecimentos vão para o Manuel G. Scotto (IST) e para a Raquel Menezes (U Minho) por redigirem as primeiras versões das páginas referidas acima.

§3. **SPE 2021**

Acabou de decorrer o XXV **Congresso da SPE—SPE 2021**. Muito obrigado mais uma vez a toda a comunidade SPE pelo seu envolvimento absoluto. Gostaria de deixar um agradecimento especial à **Comissão Organizadora da Universidade de Évora**, bem como à **Comissão Científica**, por todo o seu empenho e dedicação exemplar; de todos nós, estes colegas merecem um "Muito Obrigado!"

Devido ao contexto pandémico, este foi o primeiro Congresso da SPE a decorrer on-line - um desafio que ambicionámos converter numa oportunidade. Dessa oportunidade resultou que tivemos 200 participantes de várias partes do mundo a debater connosco os últimos desenvolvimentos científicos na nossa área, bem como o amanhã da nossa profissão. E a nossa nação teve particular destaque durante o Congresso, com duas sessões exclusivamente dedicadas às Estatísticas do nosso país (*Statistics of the Nation*, I–II), tendo como chairs Francisco Lima (Presidente do INE) e Lisete Sousa (Coordenadora CEAUL). O artigo da Comissão Organizadora que será apresentado neste Boletim contém mais detalhes sobre o Congresso gostaria apenas de acrescentar o seguinte:

- Tal como foi referido na sessão convidada do Caucus for Women in Statistics (CWS), estamos neste momento a colaborar numa série de iniciativas com o CWS, incluindo uma tentativa conjunta para criar um *Dia Mundial da Mulher na Estatística*.

- Gostaria ainda de congratular calorosamente a Professora Manuela Neves pelo Prémio Carreira SPE 2021, bem como os vencedores do Prémio Estatístico Júnior 2021 e os vencedores do prémio "melhor poster SPE 2021" (João Pereira, Constantino Caetano, Liliana Antunes, Maria Luísa Morgado, Paula Patrício, Baltazar Nunes).

- Contámos com a presença virtual de várias entidades internacionais (ex: ABE, Bernoulli Society, CWS e ISBA) bem como dos sócios institucionais da SPE.

O Congresso SPE 2023, Guimarães, que será organizado pela Universidade do Minho, fica agora no horizonte!

§4. **Outras Atividades**

Dia 20 de outubro celebrámos o **Dia Europeu da Estatística 2021**. As celebrações resultaram de uma parceria com a FENStatS e contaram com duas comunicações que fizeram alusão ao lema do *European Statistical Advisory Committee* ("Estatística, uma vacina para proteger a democracia e combater o vírus da desinformação"). Houve ainda oportunidade para prestar homenagem à Professora Manuela Neves, no âmbito do Prémio Carreira 2021, e para anunciar que está aberto o concurso para o Prémio SPE 2021. Para mais detalhes acerca das celebrações sugiro consultar a página:

https://spestatistica.pt/pt/noticias/noticia/dia-europeu-da-estatistica-o-evento-para-assinalar-a-data

Recomendo ainda a leitura do excelente artigo do Colega Tiago Marques o qual foi publicado no Jornal "O Público" no âmbito das celebrações da SPE:

https://www.publico.pt/2021/10/20/ciencia/opiniao/estatistica-arma-desinformacao-1981731

Por último faço notar que decorreu em Junho a conferência internacional **EVA 2021** a qual foi patrocinada pela SPE e que teve origem há quase 40 anos em Vimeiro, Portugal (1983); trata-se da conferência mais importante em Estatística de Valores Extremos a nível internacional e, portanto, um grande motivo de orgulho para a nossa comunidade, em termos de influência e impacto na comunidade internacional.

§5. **Reflexão**: Culmino com o ponto mais importante desta mensagem:

"O que é verdadeiramente mágico nos pontos §§1–4 é que tudo resulta da colaboração da nossa comunidade, dum esforço conjunto e continuo alinhado por uma mão invisível que leva a nossa voz mais longe e que nos torna mais fortes. Precisamos de continuar a unir esforços, a escrever o presente e a moldar o futuro da nossa profissão. A Estatística precisa de vós—o País e o Mundo precisam de vós."

Até breve,

Edimburgo, 25 de Outubro de 2021

Cordais saudações

Miguel de Carvalho

Notícias

• XXV Congresso SPE

SPE 2021: 13–16 de outubro; Évora, Portugal

Relato sobre o XXV Congresso para o Boletim de Outono SPE 2021

Preâmbulo
Em 2019, a Direção da SPE, então presidida pela colega Maria Eduarda Silva, lança o repto para realizar o XXV Congresso da SPE na cidade histórica de Évora, património da Humanidade. Desde 2004 que este magno evento não voltava à nossa cidade. Tratar-se-ia de um acontecimento especial, pois o principal objetivo seria a celebração do 40º Aniversário da SPE. A Comissão Organizadora Local integrada por Dulce Gomes, Lígia Henriques-Rodrigues, Patrícia Filipe e Russell Alpizar-Jara, com o apoio institucional e local, empenharam-se em brindar a nossa Comunidade com um evento que veio a ser moldado pelas incertezas do acaso, mas que poderemos plasmar num acontecimento histórico relatado e simplificado nos seguintes 4 atos:

Ato I: Amarante 2019 (passado)
É anunciado à Comunidade de estatísticos onde decorrerá o XXV Congresso em 2020 e o objetivo do mesmo, a celebração do 40º Aniversário da SPE. Um agradecimento muito especial à Comissão

Organizadora Local, presidida pelas colegas Maria João Polidoro e Sandra Ramos, por este fantástico Congresso e por todo o apoio que sempre nos brindaram.

Ato II: Évora 2020 (40º Aniversário da SPE)
No Ato I, não se previa no final desse ano o aparecimento da Covid19, que veio a mudar o mundo, e os Congressos da SPE não ficaram indiferentes.
O Congresso iria decorrer no Évora Hotel, já aqui tinha decorrido o XII Congresso em 2004, cuja COL foi presidida pelo nosso querido colega e amigo Carlos Braumann.
A Direção da SPE, junto com a COL, decidiu nesta ocasião que não estavam reunidas as condições para a realização de um Congresso presencial, originalmente agendado para o mês de novembro, na data do aniversário da SPE.

Ato III: Virtualidade 2021 (presente)
A atual Direção da SPE, presidida pelo *Mastermind* do evento virtual, Miguel de Carvalho, imediatamente após a sua tomada de posse no início deste ano, contacta a COL para dar andamento e continuidade ao programa de trabalhos. Apesar de algumas incertezas, mas motivados pelo entusiasmo transmitido pelo Miguel, fomos adaptando-nos e moldando-nos a esta nova realidade virtual, aprendendo cada vez mais sobre as ferramentas e plataformas disponíveis (*Sococo, Slack, Mailchimp, Zoom*, …) que representaram um desafio para todos nesta nova e única experiência.
Ainda assim, um sonho convertido em realidade "virtual", este Congresso congregou mais de 200 participantes provenientes de várias partes do mundo, mais de 140 comunicações incluindo 4 plenárias proferidas por Genevera Allen (Rice University, US), Anthony C. Davison (EPFL, Switzerland), António Pacheco (IST, Portugal), e Maurizio Sanarico (*SDG Group*, Italy).
Foram apresentados 22 posters e as restantes comunicações orais em mais de 30 sessões, algumas delas organizadas por membros coletivos, como o INE (com a participação do seu Presidente, Francisco Lima), o Banco de Portugal; sociedades e associações congéneres, como a CLAD, SGAPEIO, e RBRAS, Secção de Biometria da SPE, *Caucus for Women in Statistics* e FenStatS.
Dentro das contribuições destacaram-se temas de atualidade como Análise Estatística da Covid19, Alterações Climáticas, Estatísticas da nossa Nação, Literacia e Ensino da Estatística, Estimação da Densidade e Abundância de Populações Animais, entre outros. As metodologias estatísticas como Processos Estocásticos, Equações Diferenciais Estocásticas, Séries Temporais, Estatística Bayesiana, Estatística de Extremos, Análise de Dados Longitudinais e Censurados, Métodos Robustos e de Redução da Dimensionalidade, entre outras também tiveram destaque.
Momentos importantes deste Encontro incluíram os anúncios dos premiados: o Prémio Carreira para a nossa querida colega e amiga Manuela Neves, os tradicionais Prémios Estatístico Júnior para os nossos mais novos estatísticos do 3º Ciclo do Ensino Básico e do Ensino Secundário, e o prémio ao melhor póster patrocinado pela Springer (João Pereira e coautores).
Nesta cerimónia dos prémios fomos acompanhados por dois grandes comunicadores e figuras da divulgação matemática em Portugal, Inês Guimarães (*MathGurl*) e Rogério Martins (autor e apresentador do programa de televisão "Isto é matemática"), que partilharam as suas experiências e nos deleitaram com os seus talentos.

Ato IV: Guimarães 2023 (futuro)
Votos de sucesso para os colegas que aceitaram o próximo desafio. Esperamos que nessa oportunidade nos encontremos presencialmente como foi sempre desejado pela nossa família de sócios e simpatizantes da SPE. Estamos disponíveis para ajudar aos nossos congéneres da próxima COL e passar o testemunho de tudo aquilo que seja preciso, como tem sido perpetuado no passado. O contacto humano e as interações sociais entre os participantes nunca poderão ser substituídos por esta realidade virtual. Temos muitas saudades dos encontros presenciais e esperamos que voltem em breve.

O Congresso teve lugar através da plataforma Sococo, tendo como palco principal o auditório virtual que se apresenta a seguir.
Nas diversas salas virtuais, para além dos pósteres e do Boletim SPE, estavam também os parceiros internacionais (*Bernoulli Society*; *ISBA, Internacional Society for Bayesian Analysis*), bem como os Sócios Institucionais da SPE (Banco de Portugal, GADES, INE, IPS, PORDATA, PSE).

21 de outubro de 2021
A Comissão Organizadora Local
XXV Congresso da SPE
http://www.spe2021.uevora.pt/

• Sessão Prémio Carreira da SPE 2021

O Prémio Carreira – SPE foi instituído, em 2013, pela Sociedade Portuguesa de Estatística (SPE) e propõe-se reconhecer a atividade de estatísticos portugueses com papel de relevância no desenvolvimento científico, pedagógico e de divulgação da Estatística em Portugal.


SPE
Prémio de Carreira
Sociedade Portuguesa de Estatística
À Professora Doutora
Manuela Neves
pela obra científica e dedicação
ao desenvolvimento e divulgação
da Estatística em Portugal
2021

Em 2021, o Prémio Carreira foi atribuído à Professora Manuela Neves do Instituto Superior de Agronomia da Universidade de Lisboa. A homenagem e entrega foram feitas numa sessão virtual organizada pela SPE e integrada nas celebrações do Dia Europeu da Estatística.
O Prof. Miguel de Carvalho, Presidente da SPE, fez a "entrega virtual" do diploma comemorativo do Prémio Carreira – SPE. No final desta edição são apresentados testemunhos.

FR

• Dia Europeu da Estatística

No dia 20 de outubro, celebra-se o Dia Europeu da Estatística.
A Assembleia Geral das Nações Unidas, em 2010, decidiu que o dia 20 de outubro seja declarado como o Dia Mundial da Estatística. Celebrado pela 1.ª vez nesse ano, é assinalado de 5 em 5 anos. A comunidade estatística europeia, numa iniciativa do ESAC – Comité Consultivo Europeu da Estatística, naquele mesmo dia, celebra -se o Dia Europeu da Estatística anualmente, nos anos intercalares ao Dia Mundial da Estatística. Tem a apoio do Sistema Estatístico Europeu, do Sistema Europeu de Bancos Centrais e da Federação Europeia das Sociedades Nacionais de Estatística (FENStatS). – a Federação Europeia das Sociedades de Estatística.
A SPE como membro da FENStatS, organizou um evento virtual para assinalar a data.

A primeira parte das celebrações contou com duas comunicações sobre o tema do ESAC para este ano:

"Estatística, uma vacina para proteger a democracia e combater o vírus da desinformação"

Os oradores foram: o Prof. Miguel de Carvalho (Presidente da SPE) e o Dr. Walter Radermarcher (Presidente da FENStatS).

Miguel de Carvalho, "Combating variants of the disinformation virus with a booster of statistical literacy"
Walter Radermarcher, " Public Statistics - infrastructure and facts for the public discourse"

A segunda parte das celebrações incluiu uma cerimónia de homenagem à Professora Manuela Neves- Prémio Carreira SPE 2021, de que damos destaque neste Boletim.

FR

• Prémio IASC-LARS para Estudante de Doutoramento da UÉvora

O estudante do Doutoramento em Matemática da Universidade de Évora, Nelson T. Jamba, ganhou o IASC-LARS (International Association for Statistical Computing - Latin American Regional Section) Best Poster Award pelo poster "Mixed models for individual growth in random environment through Laplace approximation" (da autoria de Nelson T. Jamba, Patrícia A. Filipe, Gonçalo Jacinto e Carlos A. Braumann) que apresentou na reunião científica virtual V Latin American Conference on Statistical Computing (LACSC 2021, 19-21 Abril 2021, http://lacsc2020.itam.mx/).
O prémio foi anunciado oficialmente no WSC-ISI 2021 World Statistics Congress.

CB

• Credencial FENSatS de Estatístico

O Sistema de Acreditação Europeia de Estatísticos (European Statistical Accreditation) foi lançado pela FENStatS a 20 de Outubro de 2020 e está actualmente em funcionamento pleno.

Este tema foi objecto de uma Sessão Temática no XXV Congresso da SPE – conforme damos notícia neste Boletim – e que integrou uma palestra de Magnus Pettersson, do Comité para Acreditação na FENStatS – a Federação Europeia das Sociedades de Estatística.
Para obter a acreditação é necessário:

i) pelo menos um Mestrado em Estatística ou equivalente;
ii) pelo menos 5 anos de experiência profissional;
iii) actualização profissional nesse período de 5 anos;
iv) competências em comunicação;
v) conformidade com standards éticos;
vi) ser sócio de uma das Sociedades de Estatística membro da FENStatS (ser sócio da SPE, por exemplo)

Informação detalhada sobre os requisitos para obter a acreditação e sobre o processo estão disponíveis em **https://www.fenstats.eu/accreditation**.

ES

• '2020 Editors' Citation for Excellence in Refereeing for Water Resources Research

Os Editores do periódico *Water Resources Research*, uma das publicações da AGU (American Geophysical Union) recomendaram que Jessica Silva Lomba recebesse a "*2020 Editors' Citation for Excellence in Refereeing for Water Resources Research*".

Um dos serviços mais importantes realizados para a AGU é a revisão cuidadosa dos artigos submetidos. Devido à natureza do processo de revisão, este serviço também é um dos menos reconhecidos. O objetivo desta citação é expressar publicamente a gratidão da *AGU* àqueles cujas avaliações foram particularmente louváveis. Suas contribuições para o programa de periódicos foram inestimáveis para manter um padrão de alta qualidade.

Em apreciação aos excelentes revisores da AGU de 2020, os editores da AGU reconhecem as contribuições dos revisores, cuja valiosa experiência continua a elevar os altos padrões de seus periódicos.
O anúncio foi feito no Eos-*Science News by AGU*, a 30 de junho de 2021, numa publicação de Matthew Giampoala, Vice President of Publications at AGU and Carol Frost, AGU's Publications Committee Chair, [1].

[1] Giampoala, M., and C. Frost (2021), In appreciation of AGU's outstanding reviewers of 2020, Eos - Transactions American Geophysical Union, 102, https://doi.org/10.1029/2021EO159793 .

IA

• REVSTAT - Statistical Journal

Publicação científica de referência, de acesso aberto com revisão pelos pares, constituída por artigos de elevado interesse científico que contribuem para o desenvolvimento da Ciência Estatística, focada em teorias inovadoras, métodos e aplicações nas diferentes áreas do conhecimento.

Em 2020, a REVSTAT - Statisticsl Jornal lançou o Volume 18 - Números 1 a 5, incluindo um número especial, com os artigos listados abaixo (*https://www.ine.pt/revstat/tables.html*).

Volume 18, Issue 1:

• "Nonparametric regression based on discretely sampled curves" by Liliana Forzani, Ricardo Fraiman and Pamela Llop.

• "A transition model for analysis of zero-inflated longitudinal count data using generalized Poisson regression model" by Taban Baghfalaki and Mojtaba Ganjali.

• "Compound power series distribution with negative multinomial summands: characterisation and risk process" by Pavlina Jordanova, Monika Petkova and Milan Stehlík.

• "Characterization of the maximum probability fixed marginals r x c contingency tables" by Francisco Requena.

• "On the occurrence of boundary solutions in two-way incomplete tables" by Sayan Ghosh and Palaniappan Vellaisamy.

• "Depth-based signed-rank tests for bivariate central symmetry" by Sakineh Dehghan and Mohammad Reza Faridrohani.

• "Prediction intervals for time series and their applications to portfolio selection" by Shih-Feng Huang and Hsiang-Ling Hsu.

Volume 18, Issue 2 (Special Issue on "AISC 2018 - International Conference on Advances in Interdisciplinary Statistics and Combinatorics"):

• "Simultaneous inference of gene isoform expression for RNA sequencing data" by Bo Li.

• "Variance estimation using randomized response technique" by Sat Gupta, Badr Aloraini, Muhammad Nouman Qureshi and Sadia Khalil.

• "On Arnold-Villaseñor conjectures for characterizing exponential distribution based on sample of size three" by George P. Yanev.

• "Testing for trends in excesses over a threshold using the Generalized Pareto Distribution" by Sergio Juárez and Javier Rojo.

• "Robust estimation of reduced rank models to large spatial datasets" by Casey M. Jelsema, Rajib Paul and Joseph W. McKean.

• "Estimation of small area total with randomized data" by Shakeel Ahmed, Javid Shabbir, Sat Gupta and Frank Coolen.

Volume 18, Issue 3:

- "Advances in Wishart-type modelling of channel capacity" by J.T. Ferreira, A. Bekker and M. Arashi.
- "Parametric elliptical regression quantiles" by Daniel Hlubinka and Miroslav Šiman.
- "A couple of non reduced bias generalized means in Extreme Value Theory: an asymptotic comparison" by Helena Penalva, M. Ivette Gomes, Frederico Caeiro and M. Manuela Neves.
- "Minimum distance tests and estimates based on ranks" by Radim Navrátil.
- "Bayesian criteria for non-zero effects detection under skew-normal search model" by Sara Sadeghi and Hooshang Talebi.
- "Efficiency of the principal component Liu-type estimator in logistic regression" by Jibo Wu and Yasin Asar.
- "On kernel hazard rate function estimate for associated and left truncated data" by Zohra Guessoum and Abdelkader Tatachak.
- "Independence characterization for Wishart and Kummer random matrices" by Bartosz Kolodziejek and Agnieszka Piliszek.
- "Bootstrap prediction interval for ARMA models with unknown orders" by Xingyu Lu and Lihong Wang.

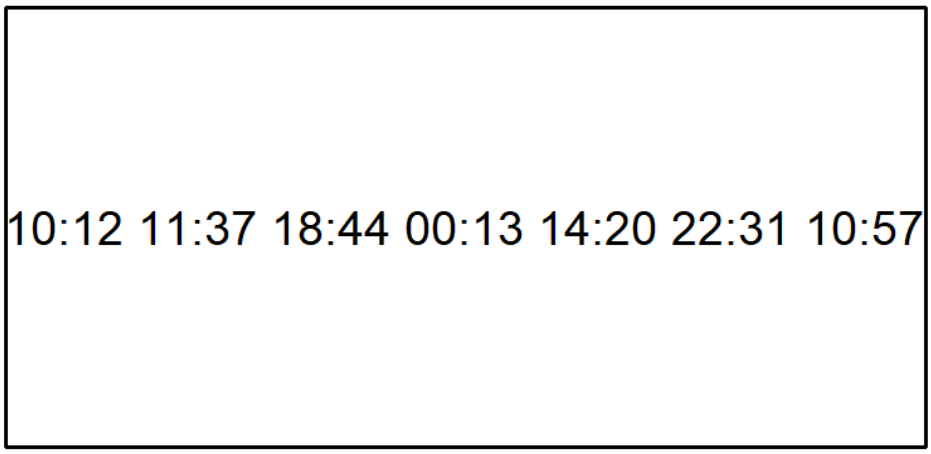
Volume 18, Issue 4:

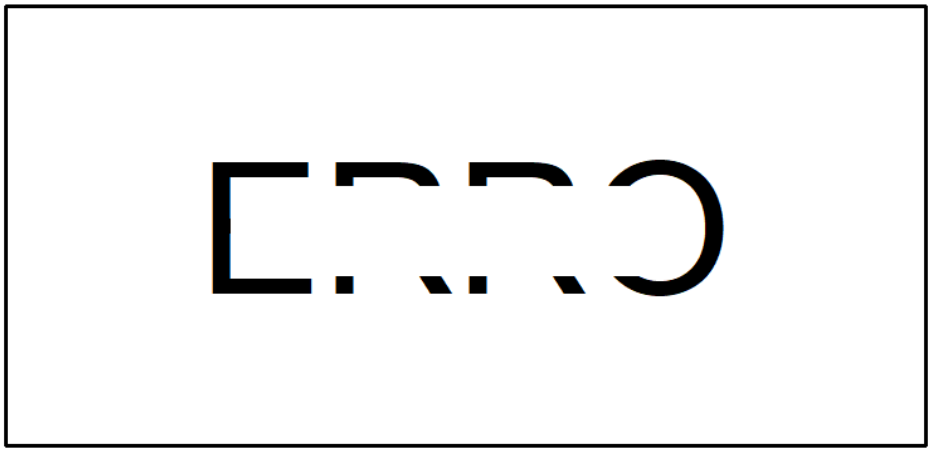
- “Block bootstrap prediction intervals for GARCH processes" by Beste Hamiye Beyaztas and Ufuk Beyaztas.
- “A quantile regression model for bounded responses based on the exponential-geometric distribution" by Pedro Jodrá and María Dolores Jiménez-Gamero.
- “Parameters estimation for constant-stress partially accelerated life tests of generalized half-logistic distribution based on progressive type-II censoring" by Abdullah Almarashi.
- “Averages for multivariate random vectors with random weights: distributional characterization and application" by A.R. Soltani and Rasool Roozegar.
- “Nonparametric CUSUM charts for circular data with applications in health science and astrophysics" by F. Lombard, Douglas M. Hawkins and Cornelis J. Potgieter.
- “Text mining and ruin theory: a case study of research on risk models with dependence" by Renata G. Alcoforado and Alfredo D. Egídio dos Reis.
- “Dissecting the multivariate extremal index and tail dependence" by Helena Ferreira and Marta Ferreira.
- “A new exact confidence interval for the difference of two binomial proportions" by Wojciech Zielinski.
- “On Bayesian analysis of seemingly unrelated regression model with skew distribution error" by Omid Akhgari and Mousa Golalizadeh.

- "Statistics in times of pandemics: the role of statistical and epidemiological methods during the COVID-19 emergency (Invited Paper with Discussion)" by Baltazar Nunes, Constantino Caetano, Liliana Antunes and Carlos Dias.
- "Endemic-epidemic framework used in COVID-19 modelling (Discussion on the paper by Nunes, Caetano, Antunes and Dias)" by M. Bekker-Nielsen Dunbar and L. Held.
- "A brief appraisal of the COVID-19 pandemic in Portugal (Discussion on the paper by Nunes, Caetano, Antunes and Dias)" by M. Gabriela M. Gomes.
- "Rejoinder (Discussion on the paper by Nunes, Caetano, Antunes and Dias)" by Baltazar Nunes, Constantino Caetano, Liliana Antunes and Carlos Dias.
- "Matched pairs with binary outcomes" by Christiana Kartsonaki and D.R. Cox.
- "The xgamma family: censored regression modelling and applications" by Gauss M. Cordeiro, Emrah Altun, Mustafa Ç. Korkmaz, Rodrigo R. Pescim, Ahmed Z. Afify and Haitham M. Yousof.
- "The Fay-Herriot model in small area estimation: EM algorithm and application to official data" by José Luis Ávila-Valdez, Mauricio Huerta, Víctor Leiva, Marco Riquelme and Leonardo Trujillo.
- "A unification of families of Birnbaum-Saunders distributions with applications" by Guillermo Martínez-Flórez, Heleno Bolfarine, Yolanda M. Gómez and Héctor W. Gómez.
- "Modelling irregularly spaced time series under preferential sampling" by Andreia Monteiro, Raquel Menezes and Maria Eduarda Silva.
- "On a sum and difference of two Lindley distributions: theory and applications" by Christophe Chesneau, Lishamol Tomy and Jiju Gillariose.
- "Nonparametric estimation of ROC surfaces under verification bias" by Khanh To Duc, Monica Chiogna and Gianfranco Adimari.

Giovani Silva

Enigmística de mefqa

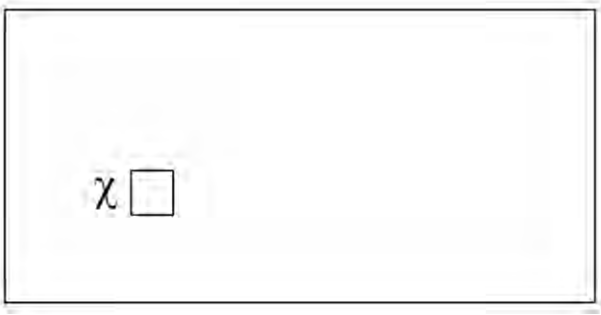




Enigmas 33 e 34

No Boletim SPE primavera de 2021 (p. 12):

Semivariograma



Qui-Quadrado não-central

SPE e a Comunidade

Em Defesa da Língua Portuguesa

Comissão Especializada de Nomenclatura Estatística da SPE

Caros Membros da Sociedade Portuguesa de Estatística:

Inspirados na Lei de Bases da Educação, que no artigo 11º explicita que é objetivo do ensino superior "Promover e valorizar a língua e a cultura portuguesas", propomos à vossa reflexão que desiderato similar seja incluído nos Estatutos da Sociedade Portuguesa de Estatística numa próxima revisão.
Daí decorre, naturalmente, que consideramos que a língua oficial da SPE deve ser a língua portuguesa, e incentivamos os participantes portugueses ou de membros da CPLP em congressos da SPE a valorizar a língua portuguesa nas suas contribuições, usando paralelamente material de apoio que permita a participantes de outras línguas entenderem as suas comunicações.

A Comissão Especializada de Nomenclatura Estatística

Carlos Daniel Paulino Dinis Pestana João António Branco

Nota adicional em 18/10/2021: Esta mensagem foi enviada para o secretariado da SPE, na manhã do dia 13/10/2021, a fim de ser divulgada aos sócios da sociedade. Decorridos que estão 5 dias, os sócios ainda não a receberam!

Comunicado da Direção da Sociedade Portuguesa de Estatística

(divulgado em 18/10/2021)

Caros Sócios,

Em seguimento do comunicado da CENE gostaríamos de clarificar o seguinte.

A preferência pela Língua Inglesa no XXV Congresso da SPE, SPE 2021, resultou do alargado número de estrangeiros que participaram no congresso; não houve, no entanto, objeções a que quem preferisse apresentar em português o fizesse (desde que os slides fossem em Inglês para que todos os participantes pudessem acompanhar a apresentação).

A Língua Portuguesa fez e sempre fará parte de todas as atividades centrais da Sociedade Portuguesa de Estatística.

P'la Direção da Sociedade Portuguesa de Estatística,

Miguel de Carvalho

Competição Europeia de Estatística (ESC 2022)

Instituto Nacional de Estatística

No dia 20 de outubro comemorou-se o Dia Europeu da Estatística. Neste ano, o lema é Estatística, a vacina para proteger a democracia e combater o vírus da desinformação. O dia é habitualmente assinalado com o lançamento de novos produtos, publicações alusivas, organização de eventos e ações dirigidas a públicos diversos.
Por essa razão, desta vez foi lançada a Competição Europeia de Estatística (ESC 2022).

Trata-se de uma competição dirigida às escolas, na qual jovens de 14 a 18 anos, organizados em duas categoriais (ensino básico-3º ciclo e ensino secundário), competem entre si fazendo pesquisa, análise e interpretação de dados estatísticos. O objetivo da competição é promover a literacia estatística entre os alunos e incentivar os professores a utilizarem novos materiais pedagógicos, baseados em estatísticas oficiais.
A ESC tem duas fases: a nacional, durante a qual os alunos competem com colegas do seu país; e a europeia, na qual os vencedores da fase nacional representam os seus países a nível europeu e produzem vídeos com base num tema estatístico.
Na última edição registaram-se mais de 11 000 participantes (dos quais 606 de Portugal) de 16 países. E duas equipas portuguesas obtiveram classificações brilhantes: 1.º lugar na categoria dos alunos do ensino secundárioe2.º lugar na categoria dos alunos do 3.º ciclo do ensino básico.
Para mais informação, aceda à página da Competição Europeia de Estatística: esc2022.ine.pt

Censos 2021 – Resultados Preliminares já saíram!

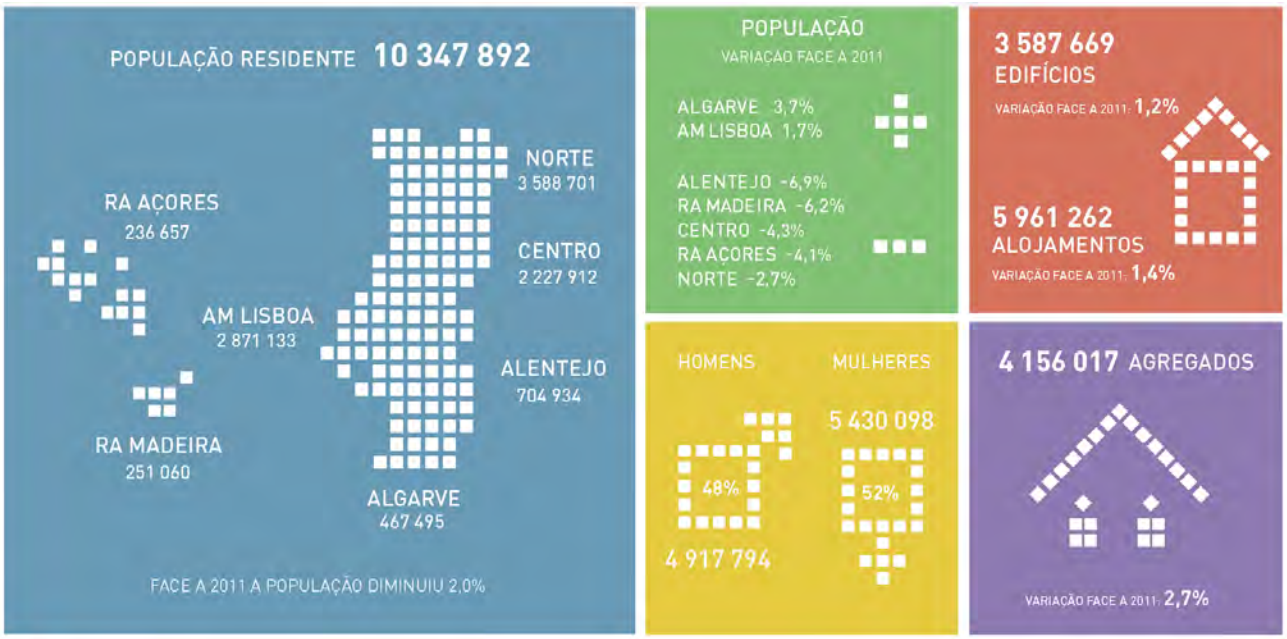
Instituto Nacional de Estatística

No passado dia 28 de julho, o INE divulgou os Resultados Preliminares do XVI Recenseamento Geral da População e VI Recenseamento Geral da Habitação - Censos 2021.

CENSOS 2021

Os Resultados Preliminares dos Censos 2021 revelam que a população residente em Portugal é 10 347 892. Na última década Portugal regista um decréscimo populacional de 2,0% e acentua o padrão de litoralização e concentração da população junto da capital. O Algarve e a Área Metropolitana de Lisboa são as únicas regiões que registam um crescimento da população, sendo o Alentejo aquela que regista o decréscimo mais expressivo. Portugal registou um ligeiro crescimento do número de edifícios e de alojamentos destinados à habitação, embora num ritmo bastante inferior ao verificado em décadas anteriores.

Os primeiros resultados dos Censos 2021 têm um carácter preliminar, na medida em que são baseados em contagens resultantes do processo de recolha (edifícios, alojamentos, agregados e indivíduos) e divulgados antes do processo final de tratamento e validação da informação recolhida, os quais fornecendo facilidade e rapidez no acesso destinam-se essencialmente a antecipar as necessidades dos utilizadores. Os Resultados Preliminares estão disponíveis até ao nível geográfico de freguesia e acessíveis na Plataforma de Divulgação dos Censos 2021 – Resultados Preliminares, disponível em censos.ine.pt.



Machine Learning e Inteligência Artificial

Inteligência Artificial e *Business Analytics*

Paulo Cortez, *pcortez@dsi.uminho.pt*

Centro ALGORITMI / Departamento de Sistemas de Informação, Escola de Engenharia, Universidade do Minho

1. Introdução

A Inteligência Artificial (IA) surgiu na década de 1950, com o trabalho de desenvolvido por diversos investigadores, incluindo Alan Turing e outros, sendo definida por John McCarthy em 1956 como a ``ciência e engenharia para construir máquinas inteligentes, em especial programas de computador inteligentes''[1] (McCarthy, 2007). Desde então, a IA desenvolveu-se em diversas subáreas, tais como (Russell & Norvig, 2020): Representação de Conhecimento, *Machine Learning* (incluindo o *Deep Learning*), Metaheurísticas, *Natural Language Processing* (NLP) ou Sistemas Multi-agente. O impacto atual da IA no mundo resulta de três fenómenos principais (Darwiche, 2018): enorme crescimento de dados (*Big Data*); aumento da capacidade de processamento (segundo a "lei'' de Moore[2]) e desenvolvimento de algoritmos sofisticados para análise de dados (e.g., *Deep Learning*).

Os avanços das Tecnologias de Informação (TI) permitem uma fácil coleta, processamento e partilha de dados sobre pessoas, organizações e do funcionamento do mundo. Assim, assiste-se hoje em dia a uma crescente Transformação Digital. São diversos os fenómenos que geram *Big Data*, tais como as redes sociais, a *Internet-of-Things (IoT)*, as *Smart Cities* e a Indústria 4.0. Estes dados podem conter um potencial conhecimento que é útil para o suporte à decisão. Nas últimas décadas, esta extração de conhecimento a partir de dados foi abordada sob perspetivas distintas envolvendo diversas áreas científicas, como a Estatística e a Matemática, a IA e as TI, dando origem a diversos termos e suas comunidades de investigação, tais como:

- *Machine Learning (ML)*, desde a década de 1960;
- Sistemas de Apoio à Decisão (SAD), desde meados da década de 1960;
- *Business Intelligence*, *Data Mining* e *Analytics*, desde os anos 1990;
- *Big Data* e *Data Science*, desde 2000; e
- *Business Analytics* (BA), desde 2006.

O termo BA é mais recente e está orientado para designar a extração de conhecimento acionável a partir de dados históricos via análises estatísticas, preditivas e de otimização (Arnott & Pervan, 2014). Em termos da IA, as subáreas do ML e das Metaheurísticas são particularmente úteis para o BA, conforme será explicado neste artigo.

Do lado do ML (por vezes também designado de *Data Mining*[3]), existe um número elevado de algoritmos (e.g., Apriori, K-means, Redes Neuronais e de *Deep Learning*, Redes Bayesianas, Árvores de Decisão, XGBoost e outros Ensembles, Máquinas de Vetores de Suporte) que permitem realizar análises multidimensionais de dados com diversos fins (e.g., aprendizagem não supervisionada e supervisionada). Cada algoritmo de ML cria automaticamente um modelo a partir de um conjunto de

[1] Tradução adotada para o excerto original em Inglês.
[2] Ver: https://pt.wikipedia.org/wiki/Lei_de_Moore.
[3] Ver por exemplo a opinião de Domingos (2012).

dados (portanto *data-driven*) e que assume uma dada forma de representar o conhecimento que foi extraído (e.g., Árvore de Decisão).
Quanto às Metaheurísticas, estas consistem em métodos iterativos de procura num espaço de soluções (Cortez, 2021). Alguns exemplos de Metaheurísticas populares são o *Hill Climbing*, o *Simulated Annealing*, a Computação Evolucionária e o *Swarm Intelligence*. Os métodos "clássicos" da Investigação Operacional (e.g., Programação Linear) (Hillier et al., 1990) tendem a usar modelos simplificados e que por isso podem não refletir bem especificidades complexas do mundo real (e.g., descontinuidades, múltiplos objetivos e que por vezes entram em conflito entre si, restrições, elevado espaço de procura, alterações dinâmicas nos dados e de relações entre variáveis). Mesmo não garantindo uma solução ótima, as Metaheurísticas podem ser adaptadas a uma vasta gama de problemas complexos do mundo real, tendendo a encontrar soluções de qualidade com um bom uso de recursos computacionais (Michalewicz et al., 2006).
Em 2013 o conhecido grupo Gartner (na área das TI, ver https://www.gartner.com/en) propôs um modelo de BA que assume quatro tipos principais de análises de dados ou *analytics*:

- descritivas, onde se descreve o que aconteceu;
- diagnósticas, onde se tenta explicar o porquê do que aconteceu;
- preditivas, tentando estimar o que se irá suceder; e
- prescritivas, tentando fazer acontecer.

O modelo do grupo Gartner assume um aumento de dificuldade e valor, à medida que se avança no tipo de análise efetuada, de descritiva até prescritiva. Nas análises descritivas, são relevantes diversos tipos de relatórios e *dashboards*, que podem basear-se em estatísticas ou gráficos simples (e.g., médias, desvio padrão, histogramas) sobre uma variável alvo ou em análises multidimensionais mais complexas (e.g., via algoritmos de ML que envolvam um Agrupamento de Dados ou Regras de Associação). Para as análises diagnósticas, o modelo assume uma análise exploratória de dados, via uma interação com um decisor. Contudo, as análises diagnósticas também podem ser obtidas via métodos mais complexos de inferência causal do domínio da estatística e da IA (e.g., Pearl, 2009). Quanto às análises preditivas, estas podem ser realizadas via modelos matemáticos (e.g., métodos de Alisamento Exponencial para séries temporais) e de ML (e.g., algoritmos supervisionados de Classificação e Regressão). Por último, existem diversas técnicas que podem ser utilizadas para realizar ações prescritivas, isto é, selecionar uma de entre diversas escolhas de decisão, tais como a Simulação e a Otimização. Em particular, na Otimização pretende-se maximizar ou minimizar um ou mais objetivos associados à decisão. Para este efeito, destacam-se aqui o uso de Metaheurísticas por serem mais flexíveis, ou seja, facilmente adaptáveis a uma vasta gama de problemas complexos do mundo real, mesmo contendo um elevado espaço de procura, restrições ou múltiplos objetivos (Cortez, 2021).
De realçar que em 2006 (portanto 7 anos antes da proposta do grupo Gartner), Michalewicz e seus colaboradores propuseram o conceito de *Adaptive Business Intelligence* (ABI) que é apresentado na Figura 1. Tal como o modelo do grupo Gartner, o conceito ABI assume que para suportar efetivamente uma decisão, um sistema inteligente tem de ser capaz de responder a duas questões fundamentais associadas ao contexto da decisão: O que vai ocorrer no futuro? E qual a melhor decisão que pode ser tomada no presente? Para responder a estes dois desafios, um sistema ABI usa dados históricos, recorrendo ao ML para obter previsões (gerando modelos *data-driven*) e a Metaheurísticas para otimizar objetivos associados à decisão (e.g., reduzir custos, maximizar lucros). Ambos os módulos, de previsão e de otimização podem ser ajustados de modo dinâmico, à medida que surgem novos dados ou novas relações entre as variáveis associadas ao contexto da decisão, daí o foco na *adaptabilidade*. De referir ainda que o conceito ABI não explicita como os módulos de previsão e otimização se devem relacionar, embora alguns dos exemplos apresentados em (Michalewicz et al., 2006) assumam a sequência proposta pelo modelo do grupo Gartner: primeiro prever, utilizando depois as previsões, em conjunto com valores de variáveis já conhecidas, na otimização.

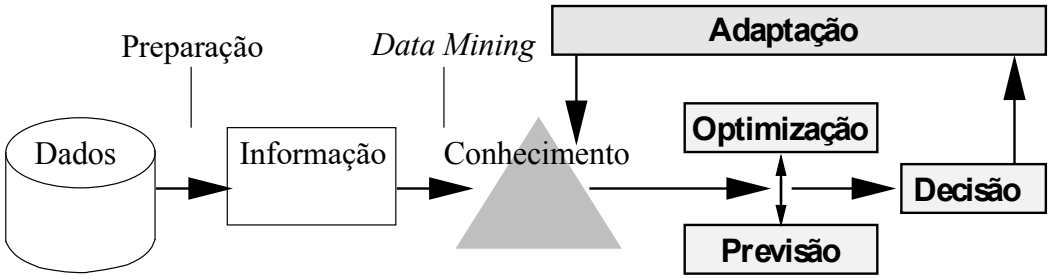


Figura 1 – Conceito ABI proposto por Michalewicz et al. (2006) para extração de conhecimento acionável a partir de dados.

2. Exemplos de Aplicações

Nesta seção, serão apresentadas diversas aplicações de BA envolvendo o autor deste artigo e relativas a análises descritivas, preditivas e prescritivas via métodos de IA. Em particular, o autor tem vindo a trabalhar no modelo designado de *Technology for Adaptive Data Analytics (TADA)* apresentado na Figura 2. Este modelo assume uma junção do conceito ABI com o mapeamento funcional do grupo Gartner, bem como diversas TI adaptativas de IA, nomeadamente ML e Metaheurísticas. Em especial, recorre a análises preditivas (via ML) e prescritivas (via Metaheurísticas) para o suporte à decisão. Quando as decisões se transformam em ações, é expectável que ocorra uma alteração do ambiente associado à decisão, pelo que o respetivo *feedback*, possivelmente associado a alterações em dados internos ou externos, pode ser utilizado para adaptar os modelos preditivos e prescritivos, de modo a suportar novas decisões.

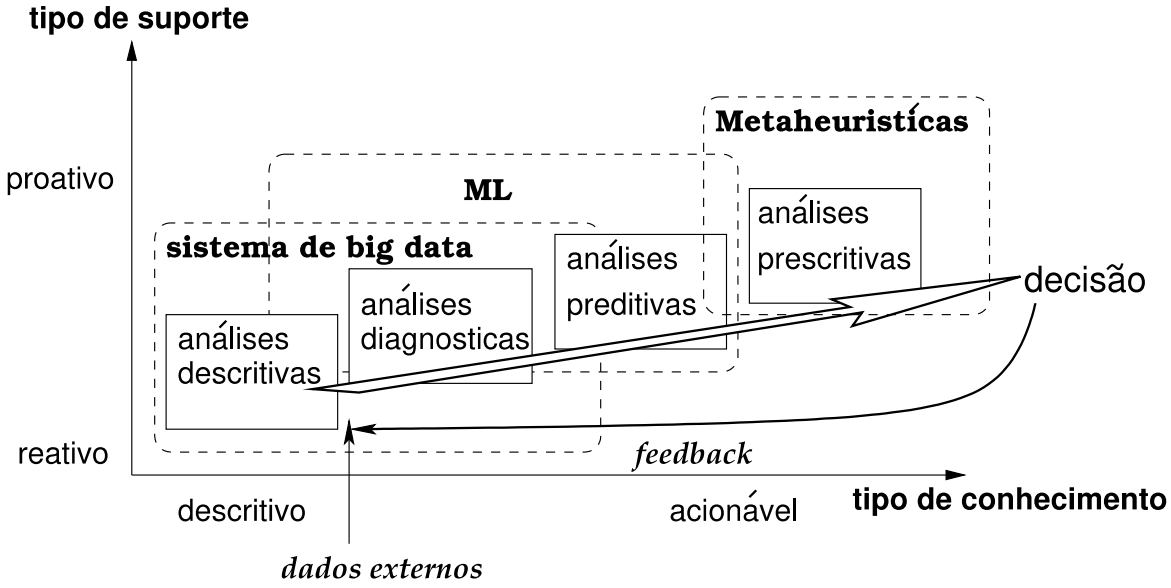


Figura 2 – Modelo TADA para suporte à decisão

Sobre aplicações que envolvem o ML, destaca-se a investigação em modelos preditivos (de Classificação e Regressão), onde é utilizada uma aprendizagem supervisionada com base num conjunto de dados. Tipicamente, dados estruturados ou tabulares (e.g., em formato CSV) são analisados via processos de *Data Mining* (DM), enquanto que dados textuais são processados via um NLP ou *Text Mining*. Alguns exemplos de modelos preditivos estudados pelo autor são: estimar a qualidade final de um vinho com base em análises físico-químicas (regressão, DM via Máquinas de Vetores de Suporte) (Cortez et al., 2009); prever se haverá uma venda durante um contacto telefónico bancário (classificação, DM via Redes Neuronais) (Moro et al., 2014); prever indicadores financeiros com base em opiniões de redes sociais (regressão, DM e NLP via Redes Neuronais, Máquinas de Vetores de Suporte e *Random*

Forest) (Oliveira et al., 2017); prever se haverá uma compra após a visualização de um anúncio num dispositivo móvel (classificação, DM via novo método de *Evolutionary Decision Trees*, uma combinação de uma Metaheurística e ML) (Pereira et al., 2021); e prever necessidades de componentes de montagem, usando *leading indicators*, no contexto da *Bosch Car Multimedia Portugal* (regressão, DM via Redes Neuronais e outros algoritmos ML) (Gonçalves et al., 2021b). Também foram estudados métodos visuais e de análise de sensibilidade para obter-se um *eXplainable Artificial Intelligence (XAI)* de modelos preditivos complexos (e.g., Rede Neuronal) (Cortez and Embrechts, 2013). Num grau menor, foi estudado o ML para extração de conhecimento descritivo (aprendizagem não supervisionada). Dois exemplos são: análise de causas de falhas em telecomunicações via regras de associação (Costa et al., 2009); e caracterização de tipos de produtos em *stock* (e.g., "high runner", "stable") via um agrupamento de dados com recurso ao algoritmo K–means (Gonçalves et al., 2021a).

Quanto às análises prescritivas, o autor também apresenta alguns trabalhos que recorrem a Metaheurísticas. Por exemplo, em (Rocha et al., 2011) otimizou-se a qualidade de serviço em redes de computadores via uma Computação Evolucionária. Mais recentemente, diversas Metaheurísticas, nomeadamente *Hill climbing*, *Simulated Annealing* e Computação Evolucionária, foram adaptadas para realizar uma realocação de *scooters* elétricas na cidade de Barcelona (Fernandes et al., 2020).
Por último, destacam-se três aplicações envolvendo uma interação entre análises preditivas (via ML) e prescritivas (via Metaheurísticas), instanciando assim o conceito ABI e modelo TADA (embora a análise adaptativa ao *feedback* ainda não tenha sido estudada). Em (Parente et al., 2015), realizou-se uma alocação de equipamento em projetos de terraplanagem. Este trabalho assumiu em primeiro lugar uma modelação via redes neuronais para estimar o rendimento de compactadores com base em diversas características de terraplanagem (e.g., tipo de solo a compactar). As estimativas obtidas foram depois incorporadas numa otimização multiobjetiva evolutiva (via algoritmo NSGAII), que minimiza de modo simultâneo (via uma abordagem de curva de Pareto) o custo e tempo do projeto de terraplanagem. Para dados históricos de uma obra real, demonstrou-se que o sistema ABI desenvolvido era capaz de produzir planos de alocações de equipamento que eram mais eficazes quando comparados com a alocação manual que foi realizada. Por sua vez, em (Fernandes et al., 2015) foi concebido um SAD para a conceção de notícias digitais. Primeiro, foi utilizado DM e NLP via um modelo *Random Forest* para prever a popularidade de uma notícia (antes de ser publicada). Este modelo foi posteriormente integrado numa procura de *Hill Climbing* capaz de sugerir alterações que aumentem a popularidade da notícia (e.g., recomendar o uso de um título mais curto ou nova *keyword*). De realçar, que este trabalho recebeu o *best paper award* da conferencia internacional EPIA 2015. Mais recentemente, em (Ribeiro, 2021) abordou-se um SAD para planeamento de subcontratação de operações sobre tecidos têxteis. Via um sistema de *Automated ML (AutoML)*, desenvolveram-se diversos modelos preditivos, incluindo um modelo capaz de estimar o tempo de produção de uma empresa subcontratada. Este modelo foi posteriormente integrado numa otimização multiobjetiva evolutiva (NSGAII) capaz de minimizar, de modo simultâneo (via curva de Parteo), o custo e o tempo total da alocação, de modo a conseguir selecionar um plano para alocar de diversas operações têxteis, que são necessárias para a produção de um artigo têxtil (e.g., corte, costura, empacotamento), a diversas empresas que estão disponíveis para realizar operações de subcontratação.

Referências Bibliográficas

Arnott, D. and Pervan, G. (2014). A critical analysis of decision support systems research revisited: the rise of design science. Journal of Information Technology, 29(4):269–293.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

Cortez, P. and Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225:1–17.

Cortez, P. (2021). Modern Optimization with R. Springer, second edition.

Costa, R., Cachulo, N., and Cortez, P. (2009). An intelligent alarm management system for large-scale telecommunication companies. In Lopes, L. S., Lau, N., Mariano, P., and Rocha, L. M., editors, Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence, EPIA 2009,

Aveiro, Portugal, October 12-15, 2009. Proceedings, volume 5816 of Lecture Notes in Computer Science, pages 386–399. Springer.
Darwiche, A. (2018). Human-level intelligence or animal-like abilities? Communications of the ACM, 61(10):56–67.
Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, *55*(10), 78-87.
Fernandes, G., Oliveira, N., Cortez, P., and Mendes, R. (2020). A realistic scooter rebalancing system via metaheuristics. In Coello, C. A. C., editor, GECCO '20: Genetic and Evolutionary Computation Conference, Companion Volume, Cancu´n, Mexico, July 8-12, 2020, pages 265–266. ACM.
Fernandes, K., Vinagre, P., and Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. In Pereira, F. C., Machado, P., Costa, E., and Cardoso, A., editors, Progress in Artificial Intelligence - 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings, volume 9273 of Lecture Notes in Computer Science, pages 535–546. Springer.
Gonçalves, J. N., Cortez, P., & Carvalho, M. S. (2021a). K-means clustering combined with principal component analysis for material profiling in automotive supply chains. European Journal of Industrial Engineering, *15*(2), 273-294.
Gonçalves, J. N., Cortez, P., Carvalho, M. S., & Frazão, N. M. (2021b). A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain. *Decision Support Systems*, *142*, 113452.
Hillier, F., Lieberman, G., and Hillier, M. (1990). Introduction to operations research, volume 6. McGraw-Hill.
McCarthy, J. (2007) What is Artificial Intelligence? http://www-formal.stanford.edu/jmc/whatisai/
Michalewicz, Z., Schmidt, M., Michalewicz, M., and Chiriac, C. (2006). Adaptive Business Intelligence. Springer.
Oliveira, N., Cortez, P., and Areal, N. (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73:125–144.
Parente, M., Cortez, P., and Correia, A. G. (2015). An evolutionary multi-objective optimization system for earthworks. Expert Systems with Applications, 42(19):6674–6685.
Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys, 3:96–146.
Pereira, P. J., Cortez, P., & Mendes, R. (2021). Multi-objective Grammatical Evolution of Decision Trees for Mobile Marketing user conversion prediction. Expert Systems with Applications, 168, 114287.
Ribeiro, Rui, Pilastri, A., Carvalho, H., Matta, A., Rocha, P., Alves, M. & Cortez, P. (2021). An Intelligent Decision Support System for Production Planning in Garments Industry. In Proceedings of 22nd International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2021), Springer.
Rocha, M., de Sousa, P. N. M., Cortez, P., and Rio, M. (2011). Quality of service constrained routing optimization using evolutionary computation. Applied Soft Computing, 11(1):356– 364.
Russell, S.J. & Norvig, P. (2020). Artificial Intelligence - A Modern Approach, Fourth Edition. Pearson Education.

Será actualmente a Estatística uma mera ferramenta utilitária?

Luísa Canto e Castro Loura, *ldloura@fc.ul.pt*

Docente do DEIO/FCUL e Directora da Pordata/FFMS

No passado dia 13 de maio fiz uma apresentação intitulada "Em busca de tesouros nos oceanos de dados" no ciclo de seminários Mundus Mathematica do ISEL e isso trouxe-me para o tema do desconforto que ainda é a vida em conjunto da Estatística, da Ciência de Dados e da Inteligência Artificial.

Isto acontece porque o mundo digital nos trouxe formas de recolher e armazenar dados em quantidades absolutamente inimagináveis há três dezenas de anos atrás. Ora, das três áreas referidas, a Ciência de Dados é a caloira mas é também a que tem as ferramentas mais potentes para minerar e retirar alguma utilidade de tanto *terabyte* de informação. De facto, a Ciência de Dados inclui tópicos que lhe são exclusivos: gestão de bases de dados, limpeza e mineração de dados, segurança de dados, criação automática de elementos de ligação entre registos de múltiplas bases de dados, ou seja, tudo o que permite traçar rotas e itinerários que nos orientem no vasto oceano de dados em que estamos mergulhados. As ferramentas tecnológicas informáticas são pois, primordiais e, entre estas, aproveito também para destacar as que têm vindo a ser desenvolvidas na área da visualização de dados, com gráficos interactivos e dinâmicos.

Quanto à área da Inteligência Artificial, uma breve nota histórica sobre a reviravolta que houve no início deste século quando passou a incluir de forma intensiva no tema da aprendizagem automática (*machine learning* e *deep learning*) as técnicas estatísticas de análise de dados e de regressão. Na verdade, foi isso que trouxe um novo fôlego à Inteligência Artificial, que estava num impasse quanto ao seu objectivo seminal de criar sistemas computacionais que repliquem a inteligência humana na resolução de problemas complexos.

Ao conseguir traçar perfis de utilizadores, estruturar índices de documentos e antecipar resultados futuros, a IA conseguiu resolver problemas complexos da indústria e da academia e voltou a estar na moda. Ora, com o advento da Ciência de Dados, tudo o que estava baseado em tratamento de grandes volumes de dados na Inteligência Artificial passou a ter "dupla tutela", com a balança a pender cada vez mais para o lado da Ciência de Dados.

Mas então e a Estatística?

Deixou de ser a ciência dos dados, por excelência, que era no século XX?

O seu papel ter-se-á resumido a preparar o caminho e as ferramentas para que esta nova área interdisciplinar da Ciência de Dados prosperasse?

É claro que não: a inferência estatística, os desenhos experimentais, os ensaios clínicos, a amostragem, a modelação, a estatística de extremos e tantos outros tópicos são ainda exclusivos da Estatística – mas isso não evita o sabor amargo de se aparecer como ferramenta utilitária da nova Ciência de Dados.

Olhando segundo outra perspectiva, pode-se dizer que sem Estatística não há Ciência de Dados, enquanto a Estatística, como ciência básica, terá sempre o seu espaço próprio.

No entanto há, a meu ver, duas questões que se colocam: o uso indiscriminado das ferramentas estatísticas no tratamento e análise de grandes bases de dados pode vir a pôr em causa a própria Estatística, pois o uso das ditas ferramentas numa modalidade de caixa negra não nos permite avaliar quão superficiais (ou erradas) podem estar as leituras e interpretações que se fazem dos dados; a lenta adaptação da Estatística ao mundo das amostras de dimensão muito elevada (criação de novas teorias e desenvolvimento de metodologias específicas para modelação e inferência), está a fazer com que perca muita da relevância científica alcançada no século passado.

Estas duas preocupações são o mote para o desenvolvimento que vos trago de seguida.

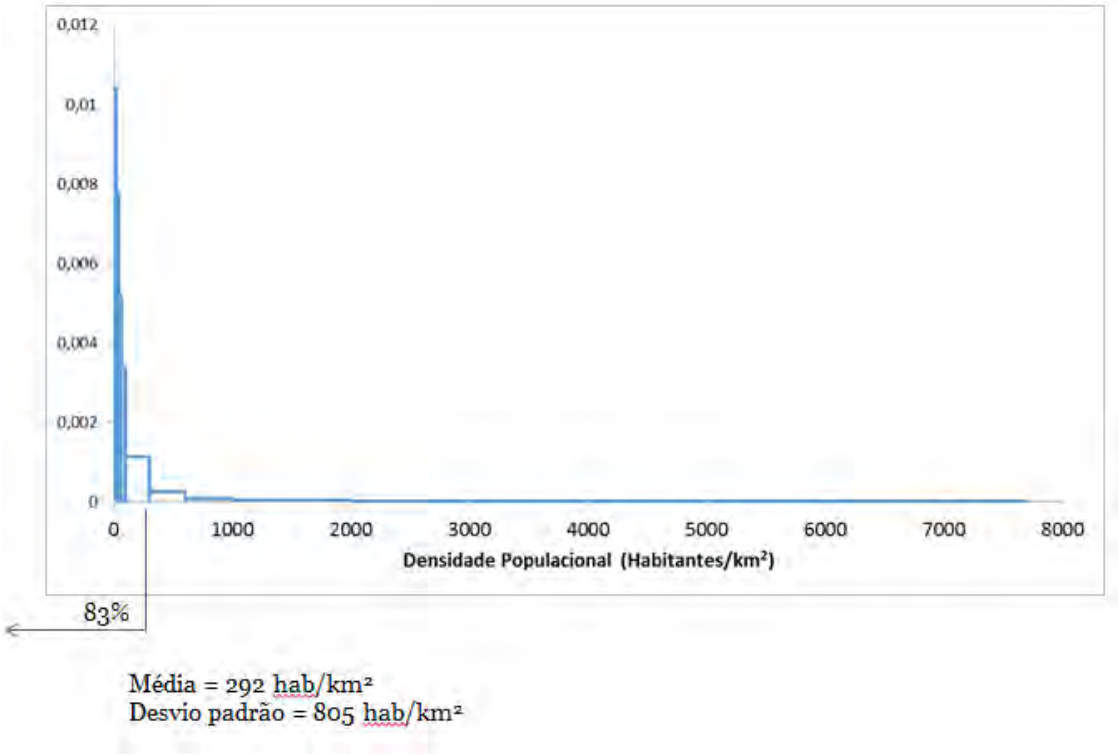
E já que é comum usar os termos oceano e tsunami como analogias para o que se está a passar no mundo digital, no já referido seminário do ciclo de palestras Mathematicae Mundi, recorri a mais uma analogia "marítima", esta a envolver drones e submarinos de investigação científica para mostrar até que ponto o conhecimento da ciência estatística e a inteligência humana são determinantes para que não se estejam a ver somente as pontas dos icebergues e, mesmo essas, quantas vezes distorcidas, e se passem a encontrar os verdadeiros "tesouros" escondidos.

Por circunstâncias diversas, a minha vida profissional mais recente tem-me obrigado a lidar com grandes bases de dados e daí os dois exemplos que, espero, sejam ilustrativos do que acima referi.

Exemplo 1: Dados dos municípios -> redução da dimensão (análise factorial) -> formação de grupos de municípios (*clusters*) usando medidas de proximidade

Os dados deste exemplo foram usados pela aluna Inês Oliveira, do Mestrado em Matemática Aplicada à Gestão da FCUL, na dissertação que concluiu no final de 2020. Ao todo, utilizou 24 variáveis para cada município, entre elas a Densidade Populacional. As restantes estão relacionadas com Envelhecimento, Finanças Autárquicas, População estrangeira, Habitação, Empresas, Trabalhadores, Salários, Nível de escolaridade dos jovens e dos trabalhadores. As respectivas taxas médias de variação anual nos últimos 5 anos também foram incluídas.

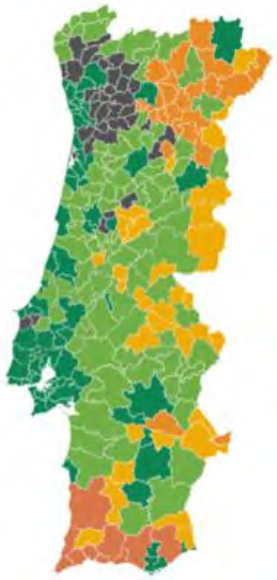
Figura 1 – Densidades populacionais dos municípios portugueses representadas graficamente num histograma com classes de diferente amplitude



Mimetizando a sequência de passos que faria um uso acrítico das ferramentas automáticas de mineração de dados e reconhecimento de padrões, a dimensionalidade reduziu de 24 variáveis para 6 factores e, a partir destes, a organização dos municípios em grupos propôs este mapa: com apenas 2 grupos! Um com os municípios de maior densidade populacional e outro com todos os restantes municípios. Utilidade zero! O facto da variável densidade populacional ter uma distribuição extremamente enviesada fez abafar qualquer possível influência das restantes variáveis em toda a cadeia de cálculos.

Pois é! Analisar a distribuição de cada uma das variáveis e encontrar a boa transformação que corrija a assimetria não faz parte do pacote da Inteligência Artificial em Ciência de Dados e por isso todas as dúvidas e reticências são poucas quando nos mostram padrões e nos apresentam projecções baseadas na análise massiva de dados.

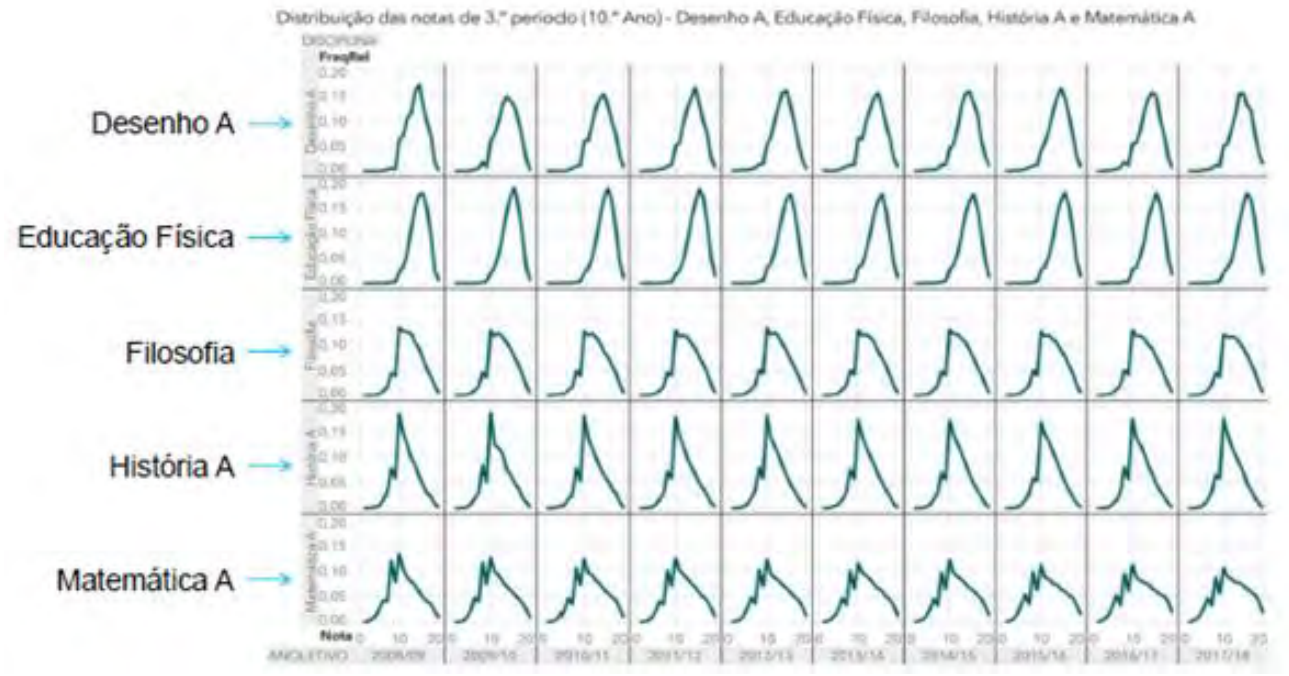
Neste exemplo, uma abordagem mais atenta à existência de enviesamentos das distribuições, permitiu reequilibrar a relevância de cada variável e o mapa de Portugal Continental passou a ter uma leitura útil.
Os factores que tiveram maior influência nesta agregação dos municípios foram os relacionados com a atracção de população estrangeira, com a proporção de população em idade activa e com a existência de recursos humanos qualificados. A escala de cores vai do verde mais forte (valores simultaneamente elevados nestes três factores) aos laranjas e ocres e ao cinzento antracite (municípios de mais baixos níveis de escolaridade). Note-se que os dados são de 2019 e muito poderá ter mudado durante estes quase dois anos de crise pandémica.



Exemplo 2: Notas dos alunos do ensino secundário -> limpeza e estruturação da base de dados -> representação em polígonos de frequência -> distância entre distribuições

Para mim este é bem um exemplo de tesouro que só conseguiria ser descoberto com recurso às tecnologias recentes de estruturação e pesquisa em grandes bases de dados. O que trago aqui é apenas uma pequena parte pois, ao todo, a base de trabalho tinha as notas de 3.º período de todos os alunos do ensino público, em 15 disciplinas dos cursos científico-humanísticos, no 10.º, 11.º e 12.º anos, nos anos lectivos de 2008/09 a 2017/18 (cerca de 15 milhões de registos).
Note-se que, no ensino secundário as classificações são dadas numa escala de 1 a 20 e a classificação que o professor dá no 3.º período deve resumir os resultados das avaliações ao longo do ano. Isto é relevante pois pode ser uma das explicações para a repetitividade dos padrões que iremos ver.

Figura 2 - Distribuição das classificações atribuídas pelos professores no 3.º período do 10.º ano (Cursos Científico-Humanísticos) nas disciplinas de Desenho A, Educação Física, Filosofia, História A e Matemática A – de 2008/09 a 2017/18



Não há quem, num primeiro momento, não fique perplexo perante exemplos tão claros de regularidade estatística. São 10 contingentes distintos de alunos, muitos dos professores também mudaram ao longo destes anos lectivos e, no final, fixada a disciplina, a distribuição das notas pouco muda (pelo menos numa comparação "a olho").
Escusado será dizer que, com amostras de dimensão a rondar os 100 mil, qualquer teste estatístico de homogeneidade das distribuições subjacentes nos leva a valores-p muito próximos de zero e muito inferiores aos níveis de significância usuais. E no entanto, as distribuições das notas a Filosofia, por exemplo, são muito parecidas, as das restantes disciplinas também, e a diferença entre disciplinas é muito clara ao olhar para os gráficos. Mas não há qualquer método da estatística clássica que nos permita fazer afirmações sobre se duas destas distribuições são semelhantes ou não com um nível de confiança pré-fixado e, como é comum dizer-se, estamos perante um desafio!

Para já, num primeiro passo, optou-se por uma alternativa mais heurística, recorrendo à distância de Hellinger (HD) cujo quadrado é dado, neste caso, por:

$$HD^2 = \frac{1}{2}\sum_{i=1}^{20}\left(\sqrt{p_i^A} - \sqrt{p_i^B}\right)^2$$

A distância de Hellinger toma valores entre 0 e 1 e só toma o valor 1 para distribuições cujos suportes são conjuntos disjuntos.
Para a disciplina de Filosofia do 10.º ano, por exemplo, a Tabela 1 mostra que as distâncias são todas muito próximas de zero.
Os anos em que a distância HD foi maior foram 2008/09 e 2017/18 e, mesmo nesse caso não excedeu 0.05 e entre 2015/16 e 2016/17 a distância foi pouco superior a 0.01.
Para se ter uma noção mais concreta de quão próximas de zero são as distâncias que surgem na Tabela 1, basta atentar para a os valores da Tabela 2, onde figuram as distâncias entre distribuições respeitantes a disciplinas diferentes.

Tabela 1 – Distância HD entre as distribuições das classificações internas de 3.º período, no 10.º ano, à disciplina de Filosofia (estão sinalizados os valores máximo e mínimo de HD)

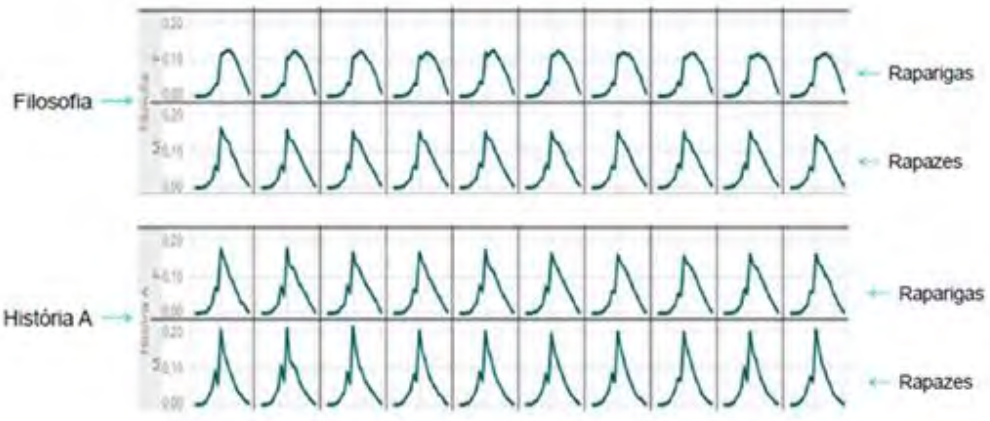
	Filosofia (10.º Ano)									
	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18
2008/09	0.000	0.026	0.031	0.029	0.030	0.031	0.027	0.036	0.036	**0.039**
2009/10		0.000	0.015	0.012	0.027	0.021	0.017	0.020	0.022	0.019
2010/11			0.000	0.012	0.025	0.020	0.016	0.015	0.019	0.019
2011/12				0.000	0.027	0.021	0.017	0.018	0.018	0.016
2012/13					0.000	0.014	0.018	0.022	0.021	0.034
2013/14						0.000	0.012	0.015	0.014	0.026
2014/15							0.000	0.014	0.015	0.022
2015/16								0.000	**0.011**	0.017
2016/17									0.000	0.020
2017/18										0.000

Tabela 2 – Distância HD entre as distribuições das classificações internas de 3.º período, no 10.º ano, às disciplinas de Desenho A e Matemática A e às disciplinas de História A e Educação Física, ao longo dos anos letivos 2008/09 a 2017/18.

	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18
Desenho A e Matemática A	0.403	0.336	0.335	0.325	0.319	0.301	0.323	0.340	0.317	0.315
História A e Educação Física	0.460	0.452	0.451	0.473	0.435	0.424	0.433	0.430	0.449	0.465

Para disciplinas diferentes a distância de Hellinger chega a ser próxima de 0.5.
Para além de muitas outras situações de proximidade e de distanciamento entre distribuições[1] há uma que tem fortes repercussões no acesso ao ensino superior e que, mais uma vez, mostra um padrão sistemático: as distribuições são visualmente distintas quando se desagregam as amostras por sexo feminino (F) e masculino (M) mas são muito semelhantes dentro de cada uma das duas categorias desta variável (Figura 3 e Tabela 3)

Figura 3 - Distribuição das classificações atribuídas pelos professores no 3.º período do 10.º ano (Cursos Científico-Humanísticos) nas disciplinas de Filosofia e História A – de 2008/09 a 2017/18 (por sexo)



[1] Ver estudo completo em:
https://www.dgeec.mec.pt/np4/292/%7B$clientServletPath%7D/?newsId=516&fileName=Avalia__o_de_conhecimentos_no_ensino_sec.pdf

Tabela 3 – Mínimo, máximo e média das distâncias HD entre as distribuições das classificações internas de 3.º período, no **10.º ano**, a Filosofia e História A, por género (F/F e M/M) e entre géneros (M/F).

10.ºANO	Mínimo, Máximo e Média dos valores de HD – 2008/09 a 2017/18								
	F/F			M/M			M/F		
Disciplina	**Min**	**Max**	**Média**	**Min**	**Max**	**Média**	**Min**	**Max**	**Média**
Filosofia	0.013	0.042	0.025	0.010	0.046	0.024	0.093	0.151	0.127
História A	0.018	0.064	0.038	0.026	0.069	0.045	0.074	0.152	0.108

Este exemplo levanta muitas questões sociológicas mas o estudo destes dados usando as habituais metodologias da estatística leva-nos sistematicamente à rejeição da hipótese de igualdade das distribuições e impede a sua divulgação em revistas da área da estatística.
Idealmente, o que se pretende, é afirmar com um grau de confiança pré-fixado que as distribuições são semelhantes quanto baste.
Na área das amostras "gigantes" há, pois, muito espaço para investigação estatística em novos conceitos, novas teorias e novas metodologias.

Inteligência Artificial e Machine Learning

Ricardo Galante, *Ricardo.Galante@sas.com*

Professor Convidado no Departamento de Estatística e Investigação Operacional, FCUL
e
Senior Analytics Customer Advisor no SAS Institute Portugal

Eu sou Professor de Data Science já há alguns anos, e não é incomum, entre alunos e colegas termos conversas e debates sobre o tema Inteligência Artificial e Machine Learning. Me lembro de uma aula onde os alunos colocaram questões tão interessantes que as respostas direcionaram aquela aula e as aulas seguintes. Algumas das questões foram:

- "Quem foi que determinou que Inteligência Artificial teria esse nome e porque?",
- "Inteligência Artificial e Machine Learning são a mesma coisa?",
- "Como as empresas, como Netflix, por exemplo, conseguem recomendar filmes utilizando Machine Learning"

e, talvez a pergunta mais complexa veio na sequência:

- "Como controlar o "poder" da Inteligência Artificial e como é definido o processo ético do uso desta abordagem no nosso dia-a-dia?".

Percebam que estes alunos, fizeram perguntas que variaram desde pontos históricos e conceitos fundamentais passando pelo uso de Inteligência Artificial no dia a dia e, chegaram a questões bastante complexas relacionadas ao bom ou mal uso destes "procedimentos".

Para escrever este artigo, eu me inspirei nestas interações que tive (e tenho) com meus alunos. De uma maneira geral, o meu objetivo é replicar as reflexões que fizemos, naquela prazerosa aula, onde percorremos, de uma forma geral, os principais tópicos relacionados ao nosso tema central.

O que é Inteligência Artificial e Machine Learning?

Vamos começar pelos conceitos fundamentais, que na verdade, deram início a toda reflexão feita durante as aulas.

Se formos ao dicionário online *dictionary.com* teremos as seguintes definições sobre ***Inteligência Artificial***:

a) a capacidade de um computador, robô ou outro dispositivo mecânico programado de realizar operações e tarefas análogas ao aprendizado e tomada de decisão em humanos, como reconhecimento de fala ou resposta a perguntas;
b) um computador, robô ou outro dispositivo mecânico programado com esta capacidade humana;
c) o ramo da ciência da computação envolvido com o projeto de computadores ou outros dispositivos mecânicos programados com a capacidade de imitar a inteligência e o pensamento humanos;

Utilizando estes conceitos como referência, podemos dizer que ***Inteligência Artificial*** (***AI***) é uma tentativa de tornar os computadores tão inteligentes, ou até mais inteligentes que os seres humanos. Trata-se de dar aos computadores comportamentos semelhantes aos humanos, processos de pensamento e habilidades de raciocínio.

E na verdade, AI está inserido nos nossos dias sem mesmo que apercebamos. Por exemplo quando utilizarmos o assistente virtual nos nossos telemóveis para verificar se irá chover no final do dia, ou quando fazemos uma simulação de crédito em um banco, quando jogamos vídeo games como *Far cry* ou *Call of Duty,* e em sistemas de segurança que detetam se os colaboradores estão adequadamente vestidos ou equipados para trabalhar utilizando *computer vision*. As aplicações no mundo corporativo são inúmeras, por exemplo, compras online, "*web search*", "smart homes, cities e infrastructure", "Cybersecurity" e *combate a "fake news*".

O termo Inteligência Artificial foi estabelecido em 1955 em uma proposta para um "*estudo de inteligência artificial de 2 meses e 10 homens*" submetido por *John McCarthy* (Dartmouth College), *Marvin Minsky* (Harvard University), *Nathaniel Rochester* (IBM) e *Claude Shannon* (Laboratórios Bell Telephone). O workshop, que ocorreu um ano depois, em julho e agosto de 1956, é usualmente considerado o momento de nascimento oficial deste termo.

Apesar do termo ter sido estabelecido em ***1955***, a origem deste contexto é muito anterior. Para ser ter uma ideia, em ***1308*** o poeta e teólogo catalão *Ramon Llull* publica *Ars generalis ultima* (The Ultimate General Art), onde são falados de métodos de uso de meios mecânicos baseados em papel para criar novos conhecimentos a partir de combinações de conceitos. Outra referência histórica data de ***1666***. Neste ano, o matemático e filósofo *Gottfried Leibniz* publica *Dissertatio de arte combinatoria* (Sobre a arte combinatória), seguindo o trabalho proposto por *Ramon Llull* ao propor um alfabeto do pensamento humano e argumentar que todas as ideias nada mais são do que combinações de um número relativamente pequeno de conceitos simples. Estes são exemplos históricos que representam o "embrião" acadêmico do que hoje chamamos de ***Inteligência Artificial***.

Qual é a relação entre Inteligência Artificial e Machine Learning?

Ainda há muita confusão entre as pessoas e na mídia sobre o que é genuinamente *Inteligência Artificial* e o que exatamente é *Machine Learning*. Frequentemente, os termos são usados como sinônimos.

Citando reitor interino da **School of Computer Science** da *CMU*, Professor e Ex-Presidente do Departamento de Machine Learning da *Carnegie Mellon University*, Tom M. Mitchell:

"Um campo científico é mais bem definido pela questão central que estuda. A área de ***Machine Learning*** (ML) busca responder à pergunta: Como podemos construir sistemas de computadores que melhoram automaticamente com a experiência, e o que são as leis fundamentais que regem todos os processos de aprendizagem?"

Podemos afirmar que ***AI*** é uma tecnologia com a qual podemos criar sistemas inteligentes que podem simular a inteligência humana. Já ***Machine Learning*** ou ***ML*** é um subcampo da ***AI***, que tem seu foco no aprendizado de máquinas de tal forma que as máquinas aprendam com dados ou experiências anteriores sem serem explicitamente programadas.

Aplicações de Inteligência Artificial e Machine Learning

Do ponto de vista prático, voltando as questões que me foram feitas durante a aula, como citei no início do artigo, como as empresas, como a ***Netflix*** utilizam ***AI*** e ***ML*** nos dias de hoje?

É interessante que, apesar de ***AI*** e ***ML*** remeterem a contextos extremamente atuais e modernos, muitos dos algoritmos que são utilizados hoje, tem a sua gênese há muitos anos no passado.

O sistema de recomendação, por exemplo, é muito utilizado por gigantes corporativos mas a sua origem não é de agora.

Relembro de um excelente livro que lí, "***AIQ – How artificial inteligence works and how we can harness it's power for a better world***", escrito por ***Nick Polson*** e ***James Scott***, onde, de forma simples e didática, os autores exemplificam como muitos dos algoritmos utilizados hoje, na verdade já haviam

sido muito explorados no passado. E um dos meus capítulos preferidos, é o primeiro chamado *The Refugee*. Este capítulo aborda de uma forma muito interessante aplicações de sistemas de recomendação no passado. Além disso, é contado a história do pesquisador e estatístico húngaro, ***Abraham Wald***.
Durante a Segunda Guerra Mundial, ***Wald*** foi membro do **Grupo de Pesquisa Estatística** (SRG) da Universidade de Columbia, onde aplicou suas habilidades estatísticas a vários problemas relacionados a guerra. Um de seus trabalhos estatísticos mais conhecidos durante a 2ª Guerra Mundial foi como minimizar os danos em aviões bombardeiros levando em consideração o viés de sobrevivência em seus cálculos.
Na verdade, podemos seguramente dizer que este trabalho, na verdade é um sistema muito interessante de recomendações.
Este sistema, era totalmente originado através de probabilidades condicionais, ou seja, a probabilidade de algum evento acontecer, sendo que algum outro evento tenha acontecido antes.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(A)P(B|A)$$

A fórmula acima representa que a probabilidade de evento ***B*** acontecer, considerando que o evento ***A*** já aconteceu, é igual à probabilidade de interseção entre o evento ***A*** e o evento ***B***, dividido pela probabilidade de evento ***A***.
Vamos considerar o problema de ***Wald***.
O objetivo era, estimar a probabilidade, do avião retornar a base considerando que ele tenha sido atingido em um ponto específico. Ou seja, o objetivo era calcular a probabilidade do avião voltar à base (***evento B***) sendo que, este avião foi atingido na fuselagem (***evento A***). A probabilidade conjunta ***A*** e ***B*** representa a probabilidade de um avião retornar à base (***B***) com relatórios de danos na fuselagem (***A***). No fundo, o que ***Wald*** propunha era "personalização" de reparos para cada tipo de avião utilizando **probabilidade condicional**.
O mais interessante é que os ***princípios*** utilizados por ***Wald***, são os mesmos aplicados hoje na ***Netflix*** para recomendar filmes aos seus clientes.
A ***Netflix*** é uma empresa que prosperou muito e, em parte, isso é devido ao uso massivo de ***Inteligência Artificial*** e ***Machine Learning***.
Por exemplo, suponha que uma analista da ***Netflix*** queira calcular a probabilidade de uma pessoa assistir ao filme: "***Life is Beautiful***" (*chamamos isso de evento E*) dado que ela já tenha assistido a outro filme: "***Amelie***" (*chamamos isso de evento F*).
Ao utilizarmos probabilidade condicional temos:

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{\frac{\#\text{people who watched both}}{\#\text{people on Netflix}}}{\frac{\#\text{people who watched } F}{\#\text{people on Netflix}}} = \frac{\#\text{people who watched both}}{\#\text{people who watched } F}$$

Ok, mas como isso é utilizado no dia a dia?
Ora, se o analista tem a informação que um cliente assistiu ao filme ***"Amelie"*** e, ao calcular a probabilidade acima, ele obtenha uma valor considerável, a empresa pode "***personalizar***" a sugestão e indicar ao cliente o filme "***Life is Beautiful***". Percebam que as ideias acima são muito similares as que ***Wald*** e sua equipe, utilizaram em ***1944*** e estão relacionadas a ***recomendações assertivas***.
De uma forma geral, as recomendações "***personalizadas***" feitas pelas empresas, resultam em uma melhoria na experiência e satisfação do cliente com a empresa, geram um aumento da receita, potencializam a "***retenção***" dos clientes e, por consequência, proporcionam um aumento na fidelização.

Ética na Inteligência Artificial - Ethical AI

Um outro ponto muito recorrente é relacionada a **ética** no uso de ***Inteligência Artificial*** e ***Machine Learning***.
À medida que mais organizações implementam ***AI*** e ***ML***, em seus processos, os líderes empresariais estão examinando mais de perto o viés destas implementações e as considerações éticas envolvidas. A isto se chama "***Ethical AI***".
Por definição, "***Ethical AI***" considera o impacto total do uso da ***AI*** e ***ML*** em todas as partes interessadas, desde clientes e fornecedores até colaboradores. Também considera o impacto do uso da tecnologia não apenas nas pessoas que a utilizam mas no mundo que as cerca, garantindo que seu uso seja justo e responsável.

Questões como *decisões automáticas*, *desemprego devido à automação*, e *práticas de vigilância, com **AI** e **ML**, que limitam a privacidade* são exemplos de pontos polêmicos de debates relacionados a este tópico.
Esse assunto é extremamente importante. Na verdade, embora ***AI*** e ***ML*** esteja mudando a forma como as empresas funcionam, há preocupações sobre como ela pode influenciar nossas vidas. Esta não é apenas uma preocupação acadêmica ou social, mas um risco para a reputação das empresas.
Nenhuma empresa quer ser exibida na mídia com exposições polêmicas relacionadas a "***Ethical AI***". Temos um exemplo "recente" disto ocorrido na Amazon.
O que ocorreu foi que houve uma reação significativa devido ao uso de um *software* lançado pela Amazon chamado ***Amazon Rekognition***. Apenas para contextualizar, ***Amazon Rekognition*** é um *software* para *computer vision* baseado em "*cloud*" que foi lançado em ***2016***. Ele foi vendido e usado por várias agências governamentais dos Estados Unidos, incluindo ***US Immigration*** bem como entidades privadas para fazer *reconhecimento facial*. O uso deste *software* sofreu alguns impactos negativos.
Isso foi seguido pela decisão da ***Amazon*** de interromper o fornecimento dessa tecnologia para as autoridades policiais e, em junho de 2020, a companhia anunciou que estava aplicando uma suspensão de um ano para o uso dessas ferramentas pela polícia, alegando que a pausa poderia dar ao Congresso dos Estados Unidos tempo para decretar garantias contra o uso indevido do reconhecimento facial.
Uma outra abordagem do "***Ethical AI***", são os receios associados a como o uso de ***AI*** e ***ML*** afetará o mercado de trabalho. Vários estudos mostram que atividades profissionais desaparecerão, sendo substituídas por atividades que até o momento são desconhecidas ou inimagináveis. Com frequência, são divulgadas listas com as profissões com maior probabilidade de desaparecer no futuro.
Mas se faz necessário ressaltar que o uso de "***Ethical AI***" é muito importante no dia a dia das empresas até para evitar a disseminação de mal conceitos relacionados, por exemplo, ao gênero, raça ou opção sexual.
É necessário a promoção do debate relacionado as implicações éticas, regulatórias e políticas que surgem do desenvolvimento tecnológico. Na minha opinião, ele deve se concentrar em como as técnicas, ferramentas e tecnologias de ***IA*** e ***ML*** estão se desenvolvendo, incluindo a consideração de onde esses desenvolvimentos podem levar no futuro.

Conclusão

Inteligência Artificial e *Machine Learning* há muito deixaram de ser ***buzzwords*** e o seu uso foi incorporado no dia a dia de todos.
Eu acredito que existam algumas *resistências* em relação ao uso de *AI* e *ML*, mas em meio ao medo de que sejam uma ameaça agora ou no futuro, é evidente que *AI* e *ML* trazem benefícios substanciais e críticos para os humanos.

Usar os sistemas que imitam a inteligência humana e animal é a próxima fronteira na solução de problemas na sociedade. Nesse sentido, já vemos a sua aplicação em campos como a medicina para buscar a cura de doenças complexas e crônicas ou, nas redes sociais, para identificar de forma prévia a ação de pedófilos ou, identificar comportamentos fraudulentos muito antes que as fraudes venha a acontecer.

Além disso, à medida que o homem aumenta as suas atividades na superfície da Terra, com o suporte de ***AI*** e ***ML***, a natureza pode se beneficiar e estar mais resiliente aos desastres naturais. Nesse caso, ***AI*** e ***ML*** são extremamente úteis e benéficos como parceiros para ajudar os seres humanos a melhorar a sua qualidade de vida, a prevenir situações não desejadas, a mitigar problemas irreparáveis e proporcionar um mundo sustentável para as gerações futuras.

Avaliação de Modelos de Regressão para a Previsão de Valores Extremos

Rita P. Ribeiro, *rpribeiro@fc.up.pt*
Nuno Moniz, *nmmoniz@inesctec.pt*

Faculdade de Ciências da Universidade do Porto, *INESC TEC*

1 Introdução

Nas tarefas de aprendizagem supervisionada é comum assumir que a correta previsão de qualquer valor do domínio da variável a prever (variável objetivo) é igualmente importante. No entanto, nem sempre isso é verdade. Em domínios como finanças, meteorologia ou ciências ambientais, o objetivo é muitas vezes a previsão de eventos extremos ou raros. Podemos imaginar cenários de deteção de fraude, diagnóstico de doenças raras, previsão de valores extremos de temperatura, poluição do ar ou mesmo de consumo de energia. As tarefas de aprendizagem supervisionada "desbalanceadas" formalizam este tipo de cenários de modelação preditiva. Estas possuem duas características principais [2]: *i)* distribuição desigual da variável objetivo e *ii)* maior importância em prever corretamente casos sub-representados.

A investigação na área das tarefas de aprendizadem desbalanceada conta com quase três décadas. Os principais tópicos dos estudos realizados nesta área (p.e. [4, 2, 9, 8, 5]) estão relacionados com a dificuldade dos algoritmos de aprendizagem *standard* resolverem este tipo de tarefas e com a proposta de estratégias para ultrapassar as limitações desses algoritmos através de técnicas de pré-processamento do conjunto de dados, pós-processamento das previsões obtidas pelos modelos, ou mesmo técnicas de modelação desenhadas especificamente para este tipo de problemas. Paralelamente, é igualmente reconhecida a inadequação das métricas de avaliação *standard* e a necessidade do desenho e utilização de métricas de avaliação que sejam mais apropriadas nestes domínios.

Mas, se por um lado, muitos estudos têm sido feitos nesta área, por outro lado, podemos afirmar que, na sua grande maioria, eles são focados em tarefas de classificação desbalanceada e, em particular, na classificação binária. Neste tipo de problemas a classe mais relevante em termos de previsões corretas é a classe minoritária. Em comparação, o número de estudos sobre tarefas de regressão desbalanceada é residual. Do nosso ponto de vista, esta disparidade prende-se com dois desafios importantes que as tarefas de regressão desbalanceada apresentam face às tarefas de classificação. O primeiro desafio tem a ver com a descrição de preferências não uniformes em domínios contínuos. Embora trivial para tarefas de classificação (por exemplo, decidir a classe positiva), as soluções *ad-hoc* para tarefas de regressão exigiriam uma especificação de preferências sobre um domínio potencialmente infinito. São assim necessários métodos automáticos para especificar essas preferências. No entanto, e ainda que a suposição de uma distribuição estática (p.e. distribuição normal) seja comum em alguns trabalhos de regressão, não será adequado fazê-lo neste contexto. O segundo desafio tem a ver com encontrar critérios de avaliação e otimização adequados, capazes de melhorar a capacidade preditiva dos modelos na previsão de valores extremos sem, contudo, um grande enviesamento do modelo. Como nas tarefas de classificação, as métricas padrão utilizadas em tarefas de regressão não são apropriadas - o seu foco é no comportamento dos modelos nos casos mais frequentes. Além disso, em relação às métricas alternativas, estas consideram todos os valores alvo como igualmente importantes, i.e. preferências uniformes, ou permitem um viés extraordinário do modelo para previsão de valores extremos, resultando numa fraca capacidade de generalização.

O objetivo principal deste artigo é fornecer uma nova base para a formalização de tarefas de regressão desbalanceada e métricas para avaliação e otimização de modelos neste contexto. A nossa primeira

contribuição consiste na proposta de um método, baseado em trabalhos anteriores [11, 15, 14] sobre regressão baseada na utilidade, para obter a chamada função de relevância para aproximar a preferência do domínio focada em valores extremos, usando uma abordagem que é automática e não paramétrica. A nossa segunda contribuição é a proposta de uma nova métrica de avaliação - $SERA$ - que permite avaliar os modelos quanto à sua capacidade de previsão de valores extremos, mas robusta a um viés severo do modelo. Esta medida não só é capaz de considerar preferências não uniformes do domínio, como fornece também uma ferramenta para análise de dominância entre modelos. Todas as contribuições estão disponíveis no pacote de código aberto `IRon` [1] desenvolvido na linguagem R [10].

2 Previsão de Valores Extremos

Nas tarefas de regressão *standard*, as estimativas fornecidas pelo erro absoluto médio ou pelo erro quadrado médio, são usadas como critérios de preferência na construção dos modelos de previsão. Estes critérios têm como alvo a minimização do erro médio em todo o domínio da variável objetivo. Dada a predominância de casos com valores próximos da tendência central das distribuições, a redução do erro nas previsões dos casos com valores comuns terá um impacto considerável no desempenho do modelo. Por essa razão, o desempenho do modelo está relacionado principalmente com a previsão precisa de tais valores em detrimento de valores extremos.

No entanto, como mencionado anteriormente, para muitos domínios do mundo real, nem todos os valores da variável objetivo são igualmente importantes. Em muitos contextos, é pertinente que a previsão de valores extremos seja, particularmente, precisa. Por exemplo, numa altura de perigo de fogos florestais uma previsão de $30°C$ para um valor real de $20°C$ ou $40°C$ incorre num erro de igual magnitude. No entanto, este último apresenta uma situação bem mais perigosa.

Neste contexto, consideramos tarefas de regressão desbalanceadas quando, dada uma distribuição da variável objetivo contínua, *i)* tal distribuição mostra a presença de valores extremos (*outliers*), *ii)* as preferências de domínio não são uniformes, e *iii)* o foco preditivo está nos valores extremos.

Debatemo-nos agora com a problemática da especificação de preferências não uniformes no domínio contínuo da variável objetivo. Tendo por base trabalhos anteriores, recorremos ao conceito de relevância e apresentamos um novo método automático e não paramétrico para a obtenção de uma função de relevância que responda à especificidade da importância dos valores extremos. Ilustramos o método num caso de aplicação cujo objetivo é a previsão de valores de concentração muito elevados de dióxido de nitrogénio $NO2$, um poluente que afeta a qualidade do ar com consequências nocivas para a saúde.

2.1 Função de Relevância

O conceito de relevância, introduzido por Torgo e Ribeiro [14], é expresso por uma função que mapeia o domínio da variável objetivo para uma escala de importância em relação às previsões do modelo. Esta função de relevância $\phi(Y) : \mathcal{Y} \rightarrow [0, 1]$ é uma função contínua que expressa o viés específico do domínio de aplicação em relação ao domínio $\mathcal{Y}$ da variável objetivo Y mapeando-o numa escala $[0, 1]$ de relevância, onde 0 e 1 representam a relevância mínima e máxima, respetivamente.

Nas tarefas de classificação desbalanceada, especificar a relevância da variável objetivo é, na maioria dos casos, escolher a classe positiva (minoritária). Nas tarefas de regressão desbalanceada, dada a natureza potencialmente infinita do domínio da variável objetivo, especificar a relevância de todos os valores é virtualmente impossível. A solução passa pela definição de uma aproximação à função de relevância através de um conjunto de valores da variável objetivo para os quais a relevância seja conhecida e uma decisão sobre qual método de interpolação usar.

O trabalho inicial de Ribeiro [11], propõe o uso de *Piecewise Interpolation Cubic Hermite Polynomials* (*pchip*) para a obtenção da função de relevância a partir de um conjunto valores de relevância conhecidos, ou seja, pontos de controlo. Esta opção teve por base a capacidade deste método de interpolação garantir a suavidade da interpolação e permitir um maior controlo sobre a forma da curva da função ger-

[1] O pacote `IRon`, desenvolvido em R, está disponível em `https://github.com/nunompmoniz/IRon`

ada ao requerer a derivada de cada ponto a interpolar. Mais detalhes sobre este método de interpolação podem ser consultados em [11, 12].

Contudo, é importante referir que, na maioria das situações, não dispomos de conhecimento suficiente do domínio para definir estes pontos de controlo de forma precisa. Nestes casos, torna-se necessária uma abordagem automática que determine quais os valores da variável objetivo com relevância mínima e máxima, de acordo com a distribuição da variável. No contexto das tarefas de regressão desbalanceadas, sendo os valores mais extremos da distribuição considerados os mais importantes a prever com precisão, devem assumir relevância máxima. Ao contrário, os valores mais comuns e bem representados da distribuição devem ter relevância mínima.

No seu último trabalho, Ribeiro e Moniz [12], propõem o uso do *boxplot* ajustado [7] (*adjusted boxplot*) para a obtenção dos pontos de controlo com base na metodologia apresentada por Ribeiro [11] – inicialmente desenhada para lidar com pontos de controlo fornecidos pelo *boxplot* proposto por Tukey [16]. Este método cumpre a meta de obter uma função de relevância contínua que mapeia o domínio da variável objetivo Y no intervalo de relevância $[0, 1]$ para que os valores extremos de Y tenham relevância máxima. Como tal, os valores adjacentes superior e inferior são considerados valores limite para a identificação de valores extremos. Além disso, o valor mediano de Y é considerado um valor de centralidade de irrelevância. Em suma, três pontos compõem o conjunto de pontos de controlo: o valor mediano de Y com valor de relevância igual a zero e os valores adjacentes superior e inferior com valor de relevância igual a um [2]. Nesta abordagem é assumido que todos estes pontos de controlo têm uma derivada igual a zero, representando assim os máximos e os mínimos locais da função de relevância. O método de interpolação *pchip* recebe este conjunto de pontos de controlo e constrói uma função de relevância $\phi()$ que cumpre as especificidades do domínio. No contexto de tarefas de regressão desbalanceadas, tal como as definimos, a utilização do *boxplot* ajustado apresenta-se como uma alternativa melhor, comparativamente ao *boxplot* de Tukey, por duas razões principais. Primeiro, é não paramétrico, e portanto, mais flexível para distribuições subjacentes. Em segundo lugar, ao usar uma medida robusta de assimetria, o método é mais adequado para evitar a perda de casos reais de valores extremos (*outliers*) [7].

De seguida, usamos um caso de aplicação onde ilustramos a obtenção da função de relevância pelo método acima descrito.

2.2 Caso de Aplicação: Previsão de Valores de Concentração de $NO2$

O tráfego rodoviário, as indústrias e os incêndios florestais estão entre os principais fatores que afetam a qualidade do ar. Devido ao aumento da percentagem de substâncias nocivas à saúde, a Organização Mundial da Saúde (OMS) determinou o estabelecimento de limites de concentração para alguns compostos tóxicos [17] - AQGs (*Air Quality Guidelines*).

O conjunto de dados [3] aqui apresentado possui 500 observações recolhidas no âmbito de um estudo que relaciona a poluição do ar numa estrada com o volume de tráfego e as condições meteorológicas. A variável objetivo (`LNO2`) são os valores da transformada logarítimica da concentração de $\mathrm{NO_2}$ medidos em $\mu g/m^3$ para cada hora, em Alnabru (Oslo, Noruega), entre Outubro de 2001 e Agosto de 2003. A Figura 1 mostra a função de densidade de probabilidade da variável `LNO2`.

A Diretiva 2008/50/EC estabelece, com base numa diretriz média anual de $ln(40\mu g/m^3) \approx 3.7$, que os valores de concentração acima de $\ln(150\mu g/m^3) \approx 5.0$, são perigosos e devem ser evitados. Antecipá-los é uma tarefa muito importante no sentido de manter os valores de concentração horária de `LNO2` abaixo do limite estabelecido. Neste contexto, é crucial obter um modelo particularmente preciso na previsão de níveis alarmantes de concentração de $\mathrm{NO_2}$ descritos por valores extremos - e felizmente raros. Na Figura 1 são apresentadas também duas funções de relevância $\phi()$ possíveis, ambas obtidas pelo método de interpolação (*pchip*). Uma, designada por *domain knowledge*, obtida definindo os pontos de controlo com base nas informações do domínio de que dispomos. A outra - *automatic*,

[2] Por omissão, é assumido que ambos os tipos de valores extremos (altos e baixos) são relevantes. No caso de apenas um tipo ser relevante, apenas o valor adjacente correspondente tem valor de relevância igual a um.

[3] Disponível no StatLib Datasets Archive: `http://lib.stat.cmu.edu/datasets/`

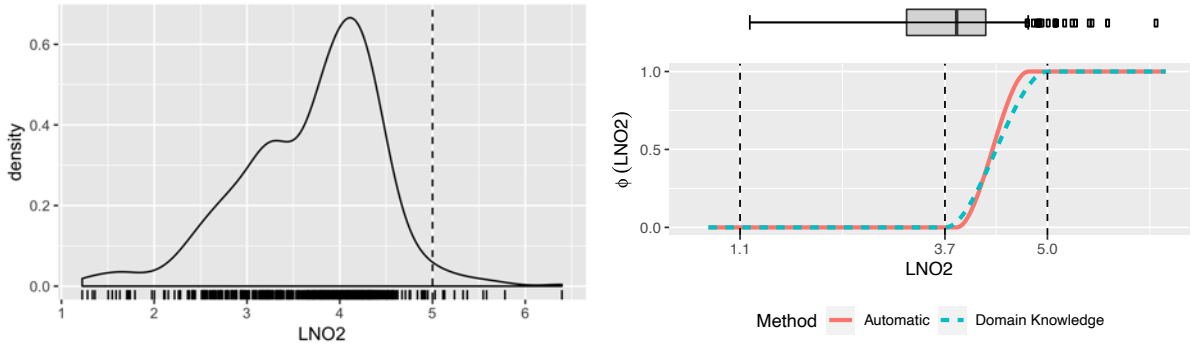


Figura 1: (**Esquerda**) Função de densidade de probabilidade da transformada logarítimica dos valores horários de concentração de NO_2 (LNO2). A linha vertical indica o limite estabelecido pela Directiva 2008/50/EC. (**Direita**) Funções de relevância para a previsão de valores de concentração extremos e elevados de LNO2 inferidas a partir de pontos fornecidos pelo no *boxplot* ajustado (vermelho), e pelo conhecimento do domínio (azul).

obtém os pontos de controlo por meio do *boxplot* ajustado [4] É interessante notar que, em relação aos valores mais críticos neste conjunto de dados em particular, o método automático proposto (vermelho) obtém uma função de relevância semelhante à função de relevância obtida pelas diretrizes estabelecidas com base na Diretiva 2008/50/CE (azul).

De facto, neste como noutros domínios, o método proposto para a obtenção automática dos pontos de controlo a partir do *boxplot* ajustado, poderá ser uma resposta efetiva à especificação da função de relevância da variável objetivo em tarefas de regressão onde a previsão acertada dos valores extremos é crucial.

3 Métricas de Avaliação

Métricas de avaliação como o erro quadrado médio (*Mean Squared Error* - MSE) ou o erro absoluto médio (*Mean Absolute Error* - MAE) são métricas habitualmente utilizadas em tarefas de regressão *standard*. Este tipo de métricas assume preferências de domínio uniformes, focando-se apenas na magnitude dos erros de previsão. Como tal, elas levantam várias questões quanto à sua adequação para avaliar tarefas de regressão desbalanceadas [2, 11, 15]. Vários autores propuseram métricas de avaliação alternativas para dar conta das preferências de domínio não uniformes. Exemplos incluem LIN-LIN [3], RROC [6], curvas REC [1], RECS [13], e versões de *precision* e *recall* para regressão [11, 15]. Contudo, estas medidas apresentam algumas limitações. As propostas assentam na avaliação dos erros agrupados por tipo, sejam eles definidos por erros por excesso/defeito, por limites de tolerância de erro ou por pontos de corte no domínio da variável objetivo para aferir a importância do erros cometidos. Qualquer que seja a opção tomada, elas conduzem à criação de uma divisão artificial no domínio de uma variável contínua.

Com base na revisão da literatura, propomos uma nova métrica de avaliação para responder aos seguintes desafios colocados na avaliação de tarefas de regressão desbalanceada: *i)* minimização de erros de previsão em casos com valores extremos, ou seja, elevada relevância, contrariando a predominância de casos de baixa relevância; *ii)* capacidade de evitar o enviesamento de modelos no sentido da previsão de valores extremos (ou quase extremos) e desconsiderando todos os outros casos; *iii)* dar uma noção assimétrica do erro, ou seja, erros de igual magnitude têm impactos diferentes dependendo de sua relevância; e, finalmente, *iv)* permitir a discriminação, comparação e análise de dominância dos modelos.

[4] Neste exemplo em particular, o foco é apenas em valores extremos altos, pois valores extremos baixos não têm impacto prejudicial à saúde humana.

3.1 SERA: *Squared Error Relevance Area*

Consideremos o conjunto de dados $\mathcal{D} = \{\langle \mathrm{x}_i, y_i \rangle\}_{i=1}^{N}$ e uma função de relevância $\phi(Y) : \mathcal{Y} \rightarrow \{0, 1\}$ definido para a variável objetivo Y. Definimos o subconjunto $\mathcal{D}^t \subseteq \mathcal{D}$ formado pelos casos para os quais a relevância do valor alvo está acima ou igual a um ponto de corte t, ou seja,

$$\mathcal{D}^t = \{\langle \mathrm{x}_i, y_i \rangle \in \mathcal{D} \,|\, \phi(y_i) \geq t\} \tag{1}$$

Podemos obter uma estimativa da relevância do erro quadrado (*Squared Error Relevance* - SER) de um modelo em relação a um corte t, da seguinte forma,

$$SER_t = \sum_{i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2 \tag{2}$$

onde $\hat{y}_i$ e y_i representam o valor previsto e o valor verdadeiro para o caso i, respetivamente. Para esta estimativa, apenas o subconjunto de previsões composto pelos casos $i \in \mathcal{D}^t$, para os quais a relevância do valor verdadeiro está acima de um ponto de corte específico t, são considerados.

Dados os limites dos valores de relevância - $\phi(y) \in [0, 1]$, podemos representar uma curva, onde cada ponto representa o valor de SER_t para um possível corte de relevância t. Esta curva tem propriedades interessantes. O maior e o menor valor de SER_t são obtidos quando todos ($t = 0$) ou apenas os casos mais relevantes ($t = 1$) são incluídos, respectivamente. Além disso, para qualquer $\delta \in \mathbb{R}^+$ tal que $t + \delta \leq 1$, temos que $SER_t \geq SER_{t+\delta}$, dado que $SER_{t+\delta}$ considera um subconjunto (ou todos) dos casos incluídos em SER_t. Estas propriedades garantem que a curva seja decrescente e monótona.

Neste artigo, propomos a utilização da área abaixo da curva de relevância do erro quadrado (*Squared Error Relevance Area* - $SERA$) como métrica de avaliação (cf. Equação 3).

$$SERA = \int_0^1 SER_t \, dt = \int_0^1 \sum_{i \in \mathcal{D}^t} (\hat{y}_i - y_i)^2 \, dt \tag{3}$$

Esta métrica representa a área abaixo da curva SER_t, obtida via integração[5]. Na Figura 2 apresentamos o exemplo de uma curva SER_t. É importante notar que esta curva fornece uma visão geral da magnitude dos erros de previsão no domínio, para diferentes valores de corte de relevância. Assim, quanto menor for a área sob esta curva ($SERA$), melhor é o modelo. Além disso, devemos notar que, quando são assumidas preferências uniformes, $\phi(\mathcal{Y}) = 1$, a métrica $SERA$ é equivalente à soma dos erros quadrados.

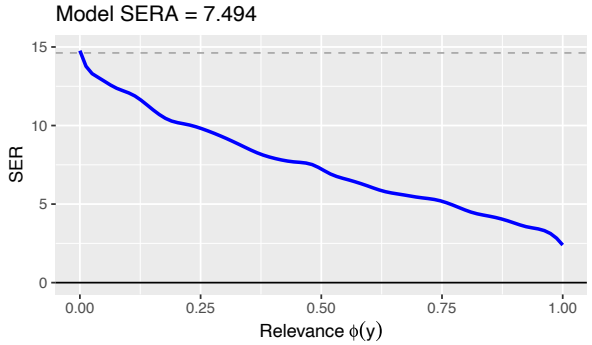


Figura 2: Um exemplo da métrica obtida pela área abaixo da curva de relevância do erro quadrado ($SERA$) para um modelo artificial, baseado na integração da curva da relevância do erro quadrado (SER_t) em relação a um corte t sobre a função de relevância $\phi(.)$. A linha tracejada cinza representa a soma dos erros quadrados para todos os casos.

Embora o nosso objetivo principal seja estimar a eficácia dos modelos na previsão de valores extremos, a métrica $SERA$ não descarta o impacto do desempenho dos modelos em relação a valores no centro da distribuição - os mais comuns. Os erros de previsão feitos em casos de menor relevância têm

[5] A aproximação é obtida pela soma de Riemann pela regra trapezoidal, com Δ igual a 0.001.

muito menos impacto do que aqueles cometidos em casos de alta relevância. Os erros destes últimos são contados mais vezes ao longo dos valores de corte de relevância na soma geral que constitui cada ponto da curva. Portanto, a estimativa $SERA$ engloba as características mencionadas anteriormente para uma métrica de avaliação apropriada em tarefas de regressão desbalanceada.

Com base na dimensão gráfica desta métrica, a sua curva permite também a discriminação, comparação e análise de dominância entre modelos, conforme mostra a Figura 3 e a Tabela 1. A título ilustrativo, usamos diferentes algoritmos de aprendizagem: *random forests* (*rf*) árvores de regressão CART (*rpart*) e *splines* de regressão adaptativa multivariada (*mars*), todos com os parâmetros padrão. A figura mostra uma comparação entre três modelos desses algoritmos de aprendizagem, usando uma métrica de avaliação padrão (MSE) e a métrica $SERA$, em três cenários distintos e comuns em regressão desbalanceada. O primeiro e o terceiro cenário representam configurações onde o melhor modelo de acordo com a métrica MSE é enviesado para a previsão de casos com um valor médio - *naive mean*, ou enviesado para a previsão de todos os casos com valores extremos (ou quase extremos) - *naive extreme*, respetivamente. O segundo cenário descreve uma situação onde as conclusões das métricas MSE e $SERA$ concordam quanto à ordenação dos modelos - *equivalence*. Os resultados das estimativas obtidas pelas métricas MSE e $SERA$ para cada um dos modelos nos três cenários são apresentados na Tabela 1.

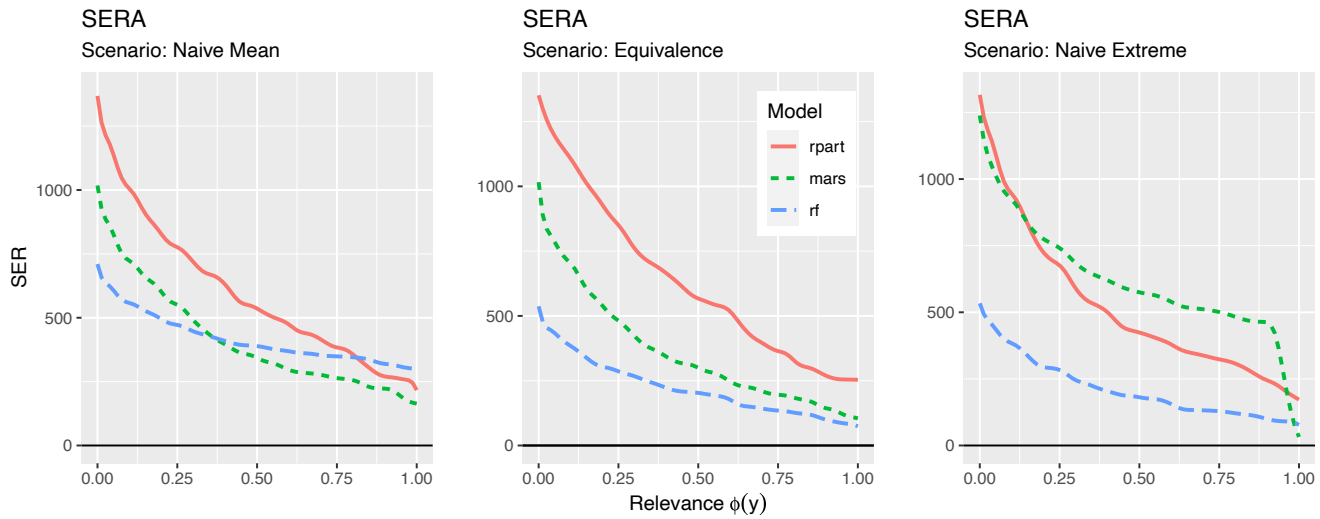


Figura 3: Comparação de três modelos de previsão de diferentes algoritmos de aprendizagem em três cenários de avaliação distintos e comuns em regressão desbalanceada: *i)* naive mean, *ii) equivalence* entre as métricas de avaliação MSE e $SERA$, e *iii) naive extreme*.

Tabela 1: Estimativas MSE e $SERA$ para os cenários de avaliação representados na Figura 3.

Modelo	**Naive Mean**		**Equivalence**		**Naive Extreme**	
	MSE	$SERA$	MSE	$SERA$	MSE	$SERA$
rpart	1392.8	596.8	1373.1	630.1	1335.6	515.4
mars	1033.2	**414.1**	1026.9	357.9	1252.7	616.8
rf	**722.8**	416.3	**548.9**	**220.2**	**540.4**	**210.9**

Para o cenário *naive mean* (esquerda), o modelo *rf* obtém a melhor estimativa MSE, mas uma estimativa $SERA$ pior quando comparado com o modelo *mars*. Com base nas capacidades gráficas da métrica $SERA$ para análise de dominância, percebemos que, embora o modelo *rf* apresente um erro geral de previsão menor (valor na relevância 0), ele demonstra um mau desempenho na previsão de casos com valores extremos, ou seja, a maioria de erros de previsão para este modelo. Isto é também ilustrado pela menor inclinação da curva ao longo do domínio face aos valores de $\phi(.)$. O modelo *mars* mostra uma vantagem preditiva para casos com relevância maior que 0.3. Tendo em conta o nosso objetivo de prever com precisão os valores extremos, podemos constatar que as estimativas de métricas mais *standard*, como é o caso da métrica MSE, podem ser enganosas em tal contexto.

Em relação cenário *naive extreme* (direita), os resultados mostram que o modelo *rf* é o melhor modelo para as métricas MSE e $SERA$. No entanto, os resultados relativos aos modelos *rpart* e *mars* são contraditórios: embora o primeiro apresente uma pior estimativa MSE, apresenta uma vantagem em relação à estimativa $SERA$. Ao analisar a curva da métrica $SERA$ observamos que, embora o modelo *mars* demonstre erros de previsão mais baixos em relação aos valores mais extremos, ele também atinge níveis consideravelmente mais elevados de erro de previsão no restante domínio, mostrando um viés considerável para prever com maior precisão valores extremos à custa de um fraco desempenho nos valores mais comuns da variável objetivo. Como tal, a otimização com a métrica $SERA$ fornece a capacidade de evitar o sobreajuste de modelos enviesados para a previsão de valores extremos (ou quase extremos).

Em relação ao cenário *equivalence*, este ilustra uma concordância entre as métricas MSE e $SERA$ quanto à classificação dos modelos *rf*, *mars* e *rpart*.

A métrica $SERA$ permite, portanto, uma análise completa dos modelos de previsão com erros de previsão em vários níveis de relevância, bem como uma análise de dominância. Estas características representam uma contribuição importante para a avaliação de tarefas de regressão desbalanceada, permitindo uma avaliação focada na capacidade preditiva dos modelos para valores extremos sem, no entanto, descurar o desempenho dos modelos nos restantes valores do domínio da variável objetivo.

4 Conclusões e Notas Finais

Este artigo aborda o problema da regressão desbalanceada e a avaliação de modelos para a previsão de valores extremos. Propomos métodos que permitem um processo mais robusto para a formalização do problema e avaliação para tarefas de regressão desbalanceada. Em primeiro lugar, dada a frequente ausência de informações sobre o domínio, propomos uma abordagem automática e não paramétrica para aproximar as preferências sobre valores extremos da variável objetivo. Em segundo lugar, propomos uma nova métrica de avaliação - $SERA$, que não só é capaz de considerar preferências não uniformes do domínio, mas também fornece uma ferramenta para análise de dominância.

No artigo [12] é apresentado um estudo experimental extenso para responder a questões mais detalhadas, como: *i)* qual o impacto no desempenho dos modelos ao usar métricas de avaliação padrão para seleção de modelo, em comparação com o uso da métrica $SERA$?; *ii)* quais os *trade-offs* preditivos associados à estimativa de desempenho dos modelos usando $SERA$ como critério?; *iii)* será a métrica $SERA$ apropriada para processos de otimização de modelos quando o objetivo é melhorar a previsão de valores extremos? Os resultados obtidos indicam que esta poderá ser uma contribuição importante e que esperamos que venha a promover trabalhos futuros em tarefas de regressão desbalanceada.

Referências

[1] Bi, J. and Bennett, K. P. (2003). Regression error characteristic curves. In *Proc. of the 20th Int. Conf. on Machine Learning*, pages 43–50.

[2] Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2):31:1–31:50.

[3] Christoffersen, P. F. and Diebold, F. X. (1996). Further results on forecasting and model selection under asymmetric loss. *Jour. of Applied Econometrics*, 11(5):561–571.

[4] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer.

[5] He, H. and Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 1st edition.

[6] Hernández-Orallo, J. (2013). Roc curves for regression. *Pattern Recognition*, 46(12):3395–3411.

[7] Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.*, 52(12):5186–5201.

[8] Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232.

[9] López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141.

[10] R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

[11] Ribeiro, R. P. (2011). *Utility-based Regression*. PhD thesis, Dep. Computer Science, Faculty of Sciences - University of Porto.

[12] Ribeiro, R. P. and Moniz, N. (2020). Imbalanced regression and extreme value prediction. *Mach. Learn.*, 109(9-10):1803–1835.

[13] Torgo, L. (2005). Regression error characteristic surfaces. In *Proc. of the Eleventh ACM SIGKDD*, KDD '05, pages 697–702. ACM.

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

[15] Torgo, L. and Ribeiro, R. (Primavera de 2010). Modelos de previsão de valores extremos e raros. *Boletim da Sociedade Portuguesa de Estatística*, pages 15–22.

[16] Tukey, J. W. (1970). *Exploratory Data Analysis*. Addison-Wesley, Prelim. ed.

[17] World Health Organization (2005). Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide.

Qualidade de Dados em Bases de Dados Anonimizadas: Uma Abordagem de Avaliação Mista

Paulo Pombinho,[1] pmimatos@fc.ul.pt
Luís Cavique[1,2], luis.cavique@uab.pt
Luís Correia[1], luis.correia@ciencias.ulisboa.pt

[1] *LASIGE, DI - Faculdade de Ciências da Universidade de Lisboa, Portugal*
[2] *Universidade Aberta*, *Lisboa, Portugal*

1. Introdução

A qualidade dos dados é crucial, tanto ao nível empresarial como governamental, para uma análise adequada da situação que aqueles representam e para as consequentes tomadas de decisão (Batini, 2009). Em projetos de prospeção de dados é especialmente relevante evitar dados com qualidade inferior uma vez que se usam algoritmos que dependem de dados corretos para criar modelos e previsões precisos. Um conjunto de dados com problemas de qualidade pode implicar custos elevados tanto a nível económico como social, com a possibilidade de decisões erróneas serem tomadas quando se olha para dados incorretos (Batini e Scannapieco, 2016; Wang e Strong, 1996). Neste texto, propomos uma abordagem de avaliação de qualidade que considera métricas que lidam com atributos individuais e, adicionalmente, uma análise longitudinal de fluxo, que permite fazer uma avaliação de qualidade que tem em consideração informação contextual. São propostas métricas de Qualidade de Dados por Entrada e Qualidade de Dados por Atributo e, finalmente, é proposta uma medida de Qualidade Global de Dados baseada nessas métricas.
A avaliação da qualidade dos dados não é um tema de investigação novo e existem diversos trabalhos que analisam as diferentes métricas de avaliação de dados. No entanto, a maioria das propostas existentes focam-se na precisão de dados que podem ser comparados com as entidades do mundo real que representam. Esta comparação não é, por vezes, possível se, por diversas razões, a informação sobre a entidade do mundo real não estiver disponível a partir de outras fontes (Coleman-Sebastian, 2013). Este problema é especialmente relevante quando, devido à anonimização dos dados, surgem problemas nos próprios identificadores das entradas na base de dados.
Neste artigo, propomos uma abordagem de avaliação de qualidade que utiliza um conjunto diversificado de métricas de qualidade. Consideramos não só a informação sobre as diferentes dimensões individuais do atributo, mas também o seu contexto semântico. Propomos métricas de avaliação que permitam classificar os dados individuais e agregados e, através desta classificação, atribuir uma qualidade global.
Na secção seguinte apresentamos o trabalho relacionado que aborda os diferentes tipos de métricas de qualidade. Na secção 3, discutimos os problemas que surgem da anonimização dos dados e, na secção 4, os tipos de erros que podem existir. A secção 5 enumera os diferentes tipos de avaliações de qualidade utilizadas e, na secção 6, as métricas de avaliação são calculadas obtendo as medições de qualidade dos dados por entrada e por atributo. Na secção 7, a qualidade global de dados é definida e, finalmente, na secção 8 apresentamos uma análise preliminar da utilização das métricas de qualidade e, na secção 9 as conclusões e o trabalho futuro.

2. Trabalho Relacionado

Embora existam diversas propostas de métricas de qualidade de dados, não existe um acordo real sobre as dimensões que definem a qualidade dos dados nem sobre a sua nomenclatura (Batini et al., 2009).

Há uma grande diversidade de terminologias sobre métricas de qualidade de dados, muitas vezes com significados semelhantes, que são frequentemente demasiado focadas apenas em métricas simples de precisão. Com uma quantidade cada vez maior de dados disponíveis, as instituições estão cada vez mais a utilizar estes dados para apoiar a sua tomada de decisão, embora com a preocupação de não utilizarem dados incorretos, que possam causar más decisões, com grande impacto social e financeiro (Heinrich, 2018).

Adicionalmente, a crescente utilização de técnicas de inteligência artificial com conjuntos de dados muito grandes também pode ver a sua precisão muito afetada pela baixa qualidade dos conjuntos de dados originais (Ding e Li, 2018). As qualidades subjacentes ao *Big Data*, caracterizadas pelos três V's: Volume, Velocidade e Variedade, são um desafio significativo para garantir dados com qualidade e a adição de um quarto V: Veracidade, é sugerida para indicar a importância da qualidade dos dados (Lukoianova and Rubin, 2013).

Em seguida, descreveremos alguns dos trabalhos mais relevantes que se relacionam com as métricas de qualidade que usamos como base para a nossa proposta de avaliação da qualidade de dados.

Pipino, Lee e Wang (2002) enumeram diferentes métricas de avaliação. Listam uma dimensão livre de erro que representa a correção de dados; a completude que pode variar desde um nível de esquema de base de dados, para indicar quão completo é o próprio desenho da base de dados, até à completude por coluna, que se refere ao número de entradas na base de dados com um valor presente vs. em falta; e consistência, que avalia se o mesmo valor é consistente entre diferentes entradas de dados redundantes.

Cai e Zhu (2015) definem cinco dimensões diferentes, sendo as quatro primeiras: disponibilidade, avaliando se os dados são atualizados regularmente; usabilidade, relativamente à fiabilidade da fonte de dados e se o intervalo de valores é aceitável; relevância, avaliando se os dados são pertinentes para o tema em que são utilizados; qualidade de apresentação, relacionada com a compreensibilidade dos dados. A dimensão final é a fiabilidade, em que os autores incluem quatro subelementos: precisão, avaliando se os dados representam o estado real da informação; consistência, estimando se os dados são consistentes com outras fontes; integridade, verificando a sua consistência com as regras estruturais e de integridade dos conteúdos; e a completude para avaliar se problemas ou falta de informação num componente individual de dados compostos por múltiplos componentes, pode ter impacto na precisão e integridade globais.

Sidi et al. (2012) classificam os problemas de qualidade em problemas de fonte única ou de múltiplas fontes, respetivamente se estiverem contidos num conjunto de dados, ou se os problemas forem originários de fontes heterogéneas e das suas interações. Diferenciam entre problemas ao nível do esquema da base de dados e ao nível da instância se o problema de qualidade estiver relacionado com o desenho da base de dados ou se estiver na própria entrada, respetivamente. Os autores também descrevem um conjunto extenso de diferentes métricas a partir das quais destacam a consistência, que avalia se os dados são apresentados nos mesmos formatos ou se são compatíveis; precisão, para avaliar se os valores dos dados correspondem às suas congéneres do mundo real; completude, para aferir se os dados podem representar todos os estados possíveis e significativos; e a atualidade, identificando se os dados estão atualizados.

Jugulum (2014) define as quatro dimensões fundamentais da qualidade dos dados como completude, uma medida que avalia quais os elementos que precisam de estar presentes para atingir os objetivos; conformidade, para perceber se os dados são compatíveis com os formatos exigidos; validade, que aborda a correspondência dos valores dos dados com os seus tipos e intervalos corretos; e a precisão, que avalia se o valor dos dados reflete o mundo real.

Os trabalhos apresentados tratam principalmente da qualidade das entradas individuais na base de dados, não considerando que algumas entradas possam estar relacionadas com outras. No nosso trabalho vamos utilizar uma combinação das métricas descritas por Pipino, Lee e Wang (2002), Sidi et al (2013) e Cai e Zhu (2015) e adaptá-las para avaliar não só as entradas individuais, mas também realizar uma análise longitudinal contextual que considera a relação entre diferentes entradas, permitindo uma deteção mais robusta dos problemas de qualidade nos dados.

3. O Problema dos Dados Anonimizados

A investigação e a tomada de decisões políticas, por exemplo nas áreas da saúde pública ou da educação, estão cada vez mais a utilizar sistemas de informação com conjuntos de dados muito grandes para

apoiar as suas decisões (Cavique, 2020). A utilização de dados de baixa qualidade tem sido, no entanto, frequentemente reportada e leva a efeitos negativos (Chen, 2014).
A utilização de dados pessoais, de acordo com o Regulamento Geral sobre a Proteção de Dados (GDPR, 2016), implica que os conjuntos de dados são frequentemente anonimizados. Isto cria um desafio para a avaliação da qualidade dos dados, uma vez que impede a identificação e confirmação de erros, tornando impossível o uso de outro conjunto de dados que possa ser usado como uma comparação de quais os valores são corretos. Além disso, a anonimização eficaz geralmente está associada a outras proteções para evitar a reidentificação, de modo que uma entrada anonimizada não possa ser identificada através de uma análise cruzada de todos os dados disponíveis. Um exemplo é o conceito de *k*-anonimato (Samarati, 2001), onde os dados divulgados não permitem que uma associação seja feita a um número de indivíduos menor que *k*.
Isto significa que, para alguns domínios de aplicação, todos os dados extraídos devem estar na forma agregada sem entradas individuais. Como tal, os dados utilizados em algoritmos externos podem utilizar apenas dados de grupos de entradas com atributos semelhantes e um contador de quantos deles estão nessas condições específicas.
Estas limitações têm como consequência tornar as métricas tradicionais inadequadas para avaliar a qualidade dos dados anonimizados (Fletcher e Islam, 2014). Apesar destas questões, a investigação tem-se centrado mais nos próprios procedimentos de anonimização, do que na utilização de dados anonimizados na prospeção de dados e na sua avaliação de qualidade. Além disso, se for realizado um procedimento de anonimização impróprio, a utilidade potencial da prospeção de dados pode ser arruinada até mesmo por ganhos marginais de privacidade (Brickell e Shmatikov, 2008). Isto é especialmente importante se considerarmos que o fator mais importante para o sucesso dos projetos de ciência de dados é a utilização das dimensões corretas e o seu tratamento, sendo a parte mais morosa destes projetos (Domingos, 2012).
Algumas soluções para ultrapassar as questões da anonimização propõem a recolha de estatísticas, que são divulgadas com os dados anonimizados, para ajudar na avaliação da qualidade (Inan et al., 2009). Outros investigadores propõem que se invista na escolha dos procedimentos de anonimização certos para resolver os problemas que lhe estão associados ou até mesmo permitir a obtenção de melhores resultados (Buratovic et al., 2012; Silva et al., 2017). Tais abordagens, no entanto, não são viáveis para serem utilizadas por investigadores que utilizam conjuntos de dados grandes e já existentes, que foram previamente anonimizados a nível institucional.
A nossa abordagem visa ser usada em conjuntos de dados anonimizados, onde não há controlo sobre os procedimentos de anonimização. Na secção seguinte, descreveremos os tipos de erros em que nos focamos, neste contexto.

4. Tipos de Erros

Embora existam várias formas de introduzir erros nos conjuntos de dados, existem três problemas de qualidade principais que originam erros: dados em falta, erros de inserção e problemas com identificadores.

Dados em Falta dizem respeito a atributos que não são preenchidos em algumas das entradas. Estes podem ser originados por causas distintas. Podem faltar informações sobre um determinado atributo devido a uma omissão ao introduzir os dados na base de dados ou devido a problemas ou escolhas na própria estrutura da base de dados que possam, em determinadas circunstâncias, permitir a existência de valores nulos. Além disso, consideramos também falta de informação em situações em que, embora o atributo esteja preenchido, tem um conteúdo "não responde", o que é o caso, por exemplo, em bases de dados com informações que são propositadamente omitidas pelas entidades em causa.

Erros de Inserção podem resultar de erros diretos ao inserir informação na base de dados como, por exemplo, escrever um algarismo extra num atributo de idade ou um número de telefone. Estes erros também podem derivar de diferentes formatos utilizados em diferentes momentos, o que introduz inconsistências nos dados, especialmente se as variáveis utilizadas não forem suficientemente restritivas. Por exemplo, escrever um valor de data, num formato não específico, alternando entre o dia/mês/ano e o mês/dia/ano.

Problemas de Identificadores são aqueles que derivam dos procedimentos de anonimização e podem, por vezes, produzir problemas nos dados. Uma vez que não temos nenhuma base de verdade para comparar os resultados, estes problemas são mais difíceis de detetar e corrigir. Os problemas de identificação podem resultar em problemas graves se estivermos a analisar a relação entre diferentes entradas da mesma entidade do mundo real e o processo de anonimização introduziu ruído nas próprias chaves primárias ou estrangeiras. Isto é especialmente relevante em conjuntos de dados que consistem em séries temporais onde cada entidade tem várias entradas de dados recolhidas em determinados intervalos temporais.
As imagens que se seguem mostram exemplos de como identificadores inconsistentes podem criar uma perceção errada e potencialmente grave do que é a realidade.
A figura 1 mostra um exemplo de um conjunto de dados de evolução da saúde de pacientes, onde uma sobreposição de dois identificadores diferentes pode incorretamente mostrar um paciente doente a ser curado quando, na verdade, ambos os pacientes permaneceram na mesma situação em dois pontos de dados temporais diferentes.

Figura 1. sobreposição de dois identificadores diferentes.

Da mesma forma, a figura 2 mostra como, num conjunto de dados de infratores criminais, se um novo identificador for incorretamente gerado para uma entidade existente, uma infração recorrente pode ser identificada como sendo de um novo infrator na sua primeira infração.

Figura 2. atribuição incorreta de um novo identificador a um existente.

5. Tipos de Avaliação da Qualidade

Embora alguns dos erros presentes na base de dados possam ser facilmente identificados, geralmente estes são apenas um pequeno subconjunto do total de erros presentes. Por exemplo, se as diferenças em dois atributos diferentes dos pacientes não forem extremas, os erros podem ser difíceis de detetar. Situações como as descritas nas figuras 1 e 2 são difíceis de sinalizar como erros, pois produzem evoluções de séries temporais que podem estar muito próximas dos casos reais.
Por conseguinte, os potenciais erros devem ser analisados utilizando o contexto envolvente. É provável que erros graves, como os mencionados, sejam associados a outras inconsistências entre os dados.

Neste trabalho, propomos a combinação de diferentes tipos de análise que permitem uma classificação mais robusta de qualidade e tendo em conta as diferentes dimensões de dados.
Consideramos diferentes tipos de análise:

- Análise Individual - considera cada entrada da base de dados isoladamente e visa identificar problemas nos valores dos atributos isoladamente.

- Análise longitudinal de fluxo – considera o fluxo temporal de cada entrada da entidade e avalia cada entrada tendo em consideração as entradas, da mesma entidade, que a precedem e sucedem, para permitir possíveis problemas serem identificados através da sua comparação. Esta análise só é possível em conjuntos de dados com séries temporais.

- Análise de atributos – Foca cada atributo para identificar aqueles que podem ter mais problemas, ao longo de todo o conjunto de dados. Em vez de realizar análises para cada entrada, todas as entradas são consideradas.

Para avaliar a qualidade de cada entrada na base de dados, utilizamos uma combinação das diferentes análises descritas.
Adaptamos a nomenclatura proposta por Pipino, Lee e Wang (2002), Sidi et al. (2013) e Cai e Zhu (2015) descritas na secção 2 e consideramos quatro tipos de métricas de qualidade. Estas métricas são fornecidas através de uma configuração adequada durante o processo ETL (*Extract, Transform, Load*).

Completude - avalia se todos os atributos de uma entrada são preenchidos ou se faltam valores em alguns atributos. É uma métrica binária.

Consistência - avalia se as informações contidas nos atributos da entrada tiveram de ser normalizadas durante a fase de processamento ETL. Ou se dados inseridos não estavam formatados de acordo com o resto do conjunto de dados. É uma métrica binária

Unicidade - identifica as entradas como duplicadas se, por exemplo, houver duas entradas no mesmo período de tempo que estão em conflito entre si. É uma métrica com valores reais contidos no intervalo entre 0 e 1.

Precisão - lida com a exatidão dos valores de atributo de uma determinada entrada. Permite a identificação de entradas com atributos preenchidos de forma incorreta. Como exemplo, se considerarmos a precisão de um atributo numérico, podemos identificar cenários em que o valor é inválido porque está fora do domínio de valores possíveis, tais como valores negativos quando apenas os positivos são válidos. É uma métrica de número real entre 0 e 1.

No que diz respeito à avaliação longitudinal do fluxo dos dados, consideramos uma avaliação de que analisa diferentes dimensões contextuais:

- Variações de atributos – avalia situações em que uma alteração do valor de um atributo, entre dois pontos de tempo consecutivos, de uma mesma entidade, é maior do que seria expectável. Isto só é possível nos conjuntos de dados de séries temporais.

- Variação do valor – executa um cálculo que avalia a probabilidade de um valor de um atributo considerando os outros atributos da mesma entrada.

- Incompatibilidade de Atributos – compara os diferentes atributos na mesma entrada, para os quais se sabe que existe uma correlação, em busca de valores incompatíveis.

Na secção seguinte descreveremos como calculamos a qualidade dos dados de cada entrada.

6. Qualidade Individual de Dados

Utilizando os tipos de avaliação definidos na secção anterior, podemos calcular quatro métricas diferentes de qualidade de dados. Todas estas métricas têm um intervalo entre 0 e 1, onde 0 representa qualidade nula e 1 uma qualidade perfeita.

6.1. Métricas de Qualidade Individual

Para calcular a métrica de **completude,** utilizamos informações relativas à contagem de atributos em falta, numa entrada, sobre o número total de atributos, M (equação 1).

$$Comp = 1 - \frac{1}{M} \sum_{i=1}^{M} [i \text{ está em } falta] \quad (1)$$

A **consistência** dos dados também é classificada, de forma semelhante à completude, utilizando a contagem de atributos com problemas de consistência sobre o número total de atributos, M (equação 2).

$$Cons = 1 - \frac{1}{M} \sum_{i=1}^{M} [i \text{ não é } consistente] \quad (2)$$

Finalmente, a **unicidade** é classificada de acordo com a identificação de entradas duplicadas, caso o identificador esteja em mais que uma entrada no mesmo ponto de dados temporais e, no caso de isto acontecer, avaliando quão diferentes são as entradas duplicadas (equação 3).

$$Unic = \begin{cases} 1, & não\ duplicado \\ 0.8, & duplicado\ com\ diferenças\ menores \\ 0, & duplicado\ com\ diferenças\ substanciais \end{cases} \quad (3)$$

A métrica de **precisão** avalia a exatidão e a probabilidade de determinados valores de atributos. Como indicado anteriormente, classificamos um atributo como inválido, se o valor ficar fora dos domínios desse atributo. No entanto, esta análise pode ser melhorada através da utilização de informação semântica dos valores de atributos esperados, ao considerar outras informações contextuais e permitir uma classificação com diferentes graus de probabilidade (equação 4). V é o valor do atributo, V_{max} e V_{min} são os valores máximos e mínimos válidos, e VP_{max} e VP_{min} são os respetivos valores máximos e mínimos prováveis para o contexto atual da entrada.

$$ProbabilidadeDeValor(V) = \begin{cases} 1, & VP_{min} < V < VP_{max} \\ \frac{V - V_{min}}{VP_{min} - V_{min}}, & V_{min} < V < VP_{min} \\ 1 - \frac{V - VP_{max}}{V_{max} - VP_{max}}, & VP_{max} < V < V_{max} \\ 0, & V < V_{min} \vee V > V_{max} \end{cases} \quad (4)$$

Como exemplo, podemos ter uma análise de um atributo com idade de utentes em idade reprodutiva numa maternidade, em que V é a idade atual, V_{max} e V_{min} são os valores absolutos máximo e mínimo que definem o intervalo de idades válidas, por exemplo valores entre 0 e 130, e VP_{min} e VP_{max} são os valores máximo e mínimos em que é expectável o valor do atributo, por exemplo entre 15 e 44, respetivamente.

Da mesma forma, um atributo pode ser classificado utilizando informações sobre a variação do valor atual V_a (obtida comparando com o ponto de dados anterior da mesma entidade) e as variações mínimas e máximas possíveis, respetivamente V_{min} e V_{max} (equação 5).

$$VariaçãoDeValor(V_a) = 1 - \frac{|V_a| - |V_{min}|}{|V_{max}| - |V_{min}|} \quad (5)$$

Como exemplo, podemos ter uma base de dados em que se registe o consumo mensal de água de clientes e em que se analisa a variação do consumo. Neste caso, V_a é a variação do consumo entre dois

pontos temporais, e V_{min} e V_{max}, as variações de consumo mínimas e máximas que são prováveis entre dois meses consecutivos.

Finalmente, a precisão de atributos em que se sabe existir una correlação pode ser classificada dependendo da comparação entre ambos os atributos e a existência de valores incompatíveis (equação 6).

$$Compatibilidade(V_1, V_2) = \begin{cases} 1, & V_1\ compatível\ com\ V_2 \\ 0, & V_1 incompatível\ com\ V_2 \end{cases} \quad (6)$$

Uma aplicação desta avaliação pode ser exemplificada num conjunto de dados que contenha dados de alunos e em que o atributo idade de um aluno não seja compatível com o ano curricular em que este esteja inscrito.

Tendo calculado os diferentes tipos de métricas de precisão, a precisão global da entrada é definida como a média dos seus diferentes componentes em todos os atributos M (equação 7).

$$Prec = \frac{\sum_{i=1}^{M} ProbabilidadeDeValor_i + \sum_{i=1}^{M} VariaçãoDeValor_i + \sum_{i=1}^{M} Compatibilidade_i}{3M} \quad (7)$$

6.2. Qualidade de Dados por Entrada

A métrica de avaliação de qualidade descrita na secção anterior permite-nos compreender, para cada entrada, a sua fiabilidade em cada dimensão.

Cada métrica é útil para ser capaz de identificar potenciais problemas com os dados. No entanto, uma vez que se espera que problemas de dados mais graves criem problemas de qualidade em várias dimensões em simultâneo (como é o caso, por exemplo, dos problemas que surgem de inconsistências com os identificadores), utilizamos a combinação das diferentes métricas para conseguir obter uma qualidade global para cada entrada na base de dados.

A Qualidade dos Dados por Entrada (QDE) é obtida como a soma ponderada de todas as métricas anteriormente descritas e é classificada como um valor entre 0 (sem qualidade) a 1 (melhor qualidade) (equação 8). Uma vez que, dependendo dos objetivos dos projetos, alguns dos problemas detetados com os dados podem ter uma influência mais severa na utilidade dos conjuntos de dados resultantes. Optámos por usar pesos em cada métrica de forma a poder dar mais importância às métricas mais importantes. Estes pesos são valores reais positivos, cuja soma é 1.

$$QDE_i = w_{Prec} \times Prec_i + w_{Comp} \times Comp_i + w_{Cons} \times Cons_i + w_{Unic} \times Unic_i \quad (8)$$

A qualidade dos dados de entrada permite-nos não só realizar a filtragem inicial das entradas de dados que têm a qualidade mais baixa, mas também, reconhecer os identificadores que estão associados a potenciais problemas de identificação e reportá-los para serem corrigidos.

7. Qualidade de Dados Global e por Atributo

Para utilizar os resultados em algoritmos de prospeção de dados e, quando apenas os dados agregados podem ser extraídos, é também necessário calcular um valor médio de qualidade de dados de entrada (QDE_{med}) de todas as entradas que compõem cada grupo de dados, com *N* entradas (equação 9).

$$QDE_{med} = \frac{1}{N} \sum_{i=1}^{N} QDE_i \quad (9)$$

Embora seja possível realizar a filtragem inicial utilizando a QDE, a utilização de uma qualidade de dados de grupo, no âmbito das técnicas de prospeção de dados, é relevante para ser capaz de filtrar os

dados, de forma dinâmica. Desta forma é possível utilizar diferentes limiares de qualidade, com os quais é possível afinar quais os que dão os melhores resultados e produzem as previsões mais corretas.
Anteriormente já foi descrita a avaliação de qualidade que pode ser alcançada para cada entrada do conjunto de dados. É também vital obter informações para a qualidade de cada atributo, utilizando uma visão global do conjunto de dados em vez de olhar para cada entrada separada.
A qualidade dos dados do atributo (QDA) pode ser obtida usando os diferentes cálculos para diferentes tipos de atributos. Para atributos para os quais só identificamos, numa classificação binária, se existe ou não um problema, utilizamos a proporção de entradas com problemas relativos ao número total de entradas *N* (equação10).

$$QDA\_bin = 1 - \frac{1}{N} \sum_{i=1}^{N} [i\ tem\ problemas] \tag{10}$$

Se, por outro lado, tivermos cálculos de precisão disponíveis, podemos utilizar a média de todos os valores de precisão das entradas nesse atributo, onde *N* é o número de entradas no conjunto de dados (equação 11). Uma alternativa que poderia ser explorada no futuro, é a utilização de um modelo não linear onde é possível, por exemplo, ter um crescimento exponencial que dependa do número de entradas que têm um erro nesse atributo.

$$QDA\_prec = \frac{1}{N} \sum_{i=1}^{N} \frac{ProbDeValor_i + VariaçãoDeValor_i + Compatibilidade_i}{3} \tag{11}$$

Por último, para obter uma visão geral da qualidade no conjunto de dados, um primeiro passo foi incluir pesos para cada tipo de métrica no cálculo da qualidade individual. Isto permite-nos definir uma maior importância para o tipo de problemas de qualidade que podem ser mais severos para os objetivos do projeto. Além disso, para obter uma avaliação verdadeiramente global do conjunto de dados (ou subconjuntos de todo o conjunto de dados), calculamos a qualidade global dos dados (QGD) adicionando as qualidades individuais médias (*N* é o número de entradas no conjunto de dados) com as qualidades médias dos atributos (*M* é o número de atributos) e usando pesos para ajustar a importância dos atributos versus as qualidades individuais (equação 12).

$$QGD = \frac{w_{QDE}}{N} \sum_{i=1}^{N} QDE_i \times + \frac{w_{QDA}}{M} \sum_{j=1}^{M} QDA_j \tag{12}$$

8. Análise Preliminar

As métricas propostas foram aplicadas numa base de dados, anonimizada, com cerca de 20 milhões de entradas, na qual era possível apontar problemas decorrentes de erros de identificadores. Estes podiam ser detetados através da comparação de atributos entre diferentes pontos temporais e da verificação da incompatibilidade dos valores existentes, nomeadamente em (i) entradas duplicadas (referentes a um mesmo ponto temporal, mas com atributos diferentes), (ii) inconsistências nos atributos que definiam uma idade, e (iii) em atributos com variação demasiado alta entre pontos temporais sequenciais.
Uma vez que, para o conjunto de dados utilizado, os erros com maior potencial de provocar efeitos negativos eram os decorrentes da **precisão** de alguns atributos, optou-se por utilizar um peso maior para esta métrica e também um peso ligeiramente maior para a **unicidade**, em detrimento das métricas de **consistência** e **completude**.
Após o cálculo das métricas, foram extraídos os dados com diferentes linhas de corte e feita uma verificação do número de entradas filtradas em cada caso (Tabela 1).

Tabela 1. Percentagem de entradas filtradas por linha de corte.

Linha de Corte	Diferença de entradas (%)
0.4	0.00%
0.5	-0.20%
0.6	-0.22%
0.7	-0.25%
0.8	-0.56%
0.9	-11.04%

Optou-se por filtrar todos os dados com uma QDE inferior a 0.7, uma vez que permitia retirar apenas 0.25% dos dados, correspondendo a menos de 40 mil entradas.
Na análise preliminar efetuada aos dados filtrados, apesar do reduzido número de entradas retiradas, foi possível verificar uma diminuição significativa de potenciais entradas com erros graves (Tabela 2). Os erros considerados mais importantes, como entradas com atributos com valores improváveis (que neste caso correspondia ao atributo idade), foram filtrados 98% dos problemas, e as entradas com duplicação de dados com inconsistências também reduzidas em 46%. As restantes métricas tinham pesos menores e, como tal, provocaram um menor número de dados filtrados. Como exemplo, a existência de um número muito alargado de entradas com atributos não preenchidos, em conjunto com o menor peso utilizado para a métrica de completude resultou apenas numa leve diminuição do número de entradas com este problema.

Tabela 2. Entradas com erro e, destas, qual a percentagem que é filtrada.

Tipo de Erro	Dados com Erro (%)	Entradas com Erro Retiradas (%)
Completude	2.25%	-0.38%
Unicidade	0.02%	-46.24%
Valor Improvável	0.23%	-97.67%
Variação de Valor	0.33%	-7.36%
Erros Combinados	**0.36%**	**-67.98%**

Apesar de a avaliação do impacto das métricas separadas ser útil para perceber que tipo de erros estão a ser detetados e resolvidos, conjuntos de dados anonimizados, com problemas de identificadores, provocam normalmente problemas em diversas métricas em simultâneo. Como tal, é importante verificar que o conjunto de entradas com diversos erros combinados, potencialmente provocados por problemas nos identificadores, foi reduzido em quase 68%.
A análise apresentada permite ter apenas uma noção de como este tipo de avaliação de qualidade pode ser utilizada para melhorar os dados, sem que seja necessário cortar um número demasiado grande de dados. É, no entanto, importante efetuar uma análise mais detalhada, comparando diferentes pesos e linhas de corte, bem como qual o ganho efetivo em cada caso. O número de entradas com erros poderia ter tido uma redução maior através da utilização de uma linha de corte mais alta, resultando, no entanto, num maior número de entradas cortadas. É, assim, necessário efetuar uma avaliação aprofundada, de forma a entender qual o maior ganho de qualidade para um menor número de entradas cortadas.
É importante destacar ainda que a utilização de pesos adequados é essencial para a obtenção de cálculos de qualidade que sejam relevantes para o tipo de utilização que se pretende para os dados, derivada da avaliação preliminar e do tipo de erros que é necessário resolver.

9. Conclusões

Neste artigo, foi apresentada uma abordagem para obter uma classificação de qualidade de dados. A nossa proposta utiliza uma agregação de diferentes tipos de avaliação de qualidade. Além disso,

permite o cálculo de uma qualidade de dados global e utiliza pesos para ajustar a importância de cada métrica usada, para refletir melhor os objetivos do caso em estudo.
Os problemas de identificação, derivados de procedimentos de anonimização, a menos que filtrados, podem ter grande impacto na criação de modelos válidos dificultando o processo, nomeadamente por poderem produzir leituras incorretas dos dados, misturando entradas de diferentes entidades.
A abordagem aqui proposta permite calcular uma métrica individual de qualidade de dados que avalia quão fiável é uma entrada específica no que diz respeito à sua precisão, consistência, completude e unicidade. Além disso, propõe-se igualmente uma métrica de qualidade de dados a utilizar com resultados agregados extraídos das bases de dados.
As avaliações propostas têm a vantagem de permitir que o conjunto de dados seja filtrado dinamicamente, para permitir remover as entradas com má qualidade de acordo com limiares de qualidade específicos.
No que diz respeito ao trabalho futuro, existe o objetivo de avaliar como a utilização das ferramentas de análise de qualidade propostas podem melhorar as previsões feitas pelos algoritmos de prospeção de dados, comparando conjuntos de dados filtrados, com diferentes limiares de qualidade, bem como com conjuntos não filtrados.
Pretendemos também comparar conjuntos de dados com qualidade heterogénea com conjuntos homogéneos, para compreender como a existência de entradas com qualidade muito diversa influência a qualidade global e explorar a adição de uma medida que avalie esta variação.
Por último, as métricas propostas utilizam uma avaliação interna, considerando apenas os dados existentes na base de dados. É relevante adicionar uma avaliação que também permita fazer uma comparação entre os valores globais do conjunto de dados e os disponibilizados por fontes externas, como é o caso das instituições que fornecem informação estatística.

Agradecimentos. Este trabalho foi parcialmente financiado pelos projetos FCT, na unidade de investigação BioISI, ref. UID/MULTI/04046/2103, unidade de investigação LASIGE, ref. UIDB, UIDP/00408/2020 e DSAIPA/DS/0039/2018.

Referências

Batini, C., Cappiello, C., Francalanci, C. e Maurino, A. (2009). In *ACM Computing Surveys,* 41(3), article 16.

Batini, C. e Scannapieco, M. (2016). Data and Information Quality: Dimensions, Principles and Techniques. *Data-Centric Systems and Applications,* Springer.

Brickell, J., e Shmatikov, V.. (2008). The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 70-78.

Buratović, I., Miličević, M. e Žubrinić, K.. (2012). Effects of Data Anonymization on the Data Mining Results. *2012 Proceedings of the 35th International Convention MIPRO*, 1619-1623.

Cai, L e Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the *Big Data* Era. In *Data Science Journal,* 14(2), 1-10.

Cavique, L., Pombinho, P., Tallón-Ballesteros, A. e Correia, L.. (2020) Data Pre-Processing and Data Generation in the Student Flow Case Study. *IDEAL 2020: International Conference on Intelligent Data Engineering and Automated Learning*.

Chen, H., Hailey, D., Wang, N., e Yu, P. (2014). A Review of Data Quality Assessment Methods for Public Health Information Systems. *International Journal of Environmental Research and Public Health.* 2014, 11, 5170-5207.

Coleman-Sebastian, L. (2013). Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. *(1st edition). Morgan Kaufmann, Elsevier*.

Ding, J. e Li, X. (2018). An Approach for Validating Quality of Datasets for Machine Learning. *2018 IEEE International Conference on Big Data*, 2795-2803.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM, 55(10),* 78-87.

Fletcher, S., e Islam M.Z.. Quality Evaluation of an Anonymized Dataset. *2014 22nd International Conference on Pattern Recognition*, 3594-3599.

GDPR, General Data Protection Regulations (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016* on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, OJ L 119, 4.5.2016, 1–88.

Heinrich B., Hristova, D., Klier, M., Schiller, A., e Szubartowicz, M. (2018). Requirements for Data Quality Metrics. *Journal of Data and Information Quality*, ISSN: 1936-1963.

Inan, A., Kantarcioglu, M., e Bertino, E.. (2009). Using Anonymized Data for Classification. *IEEE International Conference on Data Engineering,* 429-440.

Jugulum, R. (2014). Competing with high quality data: Concepts, tools, and techniques for building a successful approach to data quality. *(1st edition) Wiley*.

Lukoianova, T., e Rubin, V. (2013). Veracity Roadmap: Is *Big Data* Objective, Truthful and Credible?. *Advances in Classification Research Online* 24(1), 4–15.

Pipino, L., Lee, Y. and Wang, R. (2002). Data Quality Assessment. In *Communications of the ACM,* 45(4), 211-218.

Sidi, F., Panahy, P., Affendey, L., Jabar, M., Ibrahim, H. and Mustapha, A. (2012). Data Quality: A Survey of Data Quality Dimensions. P*roceedings of the 2012 International Conference on Information Retrieval & Knowledge Management*. 300-304.

Silva, H. O., Basso, T., e Moares, R.. (2017). Privacy and data mining: evaluating the impact of data anonymization on classification algorithms. *2017 13th European Dependable Computing Conference,* 111-116.

P. Samarati. (2001) Protecting Respondents' Identities in Microdata Release, *IEEE Transactions on Knowledge and Data,* 13, (6), 1010-1027.

Engineering, Vol 13 (6), 2001, pp. 1010–1027. International Journal on Uncertainty, Fuzziness and Knowledgebased

Systems, 10 (5), 2002; 557-570.

Wang, R. e Strong D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumer. *Journal of Management Information Systems,* 12(4), 5-34.

Ciência Estatística

• Artigos em Revistas

Papança, F. (2021). A influência do Ensino e da Arte Militar na Evolução da Matemática e da Estatística. *EasyChair Preprint no. 5075.*

Silva Lomba, J., Fraga Alves, M.I. (2020). L-moments for automatic threshold selection in extreme value analysis. *Stoch Environ Res Risk Assess* 34, 465–491 (2020). https://doi.org/10.1007/s00477-020-01789-x.

Braumann, C. A. (2021). Growth and extinction in randomly varying environments: modelling and optimization using stochastic differential equations / Crecimiento y extinción en ambientes que varían aleatoriamente: modelamiento y optimización mediante ecuaciones diferenciales estocásticas. *Revista de Modelamiento Matemático de Sistemas Biológicos - MMSB,* VOL. 1, No. 1, p. 25-36.
Artigo convidado para integrar o Número Inaugural da *Revista Sul-Americana MMSB* de Modelação Matemática.

• Capítulos de Livros

Cabral. I., Caeiro, F. & Gomes, M.I. (2021). Comparação assintótica de duas classes de estimadores de um parâmetro de forma de segunda-ordem. In P. Milheiro, A. Pacheco, B. de Sousa, I. Fraga Alves, I. Pereira, M.J. Polidoro & S. Ramos (eds.), *Estatística: Desafios Transversais às Ciências com Dados - Atas do* XXIV *Congresso da Sociedade Portuguesa de Estatística*. Edições SPE, pp. 107-120.

Gomes, M.I., Henriques-Rodrigues, L. & Pestana, D. (2021). Estimação de um índice de valores extremos positivo através de médias generalizadas e em ambiente de não-regularidade. In P. Milheiro, A. Pacheco, B. de Sousa, I. Fraga Alves, I. Pereira, M.J. Polidoro & S. Ramos (eds.), *Estatística: Desafios Transversais às Ciências com Dados - Atas do* XXIV *Congresso da Sociedade Portuguesa de Estatística*. Edições SPE, pp. 213-226.

• Livros

Título: *Boletim SPE primavera de 2021*
Editor Fernando Rosado
Ano: 2021. Edições SPE.
ISSN: 1646-5903. Depósito Legal: 249102/06
Disponível em: https://www.spestatistica.pt/publicacoes/categoria/boletim-da-spe.

Título: *Estatística: Desafios Transversais às Ciências com Dados - Atas do* XXIV *Congresso da Sociedade Portuguesa de Estatística*.
Editores: Paula Milheiro, António Pacheco, Bruno de Sousa, Isabel Fraga Alves, Isabel Pereira, Maria João Polidoro, Sandra Ramos.
Ano: 2021. Edições SPE.
ISBN: 978-972-8890-47-6
Disponível em: https://www.spestatistica.pt/publicacoes/publicacao/estatistica-desafios-transversais-ciencias-com-dados

Título: *Book of Abstracts of SPE 2021*
Editores: Russell Alpizar-Jara, Dulce Gomes, Lígia Henriques-Rodrigues, Patrícia A. Filipe.
Ano: 2021. Edições SPE.
ISBN: 978-972-8890-48-3
Disponível na página do XXV Congresso da SPE, http://www.spe2021.uevora.pt/

• Doutoramento

Título: Estimação assintoticamente centrada em Estatística de Extremos
Autora: Ivanilda Maria dos Santos Cabral Semedo, *ivanilda.cabral@docente.unicv.edu.cv*
Orientadores: Frederico Almeida Gião Gonçalves Caeiro e Maria Ivette Leal de Carvalho Gomes

A minha tese de Doutoramento teve como objectivo central a estimação semi-paramétrica do índice de valores extremos de um modelo com cauda pesada. O clássico estimador de Hill é um dos primeiros e mais conhecido estimador semi-paramétrico do índice de valores extremos positivo. Contudo, este estimador apresenta usualmente um acentuado viés assintótico ou uma variância elevada, dependendo do número k de estatísticas ordinais de topo envolvidas na estimação. A dificuldade na escolha do nível ótimo, nível que minimiza o erro quadrático médio foi mitigada com a introdução de estimadores MVRB ("minimum variance reduced bias"), sendo o estimador de Hill Corrigido o mais simples estimador MVRB.
Como ponto de partida, foi feita uma revisão de resultados relevantes da literatura. Em seguida, foram estudados mais detalhadamente alguns estimadores semi-paramétricos clássicos, como os estimadores dos Momentos, dos Momentos Mistos e de Hill Generalizado. Assumindo uma expansão de terceira ordem da cauda do modelo, foi possível obter novos resultados acerca do comportamento não degenerado dos estimadores. Esses resultados permitiram identificar, em função do modelo de cauda pesada considerado, qual dos estimadores considerados é o mais eficiente.
São também propostos novos estimadores assintoticamente centrados do índice de valores extremos. Considerou-se uma generalização recente do estimador de Hill, baseada na média de Lehmer, e fez-se a redução do seu viés. Um outro tópico abordado na tese foi o estudo de um novo estimador semi-paramétrico de Hill corrigido. Por fim, abordou-se uma nova classe de estimadores MVRB do índice de valores extremos. Essa classe de estimadores depende de um parâmetro de controlo e, se escolhermos convenientemente esse parâmetro, conseguimos anular os dois primeiros termos dominantes de viés assintótico, mantendo a variância assintótica igual à variância do estimador de Hill. Todos os desenvolvimentos teóricos foram obtidos assumindo uma condição de variação regular de terceira ordem. Este tópico foi finalizado com a apresentação de um algoritmo para determinar o valor dos parâmetros do estimador e o valor da estimativa do índice de valores extremos.
Para uma melhor compreensão da nova metodologia, a tese apresenta não só estudos de simulação de Monte Carlo, como também uma aplicação aos valores do caudal do rio Whiteadder Water em Hutton Castle, na Escócia.
Consegui realizar com mérito, todo esse trabalho, graças à presença constante e incondicional do meu orientador, Professor Auxiliar da Universidade Nova de Lisboa, Frederico Caeiro, e da minha co-orientadora, Professora Catedrática da Universidade de Lisboa, Ivette Gomes.
Obrigada, Professor Frederico!
Obrigada, Professora Ivette!

Ivanilda Cabral

• Prémio Estatístico Júnior

com o apoio do Centro de Matemática da Universidade de Coimbra - CMUC

Trabalhos Vencedores do Prémio Estatístico Júnior 2021:

- ***"Alimentação"***, Escola Básica e Secundária Artur Gonçalves, Torres Novas
  1º lugar (3º ciclo do Ensino Básico)

- ***"O Papel da Mulher na Sociedade"***, Colégio Senhor dos Milagres, Milagres, Leiria
  2º lugar (3º ciclo do Ensino Básico)

- ***"Alterações Climáticas"***, Agrupamento de Escolas do Castelo da Maia, Maia
  1º lugar (Ensino Secundário)

- ***"Abuso Sexual"***, Agrupamento de Escolas do Castelo da Maia, Maia
  2º lugar *ex-aequo* (Ensino Secundário)

- ***"Igualdade de Género"***, Agrupamento de Escolas do Castelo da Maia, Maia
  2º lugar *ex-aequo* (Ensino Secundário)

- ***"Cantina Escolar"***, Escola Profissional e Artística da Marinha Grande, Marinha Grande
  3º lugar *ex-aequo* **(**Ensino Secundário)

- ***"Métodos Contracetivos"***, Escola Mestre Domingos Saraiva, Algueirão Mem-Martins
  3º lugar *ex-aequo* (Ensino Secundário)

A cerimónia de entrega dos Prémios Estatísticos Júnior 2021 decorreu durante o XXV Congresso da Sociedade Portuguesa de Estatística - SPE 2021 (www.spe2021.uevora.pt) realizado de 13 a 16 de outubro.

Agradecemos aos membros do Júri do Prémio pela colaboração e ao Centro de Matemática da Universidade de Coimbra (**CMUC**) pelo apoio.

• Prémio Carreira SPE 2021 - Manuela Neves


SPE
Sociedade Portuguesa de Estatística

Prémio de Carreira

Sociedade Portuguesa de Estatística

À Professora Doutora
Manuela Neves
pela obra científica e dedicação
ao desenvolvimento e divulgação
da Estatística em Portugal

2021

Miguel de Carvalho
Presidente da Direção da Sociedade Portuguesa de Estatística
2021-2023

O Prémio Carreira 2021 foi atribuído à Professora Manuela Neves numa sessão integrada no programa do XXV Congresso Anual da SPE, que foi organizado pela Universidade Évora e que decorreu on-line entre 13 e 16 de outubro de 2021.

Na proposta de nomeação para este prémio pode ler-se

"A Maria Manuela Costa Neves Figueiredo, conhecida por todos nós como Manuela Neves, é uma figura ímpar da Estatística em Portugal. Reconhecida pelo seu saber diversificado e sólido nas áreas de Estatística e Matemática, alia os seus vastos conhecimentos a uma sempre completa disponibilidade para participar em quaisquer atividades para a qual é solicitada. Está sempre pronta a ajudar, mesmo que isso implique um enorme esforço da parte dela, mas sem o demonstrar. A comunidade Estatística em Portugal e a SPE, por esse facto, devem muito à Manuela.

É um nome reconhecido por todos. Participou em várias direções da SPE, foi membro de júris de atribuição de prémios SPE e tem sempre participado na avaliação dos trabalhos para atribuição do prémio Estatístico Júnior. O sorriso e a amabilidade estão sempre presentes, qualquer que seja a circunstância. Muitos foram os estudantes que lhe passaram pelas mãos, quer enquanto professora no ISA, na FCT da Nova, e em outros locais. Participou em inúmeros júris de doutoramento e de mestrado, com arguências vivas, profundas, e sempre de grande atualidade. Praticamente não há júri de associado, catedrático, agregação, professor coordenador, etc, que não conte com a participação da Manuela, julgando sempre com imparcialidade e retidão e com base em argumentos rigorosos e informados."

Na sessão de homenagem à Professora Manuela Neves, realizada pela SPE no âmbito do Dia Europeu da Estatística, em 20 de outubro, fomos convidadas a partilhar o nosso testemunho, momentos e experiências, como antigas alunas de doutoramento. A seguir incluímos excertos desta homenagem, com uma breve descrição do seu percurso académico e profissional, focando a sua contribuição para a área da Matemática no Instituto Superior de Agronomia (ISA), e com os nossos testemunhos pessoais.

Percurso académico

A Professora Manuela Neves licenciou-se em Matemática Aplicada - Ramo de Estatística e Computação, no ano letivo de 1975/1976, pela Faculdade de Ciências da Universidade de Lisboa. Entre 1975 e 1977, foi professora no Liceu Gil Vicente, em Lisboa.

Em 1977 ingressou como Assistente na Secção de Matemática do Instituto Superior de Agronomia (ISA), a sua segunda casa desde então. Doutorou-se em 1990 pela Faculdade de Ciências da Universidade Nova de Lisboa, com a tese "Estimação por Blocos dos Parâmetros da Distribuição de Fréchet - Comparação de Métodos Expeditos", sob orientação do Professor Tiago de Oliveira. De 1990 a 1993 foi Professora Auxiliar na Secção de Matemática do ISA e também Professora convidada do Departamento de Matemática da Faculdade de Ciências e Tecnologia (FCT) da Universidade Nova de Lisboa entre 1990 e 1998. Entre 1993 e 2005 desempenhou o cargo de Professora Associada do ISA. Neste período, em junho de 2003, obteve o grau de Agregação em Matemática pelo Instituto Superior de Agronomia. Desde março de 2005 é Professora Catedrática do ISA.

Contributos na área da Matemática no ISA

A Professora Manuela Neves tem sido uma acérrima defensora do ensino da Matemática no ISA, tendo exercido a sua influência nos diversos cargos que desempenhou ao longo do tempo. A Secção de Matemática do ISA reconhece e agradece o esforço despendido pela Professora Manuela ao longo dos anos na defesa da importância do ensino da Matemática nos cursos do ISA.

Em 1992 esteve ao lado do Professor St. Aubyn na criação e coordenação do Departamento de Matemática do ISA, tendo sido sua presidente por diversos períodos desde 1995.

Em 1995 colaborou ativamente na criação do Mestrado em Matemática Aplicada às Ciências Biológicas do ISA, tendo feito parte da sua Comissão Científica e tendo sido mais tarde sua coordenadora. Este mestrado teve um enorme êxito na altura, contribuindo para a formação matemática e estatística de muitos investigadores na área das Ciências Biológicas. Neste mestrado a Professora Manuela lecionou várias disciplinas, cobrindo diversos tópicos de Probabilidade e Estatística, como Modelação Estatística, Amostragem e Reamostragem, Simulação, Análise de Dados Espaciais.

No âmbito deste mestrado orientou 13 teses em diversos temas que incluem Probabilidade e Estatística fundamental, e aplicações nas áreas da pesca, genética vegetal e saúde.

A Professora Manuela Neves tem tido um papel de relevo no ensino de várias UC na área da Probabilidade e Estatística nos 3 ciclos de estudos do ISA. Salienta-se a responsabilidade da disciplina Estatística do 1º ciclo desde 1989/90, disciplina comum a todos os cursos do ISA (exceto Arquitetura Paisagista), com cerca de 300 alunos; e a participação, desde a sua criação, na UC de doutoramento

Modelos Matemáticos e Aplicações, comum aos vários terceiros ciclos do ISA. O seu entusiasmo e energia na lecionação são contagiantes, sendo muito estimada pelos alunos em geral.
A Professora Manuela foi grande impulsionadora e defensora do doutoramento em Matemática e Estatística no ISA.

Sob o tema "A Matemática nas Ciências Biológicas" organizou, em conjunto com a Professora Isabel Martins do ISA, um ciclo de "Encontros" que decorreu entre 2008 e 2010. Estes "Encontros", que tinham como objetivo divulgar o papel da Matemática e da Estatística no domínio das Ciências Biológicas, suscitaram um grande interesse por parte da comunidade do ISA. Passaram por estes "Encontros" muitos oradores, especialistas que desenvolvem e aplicam métodos matemáticos neste domínio, quer externos quer internos ao ISA.

Os nossos testemunhos pessoais

Maria João Martins (ISA, ULisboa): "Conheci a Manela em 1991, quando entrei para o ISA como Assistente Estagiária, pela mão do Professor António St. Aubyn, que foi meu professor e orientador no Mestrado em Matemática Aplicada no Técnico.
Em 1996, quando chegou a altura de escolher tema e orientador para o meu doutoramento, perguntei à Manela se ela estaria disponível e interessada em orientar-me. Generosamente a Manela apresentou-me à Professora Ivette Gomes, para quem eu era desconhecida, e ambas aceitaram orientar-me na área de Extremos, concretamente na estimação semi-paramétrica de um parâmetro de grande importância em Teoria de Valores Extremos. Rapidamente formamos uma equipa imbatível. Foram tempos muito agradáveis em que trabalhámos intensamente, com enorme entusiasmo. Passámos também muito tempo juntas, tempo durante o qual foi crescendo uma relação de amizade entre as três que nunca esquecerei. Neste período publicamos vários artigos e a tese foi tomando forma, tendo sido discutida em março de 2001.
Aproveito esta oportunidade para agradecer à Manela e à Ivette os ensinamentos, a partilha e a amizade."

Dora Prata Gomes (FCT NOVA & CMA): "Sinto-me privilegiada em poder homenagear uma grande mulher, uma mãe dedicada, inspiradora de ternura, de dádiva e de amor, minha amiga e companheira de muitas viagens, a Professora Manuela Neves. Esta homenagem reflete grande parte dos momentos vividos a trabalharmos juntas no seu gabinete, a partilharmos ideias e experiências nos Workshops e

Conferências e nalgumas aventuras que passamos juntas nas viagens para as conferências. A sua disponibilidade, o seu entusiasmo, a vontade de saber e conhecer mais, assim como as suas objeções, resistências e questões permitiram-me uma maior reflexão e um desenvolvimento pessoal que jamais alcançaria sozinha. O conhecimento só faz sentido quando partilhado. É exatamente pela humildade da Professora que se deve esta partilha. Impulsionadora de felicidade, consegue promover uma excelente relação com os alunos, criando condições para aumentar a motivação e o sucesso dos alunos. A felicidade é mais do que um estado de espírito. É uma ética do trabalho, mas que a Professora aplica-a tão bem.
Um bem Haja! Um chi  "

Clara Cordeiro (FCT-UAlg & CEAUL): "A Professora Manuela Neves foi minha orientadora de mestrado e doutoramento, na área de *bootstrap* aplicado a séries temporais. Durante e após estes períodos, foram vários os trabalhos científicos que realizámos e que divulgámos em congressos e conferências. Durante o meu doutoramento, a Professora partilhou comigo o seu gabinete, os seus livros, os seus artigos e também o seu computador de secretária, para que pudesse efetuar os cálculos computacionais. Foi também neste período que pode observar o seu relacionamento com as pessoas no ISA, desde os senhores da portaria aos alunos, aos colegas, a Professora é muito querida por todos.
A Professora Manuela abraça todos os desafios com uma grande dedicação e entusiamo, é contagiante. Muitas vezes são as suas palavras "Força minha amiga" que ainda hoje não me deixam desanimar em momentos em que as coisas não estão a correr tão bem. Espero ainda ouvir estas palavras nos próximos tempos. Obrigada, Professora Manuela, não só pelos seus ensinamentos, mas acima de tudo pela sua amizade aos longos destes anos. Um grande beijinho da sua "amiga" Clara."

Helena Penalva (ESCE-IPS & CEAUL)**:** "Quando comecei a pensar no que iria escrever à Professora não me ocorreram palavras, mas instantaneamente ocorreu-me o seu sorriso. É sempre com um sorriso nos lábios que eu a vejo quando me recebe no seu gabinete ou em qualquer outro lugar. É um sorriso iluminador e tranquilizador que espelha uma alma franca, amiga, generosa e justa. Foi com esse sorriso que me recebeu pela primeira vez no seu gabinete quando a solicitei para me orientar no meu doutoramento, a que posteriormente se juntou a Professora Ivette Gomes e a Professora Sandra Nunes, às quais também agradeço. Foi também com esse sorriso que me animou nos meus momentos de dúvida ou incerteza. Foi com esse sorriso que me consolou quando enfrentava dificuldades ou insucessos.
A longa caminhada que fiz com ela foi fácil, não porque não tenha envolvido muito trabalho, mas porque trabalhar com a Professora é um prazer, é desafiador e motivador. Sempre com uma enorme disponibilidade e generosidade. Por essa caminhada estou-lhe profunda, e eternamente, grata.
Ainda gostaria de agradecer todo o carinho e conselhos que me tem dado, tanto na vida profissional, sempre me empurrando para a frente, mas também em alguns momentos mais tristes e difíceis da minha vida pessoal.

A Professora Manuela é um exemplo que devo seguir. Ela é das poucas pessoas que conheço que consegue integrar de forma exemplar as diversas facetas de uma mulher. A de mãe, a de mulher e esposa e a faceta profissional e de carreira.
Por tudo muito obrigada!"

A homenagem terminou com o testemunho emotivo da sua filha, Ana Figueiredo, em nome dos três filhos, salientando todo o Amor e atenção que a Mãe tem dedicado à família, especialmente aos filhos e mais recentemente aos netos, conseguindo conciliar a vida familiar com a exigente carreira profissional.

Maria João Martins
Dora Prata Gomes
Clara Cordeiro
Helena Penalva

Retrospetiva do Boletim SPE

O *Boletim SPE* através dos seus "Tema Central"

Primavera de 2021 – Especial Covid: a Estatística ao serviço da sociedade

Outono de 2020 – 40 anos SPE: De onde viemos? Onde estamos? Para onde vamos?

Primavera de 2020 – INE–85 anos de estatísticas a servir o país

Outono de 2019 – Estatística nas Ciências da Saúde

Primavera de 2019 – Séries Temporais de Valor Inteiro

Outono de 2018 – Equações diferenciais estocásticas e algumas aplicações

Primavera de 2018 – Estatística Multivariada – perspetiva no século XXI

Outono de 2017 – O Tema Central da Estatística - um novo olhar

Primavera de 2017 – Incerteza em Engenharia

Outono de 2016 – O Tema Central da Estatística

Primavera de 2016 – Séries Temporais e suas aplicações

Outono de 2015 – Estatística em Genética

Primavera de 2015 – Estatística no Desporto

Outono de 2014 – Estatística no Ensino Básico e Secundário

Primavera de 2014 – (Um) Ano Internacional da Estatística

Outono de 2013 – A "Escola Bayesiana" em Portugal

Primavera de 2013 – Estatística não - paramétrica

Outono de 2012 – Métodos Estatísticos em Medicina

Primavera de 2012 – Estatística no Ensino Superior Politécnico

Outono de 2011 – Análise de Sobrevivência

Primavera de 2011 – Sondagens e Censos

Outono de 2010 – Estatística Espacial

Primavera de 2010 – Data Mining - Prospecção (Estatística) de Dados?

Outono de 2009 – Modelos Econométricos

Primavera de 2009 – Investigação (em) Estatística

Outono de 2008 – Processos Estocásticos

Primavera de 2008 – ALEA - Um sítio do nosso mundo

Outono de 2007 – Bioestatística

Primavera de 2007 – A "Escola de Extremos" em Portugal

Outono de 2006 – Ensino e Aprendizagem da Estatística

também disponíveis em *https://www.spestatistica.pt/publicacoes/categoria/boletim-da-spe*

O MUNDO DA ESTATÍSTICA

ORGANIZAÇÃO PARTICIPANTE

Índice

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR