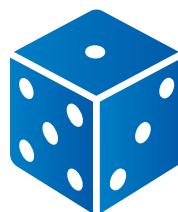




Boletim



SPE

Sociedade Portuguesa
de Estatística

Publicação semestral

primavera de 2024



Rising Stars

Explorando a lei de formação das estrelas por seleção de variáveis	<i>Ana Helena Tavares</i> 19
Misturas de modelos de regressão com efeitos aleatórios	<i>Luísa Novais</i> 22
Dados simbólicos – Os dados do futuro!	<i>Sónia Dias</i> 25
Estudar Séries Temporais através de Redes Complexas	<i>Vanessa Silva</i> 38
O software de fonte aberta na inteligência artificial e na sociedade	<i>José Pombal</i> 43
Há mar e mar, há ir e voltar	<i>Pedro Miranda Afonso</i> 48

Editorial	2
Mensagem do Presidente	4
Notícias	5
<i>Enigmística</i>	14
SPE e a Comunidade	15
<i>Rising Stars</i>	19
Ciência Estatística	49
Doutoramento	51
Prémios Estatístico Júnior 2024	52
Edições SPE - Minicursos	53
Boletim através do Tema Central	54

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 Sobreireense

Editor: Fernando Rosado, fernando.rosado@fc.ul.pt

Sociedade Portuguesa de Estatística desde 1980



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

... “fontes de juventude”...

1. O *Boletim SPE primavera 24* inicia o ciclo 2024-2026 no caminho da Sociedade Portuguesa de Estatística; com novos órgãos sociais, empossados em janeiro. É um momento de gratidão aos que “saem” e aos que “entram” – aos primeiros, pelos serviços prestados e aos segundos pela disponibilidade para servir. Com novas equipas a comandar o projeto, assim se vai construindo a vida desta associação científica desde 1980 – sempre com novos desafios e propostas que os novos tempos exigem. No *Boletim SPE outono 23* foi apresentado o Programa dos Novos Órgãos recentemente empossados (Notícias mais adiante).

2. *Rising stars*, o tema central deste *Boletim SPE primavera 24* surge na sequência de uma sessão temática com o mesmo nome, uma iniciativa conjunta SPE / CLAD, que deu “palco a jovens investigadores ascendentes” e que foi liderada e coordenada pelas colegas Maria do Rosário de Oliveira Silva e Adelaide Figueiredo. À Rosário e à Adelaide, o *Boletim SPE* e o seu Editor agradecem a colaboração na tarefa de coedição e elaboração do Tema Central desta edição. A referida Sessão Temática foi incluída no programa científico do recente XXVI Congresso SPE, realizado em Guimarães (Cf. p. 5 e segs do *Boletim SPE outono 23*, disponível [aqui](#)). O Tema, muito apelativo, conjuga juventude e ciência. Nesse boletim, por sua vez, já se “iniciou” um caminho de reflexão sobre *Educação (e) Estatística*. O *Boletim SPE*, além de uma revista, de algum modo, também pode ser uma verdadeira experiência científica, comunitária e intelectual. Uma Fonte de Juventude!

3. O *Boletim SPE primavera 2021*, no tema central, abordou *Inteligência Artificial e Machine Learning*. Estes são dois temas na ordem do dia aos mais diversos níveis da sociedade. A comunidade científica, por um lado, “cumpre a sua missão pioneira” ao mesmo tempo que, por outro lado, na sociedade civil, se sentem os reflexos da sua produção. A generalidade das “análises” enveredam pelos caminhos mais superficiais, como seria de esperar, numa temática “geneticamente nova” embora por descobrir ainda estejam, “as suas verdadeiras origens” que se confundem com os primórdios da “informática” – uma espécie de “cibernética evoluída”, agora já no século XXI. Releva as origens!

Curiosamente, para outras origens também nos remete a Mensagem divulgada, pelo Papa Francisco, para o Dia Mundial das Comunicações Sociais, em janeiro deste ano; na qual sobre aquelas áreas propõe reflexões bastante profundas, ao afirmar: “A evolução dos sistemas da chamada «inteligência artificial», [...] está a modificar de forma radical também a informação e a comunicação e, através delas, algumas bases da convivência civil. Trata-se duma mudança que afeta não só aos profissionais, mas a todos. A rápida difusão de maravilhosas invenções, cujo funcionamento e potencialidades são indecifráveis para a maior parte de nós, suscita um espanto que oscila entre entusiasmo e desorientação e põe-nos inevitavelmente diante de questões fundamentais: O que é então o homem, qual é a sua especificidade e qual será o futuro desta nossa espécie chamada *homo sapiens* na era das inteligências artificiais? Como podemos permanecer plenamente humanos e orientar para o bem a mudança cultural em curso?”

Esta mensagem leva-nos por um caminho transversal a toda a sociedade; além de nos orientar para “olhar para as origens”. De facto, para além de uma “evolução na ciência” estamos em presença de uma forte inter-relação no binómio “ciência – razão”. É um “desafio científico”!

No Editorial do referido *Boletim SPE primavera 2021*, ficaram algumas palavras interrogativas sobre a temática desafiadora que envolve a *Inteligência Artificial e a Aprendizagem Automática* que lhe é concomitante. O avassalador desenvolvimento – que também é uma característica dessas áreas de investigação – é apelativo para um contínuo aprofundamento.

Assim, fiz um pequeno exercício. Avancei para “uma pesquisa”:

“Consultei um assessor” que perante a minha “sugestão de debate” me “aconselhou”, de imediato. Com uma resposta “seguramente bem ponderada” o meu consultor transmitiu-me o seu relato de modo tão rápido que o seu tempo de resposta, veloz, foi menor do que aquele que eu consumi a escrever a questão. Além da velocidade, numa abordagem “à primeira vista”, é de realçar a organização sistemática da apresentação do resultado da “consulta” – com as propostas enquadradas em janelas de texto – e que mantive, no que abaixo transcrevo. Além da organização, a velocidade também é uma

vantagem? “Devagar que tenho pressa”, como diz a sabedoria popular, está em conflito com “o tempo do nosso tempo”? Aqui fica:

Eu

Como iniciar uma intervenção negativa ou, pelo menos, pouco positiva sobre inteligência artificial e / ou machine learning?

ChatGPT

É importante abordar críticas ou preocupações sobre inteligência artificial (IA) e machine learning (ML) de maneira equilibrada e fundamentada. Aqui estão alguns pontos que podem ser discutidos para apresentar uma perspectiva menos positiva:

1. **Viés nos Dados:**
 - Destacar o risco de viés nos dados utilizados para treinar modelos de machine learning. Se os dados históricos contêm preconceitos, o modelo pode perpetuar e ampliar essas desigualdades.
2. **Desemprego Tecnológico:**
 - Mencionar as preocupações com o impacto no emprego. A automação impulsionada por IA e ML pode levar à substituição de trabalhadores em certas indústrias, resultando em desemprego tecnológico.
3. **Falta de Transparência:**
 - Discutir a falta de transparência em alguns algoritmos de machine learning, especialmente em situações críticas como decisões médicas ou judiciais. A opacidade dos modelos pode gerar desconfiança.
4. **Privacidade e Segurança:**
 - Abordar preocupações com a privacidade, destacando como a coleta maciça de dados para treinamento de modelos pode invadir a privacidade das pessoas se não for gerenciada adequadamente.
5. **Riscos Éticos e de Segurança:**
 - Falar sobre os riscos éticos, como o uso potencialmente inadequado de IA em aplicações militares ou vigilância em massa, e ressaltar a necessidade de regulamentações mais rigorosas.
6. **Aprendizado Não Interpretável:**
 - Apontar que muitos modelos de machine learning são caixas-pretas, o que significa que suas decisões podem ser difíceis de entender. Isso levanta questões sobre responsabilidade e prestação de contas.
7. **Custo e Acesso:**
 - Destacar como o desenvolvimento e a implementação de tecnologias de IA podem ser caros, criando uma divisão digital entre aqueles que têm acesso a essas inovações e aqueles que não têm.

É fundamental lembrar que esses pontos não representam uma visão abrangente sobre IA e ML, mas são aspetos que podem ser discutidos para apresentar uma perspectiva menos otimista. Ao discutir esses temas, é importante considerar também os benefícios e os esforços em andamento para abordar essas preocupações e melhorar a tecnologia.

Aqui termina o texto que “ele sugeriu”, como já se diz, numa “referência personalizada”; quando, de facto, nos referimos a uma máquina. Perante tão “veloz eloquência” o consultor humano, deve “desafiar-se” pois “não faria melhor” que o “seu consultor” e ser uma consequência natural do seu ponto 2 – Desemprego tecnológico.

Agora, falta “apenas” dar continuidade e cumprimento às “suas propostas apelativas”.

Há 50 anos, nas nossas universidades iniciávamos os estudos de “Introdução aos Computadores e Programação” em diversas licenciaturas, não só nas engenharias, mas também nas outras ciências.

Essa iniciação, já cinquentenária, na realidade é agora continuada, mas com diferenças fundamentais nas consequências das decisões e dos “avanços que se fizeram”.

As propostas de debate que hoje nos surgem deixam em aberto um desafio: voltarmos ao tema num próximo *Boletim SPE*. Porque, de futuro se trata, então tudo se coaduna com um *Rising Stars*, como este Boletim é preenchido. Tal como se pronunciaram os novos investigadores na sessão temática do XXVI Congresso SPE: Como estaremos daqui a cinco anos? A construção da resposta é uma fonte de juventude.



O Tema Central do próximo Boletim será

Replicabilidade e Controlo de Confidencialidade

Mensagem do Presidente

Caros Sócios da SPE

É com grande entusiasmo que me dirijo a todos, expressando a minha profunda gratidão pelas mensagens de apoio recebidas desde o início deste mandato. O vosso encorajamento é vital, proporcionando uma motivação ainda maior para trabalharmos em prol do sucesso da nossa Sociedade. Gostaria de aproveitar esta mensagem para destacar a capacidade de mobilização demonstrada pela Sociedade Portuguesa de Estatística (SPE) durante o último Congresso, o XXVI, sediado na cidade de Guimarães. Este evento testemunhou uma participação notável, especialmente marcada pela presença significativa de membros mais jovens, refletindo o envolvimento e o entusiasmo desta nova geração no campo da Estatística. Além disso, o elevado número de trabalhos submetidos para as atas do congresso, que serão novamente publicados pela Springer no formato de livro, é digno de nota, contribuindo assim para a disseminação e a partilha do conhecimento estatístico gerado durante o evento. Este nível de envolvimento ativo dos membros atesta a contínua vitalidade e a relevância da SPE na promoção do conhecimento estatístico em Portugal.

Neste contexto, partilho convosco a visão que temos para a SPE e como contamos com a vossa colaboração para alcançar os objetivos delineados. Apelo à vossa colaboração na angariação de novos sócios, com especial ênfase na inclusão de membros mais jovens, visando ampliar a nossa comunidade e fortalecer a voz da estatística em Portugal. Foi neste contexto que tomamos a decisão de reintroduzir os Prémios Estatístico Júnior. Este prémio tem como objetivo estimular o interesse de alunos do ensino secundário pela Estatística, desempenhando um papel fundamental na formação das próximas gerações de estatísticos.

Durante este mandato, procuraremos consolidar a relevância da Estatística na Ciência de Dados, dinamizando essa conexão através da promoção de eventos científicos de alta qualidade, como congressos, workshops, conferências, mesas-redondas, debates e outras iniciativas, que reúnam especialistas nacionais e internacionais para discutir e partilhar avanços em Estatística, promovendo a participação ativa dos sócios da SPE. Incentivaremos também a redação de artigos na imprensa para destacar o papel fundamental da Estatística na sociedade contemporânea e a sua relação com a Ciência de Dados.

Reiteramos o compromisso em trabalhar em prol dos sócios, contando com a vossa participação, crucial para o sucesso das nossas iniciativas. Reconhecemos a importância de uma comunicação eficaz e estamos empenhados em aprimorar e ampliar a presença da SPE nas redes sociais, como o Instagram, Twitter e WhatsApp, de forma a massificar, dinamizar e expandir a estratégia de comunicação, promovendo a interligação entre Estatística e Ciência de Dados.

Valorizaremos os membros da SPE, continuando programas de reconhecimento para profissionais e investigadores em Estatística. Para 2024, destacamos o Prémio SPE 2024 e o Prémio Jornalismo de Dados SPE 2023, reconhecendo contribuições notáveis para a nossa área.

Termino esta mensagem com uma reflexão: juntos, podemos construir uma SPE mais forte, dinâmica e comprometida com o avanço da Estatística em Portugal. Conto com o vosso apoio e com a vossa colaboração para alcançarmos esses objetivos ambiciosos.

Cordiais saudações,
Luís Machado

Notícias

• Novos Órgãos Sociais da Sociedade Portuguesa de Estatística (2024-26)



MESA DA ASSEMBLEIA GERAL



DIREÇÃO



CONSELHO FISCAL

Conforme noticiado no Boletim SPE outono 23, (p. 9 e segs.), realizou-se o Processo Eleitoral para o triénio 2024-2026. Foi submetida uma única lista cuja composição e programa se podem consultar [aqui](#).

MESA DA ASSEMBLEIA GERAL

Presidente: Pedro Oliveira (Universidade do Porto)

Primeiro Vogal: Ana Luísa Trigos Papoila (Universidade Nova de Lisboa)

Segundo Vogal: Marco André da Silva Costa (Universidade de Aveiro)

DIREÇÃO

Presidente: Luís Filipe Meira Machado (Universidade do Minho)

Vice-Presidente: Lisete Maria Ribeiro de Sousa (Universidade de Lisboa)

Primeiro Vogal: Giovani Loiola da Silva (Universidade de Lisboa)

Segundo Vogal: Lígia Carla Pinto Henriques Jorge Rodrigues (Universidade de Évora)

Tesoureira: Maria João Fernandes Pereira Polidoro (Instituto Politécnico do Porto)

CONSELHO FISCAL

Presidente: Carlos Tenreiro (Universidade de Coimbra – Departamento de Matemática)

Primeiro Vogal: Isabel Maria Simões Pereira (Universidade de Aveiro)

Segundo Vogal: Arminda Manuela Pereira Gonçalves (Universidade do Minho)

No passado dia 5 de janeiro tomaram posse.

FR

• Prémios “Estatístico Júnior 2024”

Após breve interrupção, de novo, está aberto até 31 de maio de 2024 o concurso para atribuição dos prémios “Estatístico Júnior 2024”.

A atribuição dos prémios “Estatístico Júnior 2024” é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio do Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL).

Os Prémios “Estatístico Júnior” têm como objetivo, estimular e desenvolver o interesse dos alunos dos Ensinos Básico e Secundário pelas áreas de Probabilidade e Estatística.

No final deste Boletim apresenta-se o regulamento dos Prémios.

FR

• one-MAST - One-Day Meeting on Advanced Statistics Topics

No passado dia 18 de janeiro teve lugar, no Departamento de Matemática da Universidade de Aveiro, o One-Day Meeting on Advanced Statistics Topics (one-MAST) tendo sido celebrado, em forma de tributo, o 65.º aniversário da nossa colega Isabel Pereira (DMAT e CIDMA).

A parte científica do evento foi destinada a apresentações de trabalhos por colegas e investigadores que, de algum modo, têm tido trabalhos científicos em conjunto com a Isabel. Assim, a Maria Antónia Turkman (CEAUL), que foi orientadora de mestrado e de doutoramento, levou-nos numa viagem do passado ao futuro da Estatística Bayesiana, numa condução exímia cheia de partilha de conhecimento e sabedoria. A Conceição Lopes (DMat e CIDMA-UA), que foi aluna de doutoramento, e Mónica Rodrigues (TEMA-UA), que foi aluna de mestrado, apresentaram trabalhos realizados em coautoria com a Isabel abordando tópicos de séries temporais com aplicação a sistemas de alarme ótimos na área financeira e de análise da regressão no contexto florestal, respetivamente. Feita uma breve pausa para lanche, convívio e algumas fotos, na segunda parte da componente científica, a Maria Eduarda Silva (FEP e LIADD-INESC TEC) deu-nos a conhecer um tópico recente de investigação onde estuda a importância de dados de atividade online, obtidos ao longo do tempo nos motores de busca, na previsão de indicadores macroeconómicos em Portugal, e o Manuel Scotto (IST e CEMAT) apresentou-nos uma visão cronológica de estudos envolvendo modelos de séries temporais de contagem, enfatizando os contributos que têm surgidos em conjunto com a Isabel.

Detalhes do programa científico podem ser obtidos visitando a página do evento

<https://sites.google.com/view/one-daymeeting-cidma>

Antes e após estas cinco palestras, foram realizadas duas sessões de tributo à Isabel, ambas em ambiente mais descontraído. Na primeira foi realizada uma viagem, em formato quizz, pelo percurso académico e profissional da homenageada e, na segunda, foi apresentado um conjunto de testemunhos de estudantes e de docentes. Destacamos aqui o testemunho de Maria Beatriz Matias, docente de Probabilidade e Estatística durante muitos anos na UA, muito especial e importante na vida, não somente científica, da Isabel que, não podendo vir, quis juntar-se ao evento deixando uma mensagem gravada em vídeo.



No final, os participantes passaram para a sala de professores onde os esperavam um lanche ajantarado (que todos afirmaram ser mais um jantar do que um lanche!), muito bem servido e confeccionado pela

EFTA, Escola de Formação Profissional em Turismo de Aveiro, que todos os participantes apreciaram. Cantámos os parabéns, presenteámos a Isabel com um quadro com uma aguarela de uma zona de Aveiro oferecido pelos participantes. No final, a Isabel agradeceu toda a homenagem recebida. Agradeceu a presença carinhosa de todos, presentes e daqueles que não puderam estar, fazendo uma rápida passagem por momentos científicos importantes vividos ao longo dos muitos anos de carreira. Foi uma tarde e final de dia muito bem passados, em ambiente muito fraterno. Um momento em celebração vivido com muita satisfação!

*Isabel, em nome de todos os presentes e dos que não puderam vir, obrigada por seres como és.
É fácil ter carinho e amizade por ti!*

Adelaide e Magda (UA)

• IV European Conference on Queueing Theory (ECQT2024)

É com grande entusiasmo que anunciamos o IV European Conference on Queueing Theory (ECQT2024), dando continuidade ao sucesso dos encontros anteriores em Ghent (2014), Toulouse (2016) e Jerusalém (2018). O ECQT é um evento ímpar, oferecendo a cientistas e profissionais na área da teoria das filas um fórum dedicado à discussão, promoção da investigação, estímulo ao networking e troca de ideias com os seus pares.

Neste ano, entre os dias 7 e 10 de julho, temos o prazer de receber o ECQT2024 no Instituto Superior Técnico, em Lisboa, uma cidade rica em cultura e história, proporcionando o ambiente perfeito para um encontro produtivo e enriquecedor. Além de ser uma oportunidade única para explorar os avanços na teoria das filas, é também uma ocasião especial para destacar a beleza da nossa cidade, a hospitalidade do nosso país, a excelência das nossas instituições e a qualidade dos nossos investigadores.

O ECQT2024 está aberto a todas as correntes da teoria das filas, incluindo desenvolvimentos teóricos, avanços em aspectos metodológicos e computacionais, bem como aplicações inovadoras no mundo real. O evento cobre, portanto, um espectro bastante abrangente da área. A Comissão Organizadora promoverá a participação de alguns palestrantes de países em desenvolvimento, fornecendo apoio financeiro para cobrir custos de alojamento local e taxa de inscrição.

Mais informação em:

<https://sites.google.com/gcloud.fe.up.pt/ecqt2023?pli=1>

Esperamos contar com a participação ativa de todos para tornar este evento mais um marco de sucesso na trajetória do ECQT.

Contamos convosco!

Atenciosamente,
Pl'a Comissão Organizadora,
António Pacheco e Vanda Lourenço

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.

- *REVSTAT-Statistical Journal is an Open Access (OA) blind peer-reviewed journal, registered in the Directory of Open Access Journals (DOAJ), and it is published quarterly, in English, by Statistics Portugal.*
- *There are NO FEES for publishing accepted manuscripts, with Digital Object Identifier (DOI) permanent article identifier and made available in OA.*
- *All published papers are licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).*

Em 2023/2024, a REVSTAT - Statistical Journal lançou os Volume 21 – Número 4 e Volume 22 – Número 1, com os artigos listados abaixo.

Volume 21, Issue 4:

<https://revstat.ine.pt/index.php/REVSTAT/issue/view/74>

- "A Note on the Stochastic EM Algorithm Based on Left Truncated Right Censored Data From Burr XII Distribution" by Debanjan Mitra
- "Random Environment Integer-Valued Autoregressive Process with Discrete Laplace Marginal Distributions" by Bogdan A. Pirkovic, Miroslav M. Ristic, Aleksandar S. Nastic
- "Bias Reduction of Maximum Likelihood Estimates for an Asymmetric Class of Power Models with Applications" by Yolanda M. Gómez, Bruno Santos, Diego I. Gallardo, Osvaldo Venegas, Héctor W. Gómez
- "Estimation, Prediction and Life Testing Plan for the Exponentiated Gumbel Type-II Progressive Censored Data" by Kousik Maiti, Suchandan Kayal
- "New Members of The Johnson Family of Probability Distributions Properties and Application" by Piotr Sulewski
- "Exponential-Gaussian Distribution and Associated Time Series Models" by Nitha K.U, Krishnarani S.D.

Volume 22, Issue 1:

<https://revstat.ine.pt/index.php/REVSTAT/issue/view/75>

- "A Letter from the Editor-in-Chief" by Manuel Scotto
- "A Study on Zografos-Balakrishnan Log-Normal Distribution Properties and Application to Cancer Dataset" by D. S. Shibu, S. L. Nitin, M. R. Irshad
- "Bounds on Negative Binomial Approximation to Call Function" by Amit N. Kumar
- "Tales of the Wakeby Tail and Alternatives When Modelling Extreme Floods" by Jesper Rydén
- "On q-Generalized Extreme Values under Power Normalization with Properties, Estimation Methods and Applications to Covid-19 Data" by Mohamed S. Eliwa, E. O. Abo Zaid, Mahmoud El-Morshedy
- "Identifiability Analysis Using Data Cloning" by José Augusto Sartori Junior, Márcia D'Elia Branco
- "Analysis of Antibody Data Using Skew-normal and Skew-T Mixture Models" by Tiago Dias Domingues, Helena Mouriño, Nuno Sepúlveda

• A propósito de “Métodos Estatísticos em Educação”

No artigo *Métodos Estatísticos em Educação: “speaking truth to power”*, publicado no *Boletim SPE outono 23*, (p. 58 – 62) afirmei que

“Relativamente à investigação aplicada em aferição e avaliação educacional com base em microdados recolhidos através dos exames nacionais, até onde conhecemos, o organismo público proprietário dos dados nem anonimizados os cede ou disponibiliza em “sala segura” para fins de investigação científica” (Ferrão, 2023, *Boletim SPE outono 23*, p. 60 <https://www.spestatistica.pt/publicacoes/categoria/boletim-da-spe>).

Os microdados que eu havia solicitado foram, entretanto, disponibilizados na sala segura da Direção-Geral de Estatísticas da Educação e Ciência (DGEEC). O procedimento definido para o acesso aos microdados está de acordo com a legislação portuguesa para a proteção de dados pessoais. Qualquer investigador que pretenda iniciar o processo de acesso a dados da Educação e Ciência, deve enviar o pedido eletrónico à DGEEC através do formulário disponível em:

<https://www.dgeec.medu.pt/art/6499db7d9eff36f307f07bdb/65293837121f641a986cc618/65495a4a79026a41502e3e82/65495ac879026a41502e3e83>.

Maria Eugénia Ferrão

• Statistics on Health Decision Making – 6th SHDM

O próximo Encontro *Statistics on Health Decision Making* irá decorrer a 4 e 5 de julho na Universidade Aveiro. Nesta edição, as palestras convidadas são subordinadas ao tema Inteligência Artificial.

O objetivo principal do Encontro é proceder à discussão de linhas estratégicas da Estatística na decisão na área da saúde sendo uma oportunidade para a partilha de conhecimento e dinamização da investigação, nesta área, em Portugal.

Assim, convidamos todos a participarem no *6th Statistics on Health Decision Making: Artificial intelligence* a decorrer na cidade de Aveiro, capital portuguesa da cultura 2024.

Mais informações poderão ser obtidas em <https://www.ua.pt/pt/estatisticamedica/6th-SHDM>.

Até julho 2024, em Aveiro!

Ana Helena Tavares

• CIDMA - Workshop do GPE

O XII Workshop of Probability and Statistics group irá realizar-se no dia 29 de maio. O evento, inserido nas atividades do grupo de Probabilidades e Estatística do Centro de Investigação e Desenvolvimento em Matemática Aplicada (CIDMA) da Universidade de Aveiro, pretende ser uma oportunidade para a partilha de conhecimento científico, promovendo uma discussão multidisciplinar.

As sessões contarão com o contributo de investigadores da área de Probabilidades e Estatística e investigadores de outras áreas científicas que aplicam métodos probabilísticos e estatísticos nos seus trabalhos de investigação.

A participação no workshop é gratuita, mas a inscrição é obrigatória. Para informações detalhadas visite a página do evento em <https://sites.google.com/view/workshopsps-cidma/>.

Ana Helena Tavares

• Prémio Nascimento Leitão - Bayer Prize

O Prémio Nascimento Leitão – Bayer Prize visa distinguir a comunicação em poster no evento *Statistics on Health Decision Making*, que se evidencie pela originalidade, impacto para a decisão em saúde, rigor científico e qualidade da apresentação.

Aos autores do trabalho selecionado será atribuído um prémio monetário global no valor de 500 euros. A submissão de trabalhos termina a 30 de abril de 2024. Pode consultar o regulamento do prémio em https://www.ua.pt/pt/estatisticamedica/medico_antonio_nascimento_leitao.

Ana Helena Tavares

• International Symposium on Nonparametric Statistics 2024

O Centro de Matemática da Universidade do Minho (CMAT/UMinho) associa-se ao Institute of Mathematical Statistics (IMS) na realização do International Symposium on Nonparametric Statistics 2024 (ISNPS 2024), a decorrer entre 25 e 29 de junho de 2024, em Braga.

Inspirados pelo êxito das anteriores edições desta conferência, em Chalkidiki (Grécia, 2012), Cádiz (Espanha, 2014), Avignon (França, 2016), Salerno (Itália, 2018) e Paphos (Chipre, 2022), o ISNPS 2024 decorrerá a seguir ao São João de Braga, a festa sanjoanina mais antiga de Portugal. O evento terá lugar no Altice Forum Braga, um distinto local de conferências situado a apenas 15 minutos a pé do centro da cidade de Braga.

Esta conferência internacional reunirá especialistas e académicos de renome para discutir os mais recentes avanços e tendências em diversas áreas da estatística não paramétrica. Pretendemos proporcionar um ambiente propício à troca de ideias de investigação, incentivar a colaboração entre investigadores de todo o mundo e contribuir para o progresso contínuo deste campo científico. O programa do evento incluirá palestras plenárias, palestras convidadas especiais, apresentações convidadas, comunicações e uma sessão de pósteres, abrangendo todas as vertentes da estatística não paramétrica.

Oradores convidados

- Irène Gijbels (KULeuven)
- Jane-Ling Wang (UCDavis)
- Sílvia Gonçalves (McGill)
- Peter Bühlmann (ETH Zürich)
- Peter Mueller (Texas at Austin)
- Wenceslao González-Manteiga (Santiago de Compostela)
- Andrew Barron (Yale)

Datas importantes

- Data limite para envio de resumos – 4 de abril de 2024
- Notificação de aceitação para comunicação – 18 de abril de 2024
- Inscrição a custo reduzido – até 29 de abril de 2024
- Inscrição regular – de 30 de abril a 31 de maio de 2024
- Inscrição tardia – a partir de 1 de junho de 2024
- Conferência – 25 a 29 de junho de 2024

International Symposium on Nonparametric Statistics 2024 - ISNPS 2024 **Braga, 25 a 29 de junho de 2024**

Informações mais detalhadas do evento podem ser consultadas em <https://w3.math.uminho.pt/ISNPS2024>

Raquel Menezes

• Declaração sobre Ética Profissional

A SPE – Sociedade Portuguesa de Estatística é membro institucional do ISI – Instituto Internacional de Estatística. O ISI está envolvido, no estabelecimento de uma declaração sobre ética profissional, ao longo de décadas. A primeira declaração foi apresentada na Assembleia Geral na celebração do Centenário em 1985. O objetivo da declaração era fornecer orientação e não regulamentação.

Com o passar do tempo, o ISI decidiu investigar a necessidade de atualizar a declaração e, consequentemente, foi adotada uma versão substancialmente revista em 2010. Embora o conteúdo da declaração de 2010 permaneça válido, a utilização crescente de uma diversidade de fontes de dados, conjuntos de dados interligados e métodos estatísticos computacionalmente intensivos exigiram algumas atualizações que foram introduzidas em 2023. A versão em inglês da declaração atualizada está disponível ([aqui](#)); também em português.

Mais informa o ISI: “O objetivo da declaração é permitir que os julgamentos e decisões éticas individuais sejam informados por valores e princípios partilhados. Reconhece que a operação de um princípio pode impedir a operação de outro princípio. Os estatísticos podem ter de fazer escolhas entre princípios. Não tenta resolvê-los com regras rígidas a serem aplicadas. Em vez disso, fornece um quadro para ajudar os estatísticos a tomar estas decisões, compreendendo e considerando ativamente fatores concorrentes nas circunstâncias que enfrentam”.

A referida versão, no Preâmbulo, anuncia “ao que vem”:

“A Declaração de Ética Profissional do ISI consiste numa declaração de Valores Profissionais Partilhados e num conjunto de Princípios Éticos derivados desses valores. Para os efeitos do presente documento, a definição de quem é um estatístico vai muito além daqueles com diplomas formais na área, incluindo um vasto leque de criadores e utilizadores de dados e ferramentas estatísticas. Os estatísticos trabalham numa variedade de contextos económicos, culturais, jurídicos e políticos, cada um dos quais influencia a ênfase e o foco da investigação estatística. Também trabalham num dos muitos ramos diferentes da sua disciplina, cada um envolvendo as suas próprias técnicas e procedimentos e, possivelmente, a sua própria abordagem ética. Qualquer que seja a sua área de especialização, os princípios éticos dos estatísticos são parte integrante da sua competência profissional e devem fazer parte de uma formação estatística abrangente. Os estatísticos trabalham em diversas áreas como, por exemplo, a economia, a psicologia, a sociologia e a medicina, cujos profissionais possuem convenções éticas que podem influenciar a sua conduta. Mesmo dentro do mesmo contexto e ramo da estatística, os indivíduos podem enfrentar várias situações e condicionalismos em que possam surgir questões éticas. O objetivo desta declaração é permitir que os juízos e decisões éticas individuais do estatístico sejam informados por valores e experiências partilhadas e não por regras rígidas impostas pela profissão (...)”

FR

• International Conference on Statistics and Data Science (ICSIDS)

A 2023 *Institute of Mathematical Statistics (IMS) International Conference on Statistics and Data Science (ICSIDS)* – 2ª IMS ICSIDS – teve lugar no Centro Cultural de Belém, Lisboa, Portugal, de 18 a 21 de Dezembro de 2023, com o patrocínio do CEAUL (*Centro de Estatística e Aplicações, Universidade de Lisboa*) e da SPE (*Sociedade Portuguesa de Estatística*), tal como se pode ver em <https://sites.google.com/view/icsds2023/home>. De acordo com o que as organizadoras, Regina Liu e Annie Qu, escreveram no *Boletim* de Março do IMS acessível em <https://imstat.org/latest-ims-bulletin/>, a ICSIDS 2023 mais do que duplicou o sucesso da ICSIDS 2022 em Florença!

Na realidade, e relativamente ao Programa Científico, gostaríamos de mencionar que houve mais de 550 participantes de 40 países, desde estudantes e jovens no início e meados das suas carreiras, até investigadores e profissionais seniores, afiliados à Academia, Governo e Indústria. Foram cobertas amplas áreas de Estatística (a Ciência dos Dados...) e Ciência de Dados, tal como se pode ver no livro associado (<https://sites.google.com/view/icsds2023/full-program-book>), onde se podem contabilizar 475 resumos. O programa científico teve 4 sessões plenárias, 57 sessões convidadas, 7 sessões com a apresentação de artigos contribuídos, uma sessão de prémios de viagens para estudantes, uma sessão em

que foram expostos 57 posters e uma sessão sobre “Amigos da Indústria do IMS” iniciada pelo IMS para promover colaborações entre a Academia e a Indústria.

Apesar de todas as atrações turísticas de Lisboa e do inesperado período de sol em meados de Dezembro, a conferência teve uma fortíssima participação, e as 4 sessões plenárias deixaram sempre o auditório repleto de participantes, conseguindo alguns apenas lugares em pé. O que realmente vimos acontecer in loco, para além dos numerosos comentários positivos que nos foram transmitidos, atestou claramente a qualidade do programa da conferência. Além da apresentação de inúmeros avanços importantes em estatística e ciência de dados, as palestras também cobriram amplos desafios e oportunidades para esta área, e foram na realidade inspiradoras e desafiantes

Gostaria ainda de referir que, de certo modo para comemorar os 40 anos do Encontro do Vimeiro em 1983 (Statistical Extremes and Applications – SEA 1983), e também os 10 anos da Conferência EVT 2013, onde EVT não significa 'Extreme Value Theory' mas sim 'Extremes in Vimeiro Today', Congresso organizado pelos minhas colegas e grandes amigas, Antónia Amaral Turkman, Isabel Fraga Alves e Manuela Neves, aceitei organizar uma sessão intitulada, ‘Statistics of Extremes & Applications: Remembering the 40 Years of Vimeiro’. Dediquei essa sessão ao Feridun Turkman, um dos co-organizadores do SEA 1983, que participou na II ICSDS, e também à memória de Tiago de Oliveira, Presidente da Comissão Organizadora do SEA 1983, que infelizmente não está connosco desde 1992, mas que continua presente nos nossos corações. Os oradores foram: Miguel de Carvalho, da Universidade de Edimburgo, com uma palestra intitulada ‘Semiparametric Bayesian Modeling of Nonstationary Joint Extremes’; Isabel Fraga Alves, da Universidade de Lisboa (e do CEAUL), com uma palestra sobre ‘The Myth of the Kraken: When Mythology Meets EVT’; Ana Cristina Freitas, da Universidade do Porto, com uma palestra intitulada, ‘Clustering for dynamically generated stochastic processes’; Holger Rootzen, da Universidade Chalmers, Suécia, com uma palestra intitulada: ‘Is there a cap on how long a human can live? Truncation, censoring and extreme value modelling’.

Queria finalmente referir que tivemos mais de 300 participantes no banquete da conferência, realizado na Casa do Alentejo, pois é essa a única fotografia que tenho com quase todos os elementos do CEAUL que pertenceram ao ‘Local Arrangements Committee’ (Eunice Carrasquinha, Giovanni Silva, Ivette Gomes, Lisete Sousa, Soraia Pereira, Teresa Oliveira e Tiago Marques), como se pode ver na Figura seguinte. Todos os participantes puderam desfrutar de uma refeição tradicional portuguesa, bem como de um interessante espetáculo de Fado.



Foram quatro dias repletos de um programa científico muito interessante, embora muito pesado, num ambiente de saudável e alegre convívio entre os participantes, muitas vezes no belo terraço do Centro Cultural de Belém, com o Tejo e vários marcos históricos famosos à vista.

Ivette Gomes

(Membro do ‘Program Committee’ e ‘Co-Chair’ do ‘Local Arrangements Committee’, em conjunto com Eunice Carrasquinha e Teresa Oliveira)

Enigmística de mefqa

RE GR ESS ÑO

~~o~~
m~~x~~imento

Enigmas 43 e 44

No Boletim SPE outono de 2023 (p. 30):

ado ado ado ado

E	M	O	R	S	T	X
2	1	1	1	1	1	1

Enigma 41: dados truncados

Enigma 42: estatística de EXTREMOS

Literacia para a Qualidade em Estatísticas Oficiais

Pedro Campos, pedro.campos@ine.pt
Instituto Nacional de Estatística, Serviço de Métodos Estatísticos

J. Pinto Martins, pinto.martins@ine.pt
Instituto Nacional de Estatística, Serviço de Difusão

A Literacia Estatística

Vivemos imersos num "Big Data deluge", onde a quantidade de informação gerada supera a capacidade de gestão e análise eficaz. Problemas de qualidade, como outliers, dados ausentes, subcobertura e questões éticas são prevalentes nos dados disponíveis online, dificultando a sua integração como dados úteis, especialmente nas estatísticas oficiais. Neste breve artigo propomos um novo quadro conceitual em desenvolvimento, visando compreender as dimensões da qualidade em estatísticas oficiais e sua relação com a literacia estatística, que permitirá alertar o utilizador comum sobre a indispensável qualidade dos dados, capacitando-o a tomar decisões informadas¹. Não existem consensos quanto a definições de literacia estatística. Sharma (2017) faz uma revisão de literatura sobre o tema e conclui que os investigadores em todo o mundo apresentam diversas visões sobre a definição de literacia estatística. A literacia estatística é um conceito complexo que exige não só uma série de competências básicas (leitura, compreensão e comunicação), mas também competências cognitivas de ordem superior de interpretação, previsão e pensamento crítico. A capacidade de interpretar estatísticas de forma crítica e de refutar afirmações não é inata; estas competências têm de ser ensinadas para que os estudantes e os cidadãos em geral se tornem cidadãos informados.

Ao longo dos anos, a investigação em educação estatística explorou várias abordagens. A taxonomia SOLO (*Structure of the Observed Learning Outcome*) (Biggs e Collis, 1982) classifica resultados de aprendizagem. O método Aprender Fazendo em Estatística (Smith, 1998), advoga pelo desenvolvimento do raciocínio estatístico por meio de estratégias de aprendizagem ativa. O *Profile of Statistical Understanding* (Reading, 2002), procura identificar o que "pode ser" esperado dos alunos. Os modelos de literacia estatística do adulto propostos por Gal (2002, 2024), sustentam que o comportamento estatisticamente letrado envolve cinco bases de conhecimento inter-relacionadas, juntamente com disposições e crenças facilitadoras.

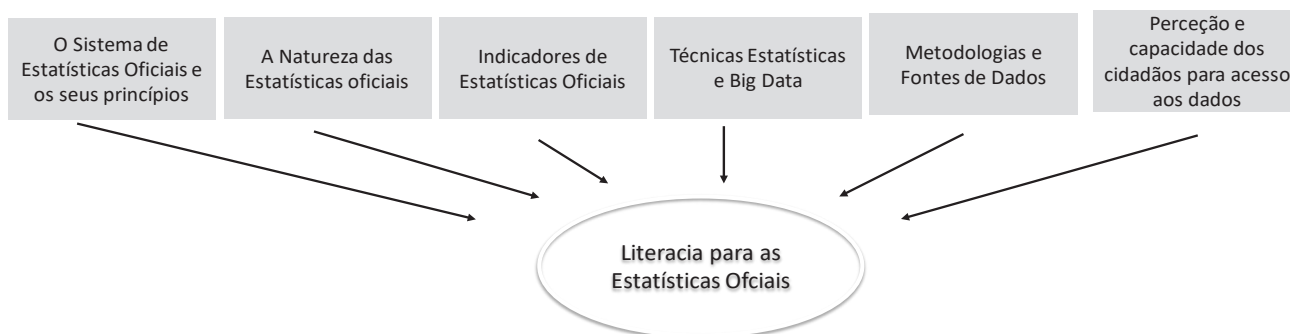


Figura 1. Competências de literacia estatística. Adaptado de Gal and Ograjensek (2017)

¹ Este trabalho foi parcialmente apresentado no Encontro “Matemática Com Vida: diferentes olhares sobre Dados e Probabilidades”, realizado na Universidade de Aveiro, em setembro de 2023.

Outras abordagens, como a literacia de dados (Gould, 2017) e o raciocínio estatístico (Sabbag, Garfield e Ziefler, 2018), têm surgido, mas poucas consideram a relação com a qualidade, especialmente na perspectiva de estatísticas oficiais.

No que respeita à literacia estatística dos adultos, nomeadamente em estatísticas oficiais, Gal e Ograjensek (2017) identificam seis elementos relacionados com as estatísticas oficiais sobre os quais os utilizadores adultos não especializados devem possuir conhecimentos.

Num mundo saturado de informações, é essencial que jovens, decisores e cidadãos possuam sólido conhecimento sobre a qualidade dos dados diários. A verificação da exatidão e compreensão do contexto são cruciais, sobretudo em um cenário onde notícias frequentemente veiculam essas informações. No entanto, é notável que as pessoas nem sempre avaliam e selecionam fontes de dados confiáveis, pois falta alguma literacia para a perceção da qualidade da informação com que lidamos no dia-a-dia.

A Literacia para a Qualidade em Estatísticas Oficiais

O Eurostat, seguindo o Código de Conduta para as Estatísticas Europeias (Eurostat, 2017), destaca dimensões cruciais como pertinência, exatidão, fiabilidade, atualidade, pontualidade, coerência, comparabilidade e clareza. Na verdade, estes são os pilares que sustentam a confiabilidade da informação estatística. Cada dimensão da qualidade definida pelo Sistema Estatístico Europeu desempenha um papel crucial na garantia de estatísticas robustas e confiáveis.

Pertinência: Garante que as estatísticas abordem as necessidades dos utilizadores, sendo relevantes para os contextos e propósitos específicos.

Exatidão: Reflete a proximidade dos resultados estatísticos à realidade que pretendem representar, assegurando a precisão das medições e cálculos.

Fiabilidade: Envolve a consistência e estabilidade dos resultados ao longo do tempo, proporcionando confiança na sua utilização e interpretação.

Atualidade: Certifica-se de que as estatísticas são oportunas e refletem a situação mais recente, sendo essenciais para decisões informadas em tempo hábil.

Pontualidade: Relaciona-se com a entrega das estatísticas dentro dos prazos estabelecidos, permitindo a sua utilização atempada e evitando a obsolescência da informação.

Coerência: Garante a consistência interna das estatísticas, evitando discrepâncias e contradições entre diferentes fontes ou períodos.

Comparabilidade: Permite que as estatísticas sejam comparadas ao longo do tempo, entre diferentes regiões ou grupos, facilitando análises e interpretações robustas.

Clareza: Envolve a apresentação transparente e compreensível das estatísticas, garantindo que os utilizadores possam entender facilmente os resultados e as metodologias utilizadas.

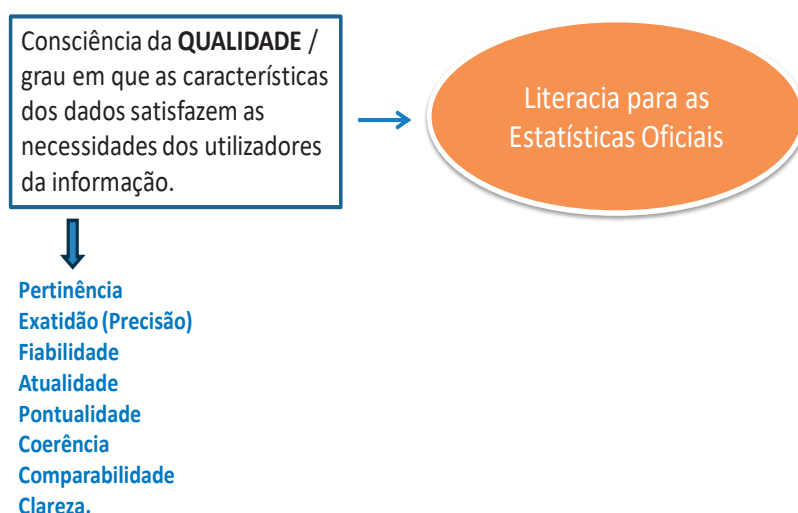


Figura 2. Competências de literacia estatística para a qualidade

Essas dimensões trabalham sinergicamente para construir a integridade e confiabilidade das estatísticas oficiais, tornando-as ferramentas valiosas para a tomada de decisões informadas, planejamento e avaliação em diversos setores da sociedade. A literacia estatística deve, assim, incorporar esses princípios, pois a compreensão da qualidade é vital para uma jornada bem-sucedida na interpretação estatística. A esses seis elementos, acrescentamos as dimensões da qualidade referidas anteriormente.

É fundamental elevar a compreensão sobre a qualidade em estatísticas oficiais, avaliando em que medida os dados atendem às necessidades dos utilizadores. Isso exige investimentos na promoção da literacia para a qualidade, uma tarefa que obriga a uma abordagem abrangente e colaborativa, envolvendo diversas atividades e setores da sociedade, tais como promover a literacia para a qualidade em estatísticas oficiais, envolvendo diferentes setores da sociedade. Em primeiro lugar, a formação dos media é vital, pois os jornalistas desempenham um papel crucial na difusão da informação estatística. Além disso, a formação nas comunidades educativas, desde o ensino básico até o secundário e até o ensino superior, é fundamental para construir uma base sólida de literacia estatística, incorporando conceitos acessíveis no currículo e destacando a necessidade de qualidade nos dados.

Para esse efeito, cursos específicos sobre literacia estatística, oferecidos em instituições de ensino superior ou plataformas online, representam uma estratégia eficaz para abordar as necessidades de diversos públicos. Envolvendo as comunidades em atividades práticas com dados estatísticos oficiais, como workshops em ciência e análise de dados, proporciona uma experiência tangível e promove a literacia para a qualidade.

Programas educacionais especializados, como o EMOS (*European Master in Official Statistics*), desempenham um papel crucial na formação de especialistas qualificados em estatísticas oficiais, contribuindo para o aprofundamento do conhecimento técnico. Iniciativas online, como o ALEA e a Explorística, o EDUCA em Espanha, ou a área *Statistics Explained* (um guia para as estatísticas europeias do Eurostat que apresenta temas estatísticos de uma forma facilmente compreensível), oferecem recursos acessíveis, como tutoriais e ferramentas interativas, atingindo um amplo público.

Notas finais

Com a grande quantidade de dados disponíveis, a literacia para compreender a qualidade dos dados tornou-se um desafio persistente nas estatísticas oficiais, sendo a qualidade dos dados estatísticos um elemento essencial a ser considerado, medido pelo grau em que as características das estatísticas atendem aos requisitos dos utilizadores da informação estatística.

Sabemos bem que a qualidade estatística mensurável refere-se à precisão, consistência e conformidade com padrões na recolha, análise e apresentação de dados. Isso inclui a minimização de erros, o uso de métodos robustos e a adesão a códigos de conduta, como o "European Statistics Code of Practice." A qualidade percebida pelos utilizadores, por sua vez, relaciona-se com aquilo que os utilizadores valorizam, como sendo dados pertinentes às suas necessidades, consistentes ao longo do tempo e comunicados de forma clara. Ambas as dimensões são interdependentes, com práticas estatísticas sólidas contribuindo para a confiança dos utilizadores e, conseqüentemente, para a qualidade percebida.

Referências

- Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.
- Eurostat. (2017). Código de conduta das estatísticas europeias. Retrieved from: <https://ec.europa.eu/eurostat/web/products-catalogues/-/european-statistics-code-of-practice-revised-edition-2017>
- Gal, I. (2002). Adults' Statistical Literacy: Meanings, Components, Responsibilities. *International Statistical Review*, 70, 1-51. <https://doi.org/10.1111/j.1751-5823.2002.tb00336.x>

- Gal, I. (2024). Adult education in mathematics and numeracy: a scoping review of recent research. *ZDM Mathematics Education*. <https://doi.org/10.1007/s11858-024-01549-z>
- Gal, I., & Ograjenšek, I. (2017). Official Statistics and Statistics Education: Bridging the Gap. *Journal of Official Statistics*, 33(1), 79-100. <https://doi.org/10.1515/JOS-2017-0005>
- Gould, R. (2017). Data Literacy Is Statistical Literacy. *Statistics Education Research Journal*, 16(1), 22–25.
- Reading, C. (2002). Profile for statistical understanding. *ICOTS 6, International Conference on Teaching of Statistics*, Cape Town, South Africa.
- Smith, G. (1998). Learning Statistics by Doing Statistics. *Journal of Statistics Education*, 6(3). <https://doi.org/10.1080/10691898.1998.11910623>
- Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning: The REALI instrument. *Statistics Education Research Journal*, 17(2), 141–160. <https://doi.org/10.52041/serj.v17i2.163>
- Sharma, S. (2017). Definitions and models of statistical literacy: a literature review. *Open Review of Educational Research*, 4(1), 118-133.



Explorando a lei de formação das estrelas por seleção de variáveis

Ana Helena Tavares, ahtavares@ua.pt

*Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro
Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA)*

A imprevisibilidade do convite para participar neste Boletim SPE, sob o tema *Rising Stars*, foi equivalente à dificuldade para decidir qual o tema a abordar. Nas simpáticas palavras do seu Editor, Professor Fernando Rosado, a ideia seria partilhar com os prezados leitores “as rotas científicas nas quais navegam os jovens investigadores”. E, na senda da ancestral tradição das descobertas, eis-me a misturar o céu e as estrelas, as rotas e os mares. A análise estatística pode fornecer conhecimentos valiosos sobre padrões, relações e fenómenos que ocorrem tanto no espaço longínquo quanto no fundo do mar. E, neste percurso, o meu contributo caminha, paulatinamente, para contextos de grandes volumes de dados, muitas variáveis, muito ruído, e outros problemas que dificultam a extração da informação.

“Explorando a lei de formação das estrelas por seleção de variáveis” (tradução livre), é o título de um estudo publicado em 2019 cujo objetivo principal é o de identificar preditores significativos da formação de estrelas em galáxias, beneficiando de uma valiosa base de dados relativa a 126 galáxias [1]. Neste contexto, os preditores são propriedades físicas das estrelas (e.g. conteúdo de gás, densidade da superfície estelar, metalicidade) e a variável de resposta é a taxa de formação estelar. Ora, mesmo um leigo em Física e Astrofísica, desconfiará que a análise a efetuar incide sobre um espaço de alta dimensão e com muitos parâmetros físicos correlacionados entre si. E a generalidade dos estatistas reconhecerá que a seleção de variáveis neste contexto não é tarefa fácil!

Para obter informações sobre sistemas complexos, como a formação das estrelas, o metabolismo ou a regulação genética, são necessárias considerações multivariadas, sendo a geração massiva de dados de extrema importância para a recolha de informação relativa a um grande número de variáveis. A desvantagem inerente é, obviamente, o risco de incluir variáveis irrelevantes nos modelos estatísticos. De modo a minimizar a influência de tais variáveis ruidosas é, geralmente, necessária alguma redução dos dados, seja através de métodos de projeção, seleção de variáveis ou uma combinação de ambos.

Os procedimentos mais populares para a seleção de variáveis em problemas de modelação são os testes de hipóteses. A técnica de regressão passo a passo permite o ajuste a um modelo de regressão efetuando, em cada passo, a remoção (método regressivo) ou adição (método progressivo) de uma única variável com base em algum critério estatístico. Normalmente, assume a forma de uma sequência de testes de hipóteses, mas poderá basear-se também em medidas de informação, como o critério de informação de Akaike, Cp de Mallows ou Kullback-Leibler. Estes permitem selecionar o modelo mais informativo a partir de um conjunto plausível de modelos. Ao selecionar o “melhor” modelo está, portanto, a selecionar a melhor combinação de variáveis predictoras. A regressão passo a passo é uma forma simples e direta de fazer a seleção. Contudo, em diversas situações, como na existência de variáveis de ruído ou de

colinearidade, não garante uma seleção consistente. Em dados de alta dimensionalidade, as fragilidades da regressão passo a passo são exacerbadas e requerem um tempo de computação intensivo [2].

Outros métodos amplamente utilizados na seleção de variáveis são, por exemplo, a Regressão por Componentes Principais e a Regressão por Mínimos Quadrados Parciais (PLS). Embora estes métodos tenham bom desempenho em situações de multicolinearidade, é sabido que as estimativas dos modelos são afetadas pela presença de um grande número de variáveis irrelevantes. Apesar destes métodos não requererem a remoção de preditores irrelevantes, a imposição de esparsidade favorece a seleção de variáveis em contextos de alta dimensionalidade [3, 4].

Na análise de dados moderna, a seleção esparsa de variáveis torna-se inevitável quando o número de variáveis preditoras é muito elevado. Uma forma de ultrapassar o problema da dimensionalidade na seleção de variáveis é a aplicação de técnicas de regularização, que ajustam o modelo com todos os preditores forçando alguns dos coeficientes estimados a serem iguais ou próximos de zero. Por exemplo, os modelos de regressão penalizada foram já utilizados em estudos de associação genética, com objetivo de identificar variantes genéticas associadas a características específicas, como doenças, traços físicos ou comportamentais, em genomas completos [5, 6]. Estes estudos são fundamentais para aumentar o conhecimento sobre a base genética das doenças e podem, eventualmente, levar ao desenvolvimento de novas abordagens de diagnóstico, prevenção e tratamento.

Voltando ao artigo “Explorando a lei de formação das estrelas por seleção de variáveis”, o problema da dimensionalidade e multicolinearidade é ultrapassado, precisamente, por aplicação de uma técnica de regularização por penalização: LASSO.

O *Least Absolute Shrinkage and Selection Operator* (LASSO) é, talvez, o mais conhecido dos modelos de regularização. Este permite uma simplificação explícita do modelo uma vez que muitos coeficientes são forçados a ser iguais a zero. No entanto, apresenta algumas desvantagens, como por exemplo, não permitir a seleção de mais variáveis do que a dimensão da amostra. Além disso, na presença de variáveis altamente correlacionadas, seleciona arbitrariamente uma ou alguma das variáveis e ignora as restantes, o que conduz a seleções instáveis [7].

Outra abordagem para a seleção de variáveis nestes contextos de alta dimensionalidade é a avaliação da quantidade de informação de cada variável por estimação da entropia normalizada, um conceito proveniente da Teoria da Informação e da Infométrica. A utilização da entropia normalizada num procedimento de seleção de variáveis foi proposta pela primeira vez em 1996 [8, 9]. Este procedimento apresenta diversas vantagens, como a simplicidade de aplicação, a utilização de estimadores de contração e a possibilidade de implementação em modelos mal-postos. Concernente à implementação desta abordagem, a literatura é exígua, possivelmente devido a dificuldades no uso de informações prévias e a falta de recursos computacionais adequados.

A investigação na área da seleção de variáveis está ainda em desenvolvimento, principalmente devido à disponibilidade de uma quantidade de dados sem precedente, com todas as especificidades que cada estrutura de dados acarreta. Em dados de alta dimensão, a instabilidade dos métodos de seleção é, quase sempre, uma realidade. Numa perspetiva de desenvolvimento da área há, certamente, margem para realizar investigação inovadora.

Como podemos contribuir?

O estudo de novas metodologias de seleção de variáveis e modelação, no contexto de um grande número de variáveis, promovendo a estabilidade na seleção das variáveis e a precisão na estimação da magnitude dos coeficientes dos modelos preditivos é, certamente, uma via que deve ser desenvolvida. Por exemplo, Rodrigues *et al.* [10] avaliam a estabilidade de um novo procedimento que identifica a associação entre

preditores genéticos relevantes e Doença de Alzheimer numa estrutura onde o número de variáveis excede, por muito, o tamanho da amostra.

É óbvio o interesse no desenvolvimento de procedimentos, adaptados a grandes volumes de dados, que sejam estáveis na identificação de variáveis preditoras influentes e na estimativa do seu efeito explicativo. Combinar procedimentos estatísticos de regularização e de máxima entropia poderá, num futuro próximo, transformar alguns procedimentos tradicionais de inferência estatística, contribuindo para uma discussão unificadora de procedimentos de inferência [11]. A este propósito, é importante realçar que o desenvolvimento de procedimentos computacionais na área de inferência com máxima entropia, suprirão uma ausência atualmente sentida pela comunidade científica.

A imensidão é apaixonante. Seja do mar, seja do espaço, seja dos dados. E como tal, é tão cativante procurarmos hoje soluções para estes problemas, como terá sido partir à descoberta do mundo com um astrolábio e um quadrante, a confiar nas estrelas e no espaço. Quem quer partir nesta epopeia?

Agradecimentos

Este trabalho é suportado pelo Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), através da Fundação para a Ciência e a Tecnologia (FCT), com referências UIDB/04106/2020 e UIDP/04106/2020.

Referências

- [1] Dey, B., Rosolowsky, E. & García-Benito, R. (2019). The EDGE-CALIFA survey: exploring the star formation law through variable selection. *Monthly Notices of the Royal Astronomical Society*, 488(2), 1926-1940.
- [2] Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, 5(1), 32.
- [3] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis, *Journal of Computational and Graphical Statistics* 15(2): 265-286.
- [4] Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*
- [5] Cherlin, S., Howey, R. A. J., and Cordell, H. J. (2018), “Using penalized regression to predict phenotype from SNP data,” *BMC Proceedings*, 12, 0–5.
- [6] Waldmann P, Gredler GMB, Fürst C, et al. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*. 2013;4.
- [7] Tibshirani, R. (1996), “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
- [8] Golan, A., Judge, G., and Miller, D. (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley & Sons, Chichester.
- [9] Golan, A. (2001). A simultaneous estimation and variable selection rule. *Journal of Econometrics*, 101: 165-193.
- [10] Rodrigues, L., Cabral, J., Tavares, A., Enes, V., Pinheiro, M., Moura, G., & Afreixo, V. (2022). Assessment of the stability of a procedure for variables selection in high dimensionality data: an application to genomic data - Alzheimer’s Disease. *Journal of Statistics on Health Decision*, 4(1), 48-49.
- [11] Caticha, A. (2021). Entropy, Information, and the Updating of Probabilities. *Entropy*, 23(7), 895



Misturas de modelos de regressão com efeitos aleatórios

Luísa Novais, *luisa_novais92@hotmail.com*

Departamento de Matemática, Universidade do Minho

Em outubro de 2023 tive a oportunidade de apresentar uma comunicação intitulada “A utilização de funções de penalização na seleção de variáveis em misturas de modelos de regressão com efeitos aleatórios” na sessão temática *Rising Stars*, organizada pela SPE/CLAD no XXVI Congresso da Sociedade Portuguesa de Estatística.

A seleção de variáveis possui um papel fundamental no estudo de modelos de mistura, envolvendo a procura de um modelo o mais simples possível, mas que descreva adequadamente os dados observados. No entanto, os métodos de seleção de variáveis clássicos requerem intensa computação, em particular em modelos de mistura, mesmo na modelação com um número moderado de variáveis explicativas. Como tal, dada a grande complexidade da generalidade das bases de dados atuais, surgiu a necessidade de desenvolver novas metodologias mais eficientes e que permitam acomodar a complexidade computacional existente, como é o caso dos métodos baseados em funções de penalização.

Assim, na realização do trabalho apresentado investigaram-se diferentes métodos de seleção de variáveis baseados em funções de penalização que atuam sobre os coeficientes das variáveis, em particular os métodos *Least Absolute Shrinkage and Selection Operator* (LASSO), *Adaptive Least Absolute Shrinkage and Selection Operator* (ALASSO), HARD e *Smoothly Clipped Absolute Deviation* (SCAD), comparando-se o seu desempenho na identificação do subconjunto de variáveis mais relevantes em misturas de modelos de regressão com efeitos aleatórios, recorrendo aos algoritmos *Expectation-Maximization* (EM) e *Classification Expectation-Maximization* (CEM).

Tal como se esperava, a seleção de variáveis em modelos de mistura é um problema complexo e os diferentes cenários experimentais influenciaram o desempenho dos diferentes algoritmos e das distintas funções de penalização aplicadas. Contudo, como principais conclusões a retirar, verificou-se que o algoritmo CEM demonstrou a sua superioridade relativamente ao algoritmo EM e, no que concerne aos quatro métodos baseados em funções de penalização, a nível de seleção de variáveis constatou-se a superioridade do método ALASSO, enquanto o método HARD obteve o pior desempenho. Em resumo, a utilização do algoritmo CEM é aconselhada em qualquer cenário, em detrimento do algoritmo EM, dada a sua superioridade para todas as medidas de desempenho estudadas. Já a nível de funções de penalização, recomenda-se a seleção de variáveis recorrendo ao método ALASSO, dados os bons resultados apresentados para qualquer dos cenários em estudo.

O trabalho apresentado no XXVI Congresso da Sociedade Portuguesa de Estatística nada mais foi do que uma pequena parte do trabalho desenvolvido ao longo da minha tese de Doutoramento, intitulada “Misturas de modelos de regressão com efeitos aleatórios”, a qual defendi em julho de 2022 na Universidade do Minho.

No trabalho de Doutoramento dei continuidade ao trabalho iniciado na dissertação de mestrado, de título “Misturas de modelos lineares mistos”, em que se abordaram distintas problemáticas relacionadas com este tipo de modelos de mistura. Deste modo, o principal objetivo da minha tese de Doutoramento consistiu em fornecer uma visão geral dos modelos de mistura nas suas diversas vertentes, com o intuito de demonstrar todo o seu potencial estatístico e a sua aplicabilidade às mais diversas áreas.

Como é sabido, nas últimas décadas a análise de regressão tem assumido particular relevância nas mais diversas áreas científicas, tornando-se numa das técnicas estatísticas mais utilizadas.

Paralelamente, a crescente evolução computacional, aliada ao aparecimento de conjuntos de dados que possuem grande heterogeneidade, desencadeou a utilização de modelos de mistura, em particular de modelos de mistura de regressões, possibilitando a sua utilização em situações em que a estimação de um único modelo de regressão não é adequada.

Por outro lado, quando as observações do conjunto de dados se encontram agrupadas de acordo com um determinado fator verifica-se a necessidade de introdução de efeitos aleatórios. Por tudo isto, surgiu a necessidade do desenvolvimento de modelos que incorporem a heterogeneidade não observada da população e que expliquem, simultaneamente, as correlações entre as observações provenientes do mesmo indivíduo, como é o caso das misturas de modelos de regressão com efeitos aleatórios.

Assim, ao longo da tese aprofundaram-se as problemáticas relacionadas com a determinação do número de componentes e com a estimação dos parâmetros, abordando-se igualmente a complexa questão da seleção de variáveis em modelos de mistura. No que diz respeito à determinação do número de componentes, desenvolveu-se com maior profundidade o trabalho realizado na dissertação do mestrado, através da implementação de novos critérios de informação e de classificação na estimação do número de componentes de um modelo de mistura e da implementação de métodos robustos. Nesse sentido, dado que os critérios de informação e de classificação são bastante sensíveis a observações *outlier*, procedeu-se ao desenvolvimento dos mesmos critérios numa versão robusta, comparando-se o desempenho de ambas as versões de cada critério recorrendo a modelos de mistura contaminados.

Relativamente à estimação dos parâmetros, através da aplicação do método da máxima verosimilhança, o algoritmo *Expectation-Maximization* (EM) é o algoritmo iterativo mais utilizado na determinação de estimativas de máxima verosimilhança em problemas de dados incompletos. Contudo, um dos principais inconvenientes deste algoritmo é a sua elevada sensibilidade ao conjunto de valores iniciais para as estimativas dos parâmetros, dado que o algoritmo EM apenas garante a seleção de um máximo local e não global, caso exista. Por este motivo, começou-se por estudar o problema da inicialização do algoritmo EM através da comparação de duas estratégias de inicialização distintas, a inicialização através dos verdadeiros valores dos parâmetros e uma inicialização recorrendo a valores aleatórios. Finalmente, investigou-se o desempenho do algoritmo EM e de duas das suas versões modificadas, o algoritmo *Classification Expectation-Maximization* (CEM) e o algoritmo *Stochastic Expectation-Maximization* (SEM) na determinação das estimativas de máxima verosimilhança dos parâmetros.

Por último, os avanços tecnológicos dos últimos anos provocaram a existência de conjuntos de dados contendo um elevado número de observações e/ou um elevado número de variáveis, pelo que a seleção de variáveis explicativas assume um papel fulcral no estudo de modelos de mistura. Tal como já referido, para a estudar investigaram-se diferentes métodos de seleção de variáveis baseadas em funções de penalização que atuam sobre os coeficientes das variáveis, em particular os métodos LASSO, ALASSO, HARD e SCAD, comparando-se o seu desempenho na identificação do subconjunto de variáveis explicativas mais relevantes recorrendo aos algoritmos EM e CEM.

A dissertação de mestrado e a tese de Doutoramento encontram-se disponíveis no repositório da Universidade do Minho, bem como os quatro artigos que resumem as principais conclusões deste trabalho. Resumir o trabalho de anos em algumas linhas não é tarefa fácil, mas mais do que olhar para o passado, o importante é pensar no futuro. Assim, apesar da tese estar concluída, há ainda muito por investigar nesta área, pelo que existem várias temáticas sobre as quais gostaria de trabalhar no futuro.

Nomeadamente, a temática da seleção de variáveis em modelos de mistura tem ainda muito por descobrir, especialmente no que diz respeito a métodos baseados em funções de penalização, dado tratar-se de um tópico relativamente recente para este tipo de modelos e onde ainda há muito por desenvolver. Em particular, será interessante ampliar a abordagem efetuada à seleção de variáveis em misturas de modelos lineares mistos, uma vez que se analisaram métodos de seleção de variáveis baseados em funções de penalização para a seleção de um subconjunto de variáveis associadas aos efeitos fixos, pelo que será de interesse alargar os métodos propostos também à seleção de variáveis associadas aos efeitos aleatórios. Paralelamente, pretendo igualmente ampliar os métodos propostos à seleção de variáveis categóricas, quer estejam associadas aos efeitos fixos, quer estejam associadas aos efeitos aleatórios.

Por outro lado, outra contribuição futura interessante consiste em aplicar as mesmas metodologias a outras famílias de misturas de modelos de regressão com efeitos aleatórios, como é o caso das misturas de modelos lineares generalizados mistos e, simultaneamente, tendo em consideração o exponencial avanço tecnológico da última década e a consequente expansão do *Big Data*, será de interesse aplicar as diversas metodologias a conjuntos de dados de elevadas dimensões.

Em suma, estas e outras contribuições fazem parte de uma (extensa) lista de ideias a trabalhar futuramente na área dos modelos de mistura. Haja tempo, porque vontade não falta!



Dados simbólicos – Os dados do futuro!

Sónia Dias, sdias@estg.ipvc.pt

ESTG- Instituto Politécnico de Viana do Castelo
e LIAAD INESC-TEC

1. INTRODUÇÃO

A evolução tecnológica e científica trouxe consigo um aumento exponencial de dados muitas vezes gerados em tempo real. Estes grandes conjuntos de dados precisam de ser tratados e analisados para que possam transmitir informação útil e gerar valor para as mais diversas áreas: económica, científica, política. Consequentemente, é necessário inovar e desenvolver novos métodos adaptados a grandes volumes de dados. Para além dos métodos de mineração de dados, têm vindo a emergir abordagens estatísticas alternativas. A Análise de Dados Simbólicos (SDA) proposta em 1988 por Edwin Diday (Diday, 1988) é uma das abordagens atualmente em expansão, onde os dados clássicos e métodos estatísticos são generalizados para dados de maior complexidade.

2. DADOS CLÁSSICOS vs DADOS SIMBÓLICOS

As tabelas de dados onde as células contêm um único valor quantitativo ou categórico são desadequadas quando pretendemos estudar grandes volumes de dados. São necessárias tabelas de dados mais complexas onde cada célula expressa a variabilidade de cada unidade observada. As unidades observadas nas tabelas de dados simbólicas surgem por vezes da agregação de observações recolhidas ao nível dos microdados. A agregação e análise dos dados a um nível superior (macro) é que permite muitas vezes transmitir informação útil e relevante.

De acordo com Arroyo & Maté (2009) a agregação dos dados pode ser temporal ou contemporânea. Temos uma agregação contemporânea quando os registos são recolhidos no mesmo momento temporal ou o momento temporal não é relevante. Nestes casos o estudo pretende analisar as unidades (de 2ª ordem) formadas por classes de indivíduos (conjuntos de unidades de 1ª ordem) agrupados de acordo com características específicas.

Ao nível das estatísticas oficiais, por exemplo nos censos, é recolhida informação dos cidadãos – unidades de 1ª ordem, mas a análise que se pretende fazer *à posteriori* é mais global, a nível dos municípios/concelhos ou NUTS - unidades de 2ª ordem.

Tabela 1: Dados (simulados) dos Censos – Microdados

ID Cidadão	Concelho	Nº elementos do agregado familiar	Tempo deslocação diária casa – trabalho (minutos)	Estado Civil
AR1234	Braga	3	5 minutos	Casado
CV4537	Lisboa	2	10 minutos	Divorciado
BG6783	Coimbra	4	15 minutos	Casado
HI8965	Lisboa	1	30 minutos	Solteiro
KM3957	Porto	5	15 minutos	Casado

A agregação será temporal se o tempo é um critério de agregação e os registos são agrupados por uma unidade de tempo, por dia/mês/ano, mas a ordenação temporal não é relevante. Neste caso as unidades em estudo são caracterizadas por conjuntos de valores recolhidos num dado período de tempo. Nas estações meteorológicas dispersas pelo país, a informação recolhida das variáveis meteorológicas é feita a cada hora mas no geral o interesse não é o registo a cada hora, mas sim o conjunto dos valores recolhido para um dado dia/mês/ano. Neste caso, as unidades de 1ª e 2ª ordem são as estações meteorológicas.

Tabela 2: Dados de variáveis meteorológicas 8 Março 2024 – Microdados (IPMA, 2024)

Estação Meteorológica	Temperatura (° C)							Precipitação (mm)						
	08/03 0h	08/03 1h	08/03 2h	...	08/03 14h	08/03 15h	...	08/03 0h	08/03 1h	08/03 2h	...	08/03 14h	08/03 15h	...
Bragança	3.0	3.0	2.9		6.9	7.0		0	0.1	0.6		0.5	0	
Cabo Carvoeiro	10.8	11.9	12.5		13	12.9		2.2	0.9	0.6		0.8	1.3	
Alcácer do Sal	10.7	9.7	9.0		15.4	15		0.5	1.1	1.0		0	0.1	

Em termos clássicos, após agregação, a informação associada a cada unidade de 2ª ordem (indivíduo ou classe de indivíduos) é um valor único, que tipicamente é a média, moda, máximo/mínimo, total acumulado. Contudo, ao fazer esta análise a variabilidade da informação associada a cada unidade macro (unidade de 2ª ordem) não é considerada.

Nas Figuras 1 e 2 são apresentados dois exemplos de abordagens clássicas obtidas a partir da agregação dos microdados, considerando os contextos apresentados nas Tabelas 1 e 2.

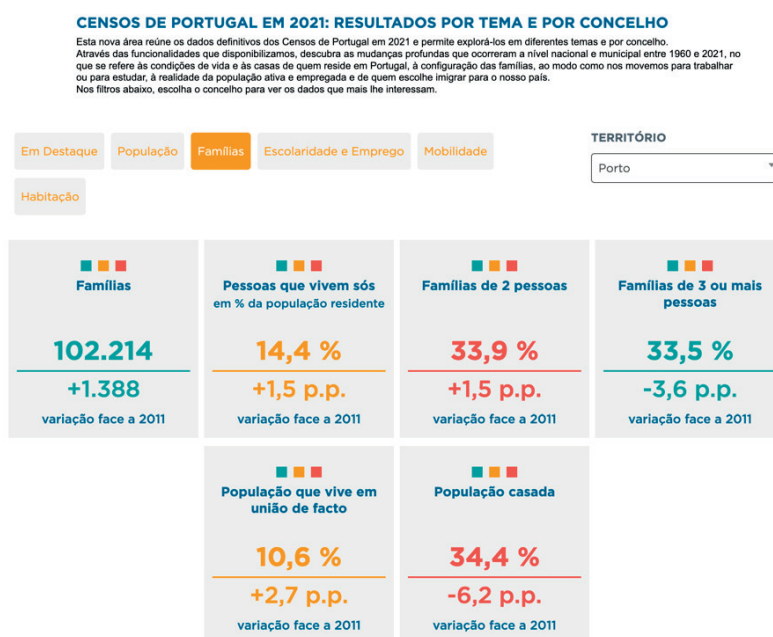


Figura 1: Apresentação clássica – Resultados dos Censos 2021 (PORDATA, 2024)



Figura 2: Apresentação clássica – Informação meteorológica, 7 Março 2023, na Estação de Bragança (IPMA, 2024)

Na SDA trabalha-se sobre tabelas dados simbólicas isto é, para cada unidade e variável, já não se associada apenas um valor/categoria mas sim um intervalo, um histograma, um conjunto de valores ou categorias (Bock & Diday, 2000). Mantendo os contextos apresentados anteriormente nas Tabelas 1 e 2, nas Tabelas 3 e 4 são apresentados dois exemplos de tabelas de dados simbólicos.

Tabela 3: Dados simbólicos (macrodados) - Resultados dos Censos por concelho (dados simulados)

Concelho	Nº elementos agregado familiar	Tempo deslocação diária casa – trabalho (minutos)	Estado Civil
Braga	[1,4]	{[0,16[, 0.4; [16,31[, 0.2; [31,60[, 0.3; [60,90], 0.1}	{Solteiros, 45%; Casados, 42%; Viúvos, 5.2%; Divorciados, 6.9%}
Coimbra	[1,6]	{[0,16[, 0.6; [16,31[, 0.2; [31,60], 0.2}	{Solteiros, 42.5%; Casados, 41%; Viúvos, 7.2%; Divorciados, 9.7%}
Lisboa	[1,4]	{[0,16[, 0.1; [16,31[, 0.2; [31,60[, 0.4; [60,90], 0.25; >90, 0.05}	{Solteiros, 51.6%; Casados, 38.1%; Viúvos, 7.5%; Divorciados, 9.2%}
Porto	[1,5]	{[0,16[, 0.1; [16,31[, 0.25; [31,60[, 0.35; [60,90], 0.2; >90, 0.1}	{Solteiros, 47.5%; Casados, 34.4%; Viúvos, 8.1%; Divorciados, 10.1%}

Tabela 4: Dados simbólicos (macrodados) – Variáveis atmosféricas no dia 8 Março 2023, por estação meteorológica

Estação Meteorológica	Temperatura (° C)	Precipitação (mm)
Bragança	{[2,5[, 0.4; [5,8[, 0.5; [8,10], 0.1}	[0, 0.8]
Cabo Carvoeiro	{[8,10[, 0.1; [10,12[, 0.7; [12,14], 0.2}	[0.5, 2.2]
Alcácer do Sal	{[8,10[, 0.25; [10,12[, 0.5; [12,14[, 0.15; [14,16], 0.1}	[0, 1.2]

3. VARIÁVEIS SIMBÓLICAS – CONCEITO E CLASSIFICAÇÃO

A SDA generaliza o conceito de variável clássica de modo que a cada unidade de 2ª ordem (elemento de um conjunto E) pode corresponder um conjunto de números reais, um conjunto de categorias ou intervalos, uma distribuição de frequências/probabilidades (elementos de um conjunto \mathbb{B}). A definição de variável simbólica assim como a suas classificações (Bock & Diday, 2000; Billard & Diday, 2007; Brito, 2014) são apresentadas no esquema da Figura 3.

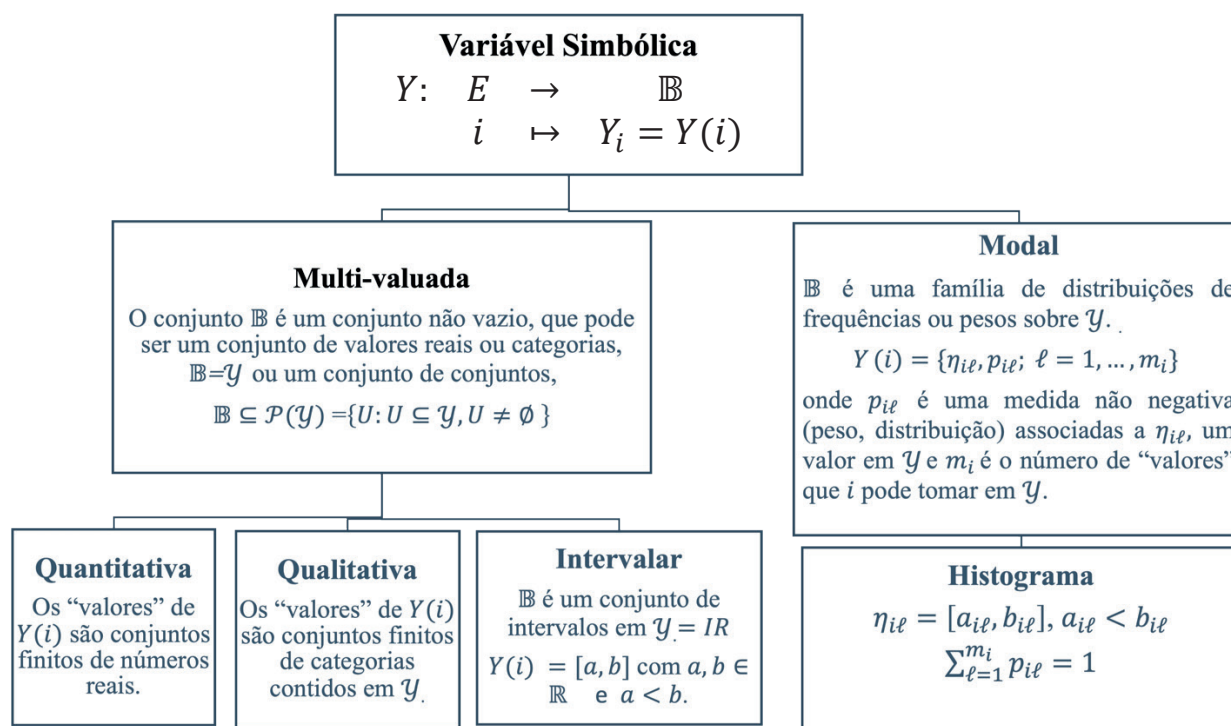


Figura 3: Definição e classificação das variáveis simbólicas.

Nos exemplos apresentados anteriormente, Tabelas 3 e 4, as variáveis *Nº elementos agregado familiar* e *Precipitação* são intervalares; o *Tempo deslocação diária casa – trabalho* e *Temperatura* são variáveis histograma e o *Estado Civil* é uma variável modal.

Inicialmente o desenvolvimento de conceitos e métodos apresentou uma maior incidência nas variáveis intervalares, mas mais recentemente as variáveis quantitativas distribucionais, nas quais se incluem as variáveis histograma, têm tido uma crescente evolução.

Definição 1 (Bock & Diday, 2000): *Uma variável Y é designada de variável histograma se a cada observação i dessa variável está associado um histograma*

$$Y_i = \left\{ [l_{Y_{i1}}, u_{Y_{i1}}], p_{i1}; [l_{Y_{i2}}, u_{Y_{i2}}], p_{i2}; \dots; [l_{Y_{im_i}}, u_{Y_{im_i}}], p_{im_i} \right\}$$

onde $p_{i\ell}$ é o peso associado ao intervalo $[l_{Y_{i\ell}}, u_{Y_{i\ell}}]$ com $\ell \in \{1; 2; \dots; m_i\}$; sendo m_i o número de subintervalos relativo à observação i ; $l_{Y_{i\ell}} \leq u_{Y_{i\ell}}$, $u_{Y_{i\ell}} \leq l_{Y_{i(\ell+1)}}$ e $\sum_{\ell=1}^{m_i} p_{i\ell} = 1$.

Assume-se que no interior de cada subintervalo $[l_{Y_{i\ell}}, u_{Y_{i\ell}}]$ os valores da variável Y para cada observação i estão uniformemente distribuídos (Billard & Diday, 2003).

De salientar que as variáveis intervalares podem ser consideradas um caso particular das variáveis histograma. Isso acontece quando para todas as observações i da variável, temos histogramas com apenas um subintervalo, $m_i = 1$, de peso associado igual a 1, $p_i = 1$. Assim, os métodos desenvolvidos para variáveis histograma, apesar de mais complexos, devem poder ser particularizados para variáveis intervalares.

De entre as dificuldades associada a trabalhar com histogramas destacamos: 1) o facto de em diferentes observações da mesma variável histograma, o número de subintervalos poder ser diferente; 2) a complexidade de efetuar operações usando a aritmética de histogramas.

4. VARIÁVEIS HISTOGRAMA

4.1. ESTATÍSTICAS DESCRITIVAS

Ao nível da estatística descritiva univariada e multivariada, estão definidos para variáveis histograma os principais conceitos. No caso univariado estão definidos os conceitos de frequência observada e relativa; função densidade empírica; função de distribuição empírica (Bock & Diday, 2000; Billard & Diday, 2003); medidas de localização como a média; medidas de dispersão como variância (mais do que uma definição) e desvio padrão (Billard & Diday, 2002; Billard & Diday, 2003; Dias, 2014). No caso multivariado estão definidos os conceitos de frequência observada e relativa; função densidade empírica e distribuição conjuntas; medidas de dependência como covariância (mais do que uma definição) e coeficiente de correlação (Billard & Diday, 2002; Billard & Diday, 2003; Billard & Diday, 2007; Arroyo, 2008).

De referir que algumas destas definições para variáveis histograma podem ser facilmente obtidas por aplicação dos conceitos da estatística clássica aos pontos médios dos subintervalos, devidamente ponderados pelos pesos associados a cada subintervalo. Note-se que, apesar das observações das variáveis serem histogramas, estes conceitos são definidos à custa de valores reais associados aos histogramas. Para exemplificar esta abordagem, na definição que se segue é apresentado o conceito de valor médio de uma variável histograma.

Definição 2 (Billard & Diday, 2003): *Consideremos a variável histograma Y (de acordo com a definição 1). O valor médio desta variável é dado por:*

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n \left(\sum_{\ell=1}^{m_i} \left(\frac{l_{Y_{i\ell}} + u_{Y_{i\ell}}}{2} \right) p_{i\ell} \right)$$

Posteriormente, Irpino e Verde (Irpino & Verde, 2006; Irpino & Verde, 2015A), apresentaram conceitos alternativos para as estatísticas descritivas. Por exemplo a média de uma variável histograma é o histograma médio tal que o critério de distância é minimizado. A variância é o escalar calculado pela média dos quadrados das distâncias entre as observações e a distribuição média. Para o desenvolvimento do seu trabalho, estes autores precisaram de encontrar uma forma mais simples de representar os usuais histogramas e de encontrar uma distância adequada para medir a “semelhança” entre distribuições.

4.2. REPRESENTAÇÕES DOS HISTOGRAMAS

Para resolver a dificuldade em trabalhar com histogramas e com a sua aritmética, Irpino e Verde (Irpino & Verde, 2006) propuseram como representação alternativa ao usual histograma a função quantil - inversa da função distribuição acumulada.

Definição 3: *As observações de uma variável histograma Y com m subintervalos, podem ser representada por:*

Histogramas

$$Y_i = \{[l_{i1}, u_{i1}[, p_{i1}; [l_{i2}, u_{i2}[, p_{i2}; \dots; [l_{im}, u_{im}[, p_{im}\}$$

onde cada subintervalo ℓ é definido à custa dos extremos inferior $l_{i\ell}$ e superior $u_{i\ell}$, $l_{Y_{i\ell}} \leq u_{Y_{i\ell}}$, $u_{Y_{i\ell}} \leq l_{Y_{i(\ell+1)}}$ e $\sum_{\ell=1}^{m_i} p_{i\ell} = 1$.

Ou alternativamente,

$$Y_i = \{[c_{i1} - r_{i1}, c_{i1} + r_{i1}[, p_{i1}; [c_{i2} - r_{i2}, c_{i2} + r_{i2}[, p_{i2}; \dots; [c_{im} - r_{im}, c_{im} + r_{im}[, p_{im}\}$$

onde cada subintervalo ℓ é definido à custa do centro $c_{i\ell} = \frac{u_{i\ell} + l_{i\ell}}{2}$ e do raio $r_{i\ell} = \frac{u_{i\ell} - l_{i\ell}}{2}$.

Funções distribuição acumulada

Assumindo a uniformidade nos subintervalos a função distribuição acumulada $F_i(x)$ fica definida por:

$$F_i(x) = \begin{cases} 0 & \text{se } x < l_{i1} \\ \frac{x - l_{i1}}{u_{i1} - l_{i1}} p_{i1} & \text{se } l_{i1} \leq x < u_{i1} \\ p_{i1} + \frac{x - l_{i2}}{u_{i2} - l_{i2}} p_{i2} & \text{se } l_{i2} \leq x < u_{i2} \\ \vdots & \\ 1 & \text{se } x \geq u_{im} \end{cases}$$

Funções Quantil (Irpino & Verde, 2006)

A função quantil $\Psi_i(t)$, inversa da função distribuição acumulada $F_i(x)$ onde é assumida a uniformidade nos subintervalos, fica definida por:

$$\Psi_i(t) = \begin{cases} l_{i1} + \frac{t}{w_{i1}}(u_{i1} - l_{i1}), & 0 \leq t < w_{i1} \\ l_{i2} + \frac{t - w_{i1}}{w_{i2} - w_{i1}}(u_{i2} - l_{i2}), & w_{i1} \leq t < w_{i2} \\ \vdots & \\ l_{im} + \frac{t - w_{i(m-1)}}{1 - w_{i(m-1)}}(u_{im} - l_{im}), & w_{i(m-1)} \leq t \leq 1 \end{cases} \quad \text{ou } \Psi_i(t) = \begin{cases} c_{i1} + \left(\frac{2t}{w_{i1}} - 1\right) r_{i1}, & 0 \leq t < w_{i1} \\ c_{i2} + \left(\frac{2(t - w_{i1})}{w_{i2} - w_{i1}} - 1\right) r_{i2}, & w_{i1} \leq t < w_{i2} \\ \vdots & \\ c_{im} + \left(\frac{2(t - w_{i(m-1)})}{1 - w_{i(m-1)}} - 1\right) r_{im}, & w_{i(m-1)} \leq t \leq 1 \end{cases}$$

onde $w_{i\ell} = \begin{cases} 0, & \ell = 0 \\ \sum_{h=1}^{\ell} p_{ih}, & \ell = 1, \dots, m \end{cases}$

Usar a representação das observações das variáveis histograma através da inversa da função distribuição acumulada – função quantil, apresenta algumas vantagens face à representação da distribuição pelo usual histograma. Estas são funções definidas por ramos, lineares em cada ramo, não decrescentes e com domínio entre 0 e 1 (Irpino & Verde, 2006; Dias, 2014; Brito & Dias, 2022).

Considerando os histogramas $H_X = \{[1,3[, 0.1; [3,5[, 0.6; [5,8[, 0.3\}$ e $H_Y = \{[0,1[, 0.8; [1,4[, 0.2\}$, estes podem ser representados pelas funções quantil:

$$\Psi_X(t) = \begin{cases} 1 + \frac{t}{0.1} \times 2, & 0 \leq t < 0.1 \\ 3 + \frac{t-0.1}{0.6} \times 2, & 0.1 \leq t < 0.7 \\ 5 + \frac{t-0.7}{0.3} \times 3 & 0.7 \leq t \leq 1 \end{cases} \quad \Psi_Y(t) = \begin{cases} \frac{t}{0.8}, & 0 \leq t < 0.8 \\ 1 + \frac{t-0.8}{0.2} \times 3 & 0.8 \leq t \leq 1 \end{cases}$$

Na Figura 3 estão representados os histogramas H_X e H_Y assim como as respectivas funções quantil.

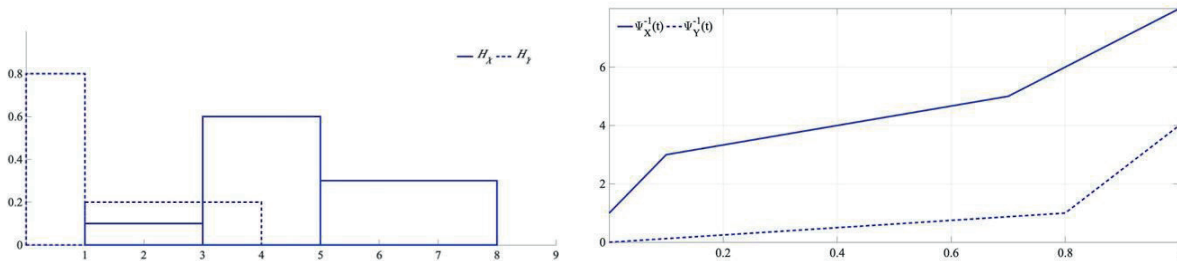


Figura 3: Representação gráfica dos histogramas H_X e H_Y e das respectivas funções quantil.

No entanto, apesar das vantagens em representar os histogramas pelas respectivas funções quantil, estas funções apresentam algumas limitações quando pretendemos efetuar operações aritméticas com este tipo de elementos.

A adição de funções quantil exige que as funções envolvidas no cálculo tenham sempre o mesmo número de ramos e os mesmos pesos acumulados, o que equivale a os histogramas tenham que ter sempre o mesmo número de subintervalos e os mesmos pesos associados a cada subintervalo i . Para resolver esta situação todos os histogramas que representam as observações das variáveis histogramas envolvidas no cálculo terão que ser reescritos nas condições exigidas. Este procedimento foi proposto por Irpino e Verde (Irpino & Verde, 2006) e baseia-se no facto de se assumir a distribuição uniforme em cada subintervalo. Aplicado este processo dizemos que as variáveis histograma ficam regularizadas.

Considerando novamente os histogramas:

$$H_X = \{[1,3[, 0.1; [3,5[, 0.6; [5,8], 0.3\} \text{ e } H_Y = \{[0,1[, 0.8; [1,4], 0.2\},$$

para reescrevê-los com o mesmo número de subintervalos e os mesmos pesos associados a cada um desses subintervalos, começamos por definir o conjunto dos pesos acumulados de H_X e H_Y , isto é, $W = \{0, 0.1, 0.7, 1, 0, 0.8, 1\}$. Selecionando os pesos sem repetição temos $Z = \{0, 0.1, 0.7, 0.8, 1\}$. Usando estes pesos e assumindo a uniformidade dentro dos subintervalos é possível reescrever os histogramas H_X e H_Y com 4 subintervalos de pesos: 0.1; 0.6; 0.1, 0.2, respetivamente. Ou seja,

$$H_X = \{[1,3[, 0.1; [3,5[, 0.6; [5,6[, 0.1; [6,8], 0.2\} \text{ e } H_Y = \{[0, \frac{1}{8}], 0.1; [\frac{1}{8}, \frac{7}{8}], 0.6; [\frac{7}{8}, 1[, 0.1; [1,4], 0.2\}.$$

Definido as funções quantil que representam os histogramas H_X e H_Y regularizados obtemos a função quantil:

$$\Psi_X(t) + \Psi_Y(t) = \begin{cases} 1 + \frac{t}{0.1} \times \frac{17}{8}, & 0 \leq t < 0.1 \\ \frac{25}{8} + \frac{t-0.1}{0.6} \times \frac{22}{8} & 0.1 \leq t < 0.7 \\ \frac{47}{8} + \frac{t-0.7}{0.1} \times \frac{9}{8}, & 0.7 \leq t < 0.8 \\ 7 + \frac{t-0.8}{0.2} \times 5 & 0.8 \leq t \leq 1 \end{cases}$$

A multiplicação de uma função quantil por um escalar positivo permite sempre obter uma função quantil, mas se a multiplicação for por um escalar negativo a função obtida é não decrescente, e conseqüentemente não é uma função quantil. Por este motivo o espaço definido pelo conjunto das funções quantil, e as operações usuais de adição e multiplicação por um escalar é apenas um espaço semi-vetorial (Dias, 2014; Dias & Brito, 2015).

Os histogramas H_X e respetivo simétrico $-H_X$, estão representados na Figura 4. No entanto se considerarmos a representação do histograma H_X pela respetiva função quantil $\Psi_X(t)$, a função quantil que representa $-H_X$ não é obtida multiplicando a função quantil $\Psi_X(t)$ por -1.

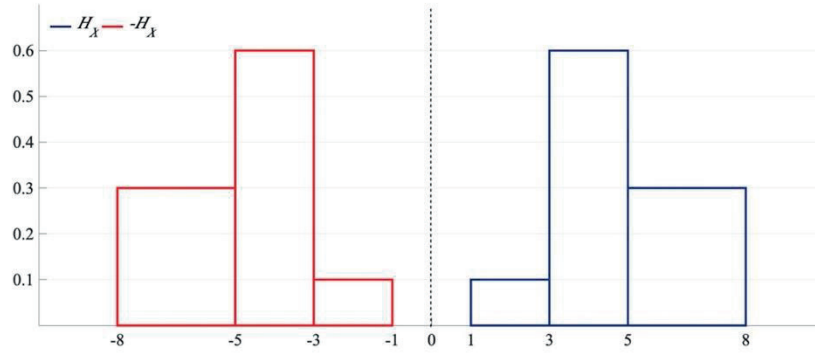


Figura 4: Representação do histograma H_X e do respectivo simétrico.

Quando multiplicamos a função quantil $\Psi_X(t)$ por um escalar negativo, neste caso -1 , obtemos uma função não decrescente (Figura 5), esta função, $-\Psi_X(t)$, não pode representar o simétrico do histograma H_X . No entanto, se a partir da função quantil $\Psi_X(t)$ efetuarmos a transformação $-\Psi_X(1-t)$, $t \in [0,1]$ obtemos a função quantil que representa o simétrico do histograma H_X (Figura 5).

$$-\Psi_X(t) = \begin{cases} -1 + \frac{t}{0.1} \times (-2), & 0 \leq t < 0.1 \\ -3 + \frac{t-0.1}{0.6} \times (-2) & 0.1 \leq t < 0.7 \\ -5 + \frac{t-0.8}{0.3} \times (-3) & 0.7 \leq t \leq 1 \end{cases} \quad -\Psi_X(1-t) = \begin{cases} -8 + \frac{t}{0.1} \times 3, & 0 \leq t < 0.3 \\ -5 + \frac{t-0.3}{0.6} \times 2 & 0.3 \leq t < 0.9 \\ -3 + \frac{t-0.9}{0.1} \times 2 & 0.9 \leq t \leq 1 \end{cases}$$

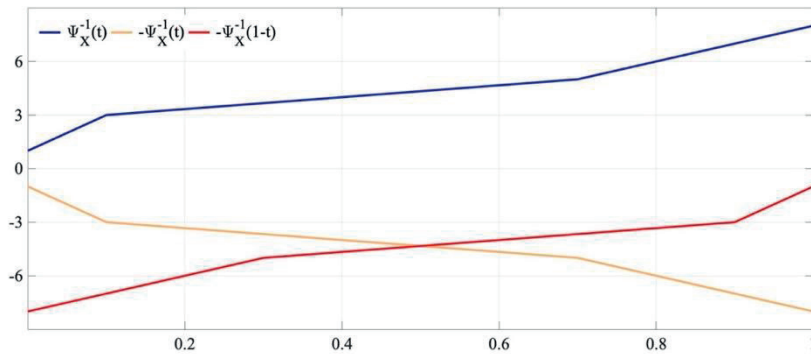


Figura 5: Representação das funções $\Psi_X(t)$; $-\Psi_X(t)$ e $-\Psi_X(1-t)$.

4.3. DISTÂNCIAS ENTRE DISTRIBUIÇÕES

Na literatura foram analisadas várias distâncias e o seu comportamento para a medição da divergência entre distribuições, sendo a Distância de Mallows (distância L_2) uma das que se mostrou estar mais adequada para este tipo de estudo (Irpino & Verde, 2006; Arroyo, 2008; Dias, 2014).

Definição 4 (Irpino & Verde, 2006): A distância de Mallows, D_M , entre duas funções quantil $\psi_X(t)$ e $\psi_Y(t)$ é dada por

$$D_M(\psi_X(t), \psi_Y(t)) = \sqrt{\int_0^1 (\psi_X(t) - \psi_Y(t))^2 dt}$$

Assumindo a uniformidade dentro dos subintervalos, cada ramo da função quantil é uma função linear, e considerando os histogramas já regularizados o quadrado da distância de Mallows entre duas funções quantil é dado por:

$$D_M^2(\psi_X(t), \psi_Y(t)) = \sum_{\ell=1}^m p_\ell \left[(c_{X_\ell} - c_{Y_\ell})^2 + \frac{1}{3} (r_{X_\ell} - r_{Y_\ell})^2 \right]$$

onde c_{X_ℓ}, c_{Y_ℓ} e r_{X_ℓ}, r_{Y_ℓ} com $\ell \in \{1; 2; \dots; m\}$ são os centros e raios do subintervalo ℓ das distribuições X e Y , respetivamente.

A distância de Mallows tem a particularidade de poder ser vista como uma generalização da distância Euclidiana. Se considerarmos que todas as observações dos histogramas contêm apenas um subintervalo degenerado e peso associado igual a 1, a expressão da distância de Mallows reduz-se à distância Euclidiana.

Encontrada uma distância adequada para medir a “semelhança” entre distribuições Irpino e Verde (Irpino & Verde, 2006) propõem uma extensão do conceito de média de Fréchet para variáveis distribucionais. Neste caso média de uma variável histograma será um histograma (baricentro).

Definição 5 (Irpino & Verde, 2006): *O baricentro da variável histograma Y , baseado na distância de Mallows e assumindo que a variável Y está regularizada é o histograma Y_B que é solução do problema de otimização*

$$\min \sum_{i=1}^n D_M^2(\psi_Y(t), \psi_{Y_B}(t))$$

O histograma baricentro, representado pela função quantil $\psi_{Y_B}(t)$, é o histograma onde o centro e raio de cada subintervalo é dado pela média clássica dos centros e raios do respetivo subintervalo, de todas as observações da variável Y .

De forma a ilustrar mais claramente o conceitos de histograma baricentro, considerou-se apenas duas observações de uma variável Y :

$$H_{Y_1} = \{[1,2[, 0.7; [2,3[, 0.2; [3,4[, 0.1\} \text{ e } H_{Y_2} = \{[11,12[, 0.1; [12,13[, 0.2; [13,14[, 0.7\}.$$

O histograma baricentro que se obtêm e que está representado na Figura 6 é:

$$H_{Y_B} = \{[6,6.6[, 0.1; [6.6,7.2[, 0.2; [7.2,7.8[, 0.4; [7.8,8.4[, 0.2; [8.4,9.1[, 0.1\}$$

Note-se que o valor médio de uma variável histograma é sempre igual ao valor médio do histograma baricentro (Irpino & Verde, 2006; Dias, 2014). Neste caso, $\bar{Y} = \bar{Y}_B = 7.5$.

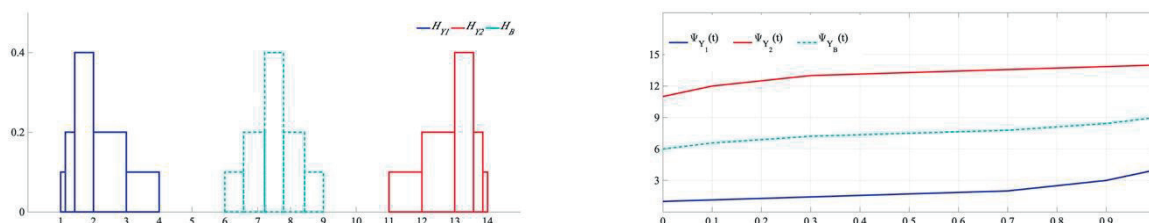


Figura 6: Representação gráfica do histograma baricentro e da respetiva função quantil.

A partir da definição de histograma baricentro, Irpino e Verde (Irpino & Verde, 2015A) apresentam definições alternativas para outras medidas descritivas tais como variância e covariância. Por exemplo, a variância é o escalar calculado pela média dos quadrados das distâncias de Mallows entre as observações e o baricentro.

5. MÉTODOS MULTIVARADOS – REGRESSÃO LINEAR

5.1. MODELO DE REGRESSÃO LINEAR PARA VARIÁVEIS HISTOGRAMAS

Nos últimos anos foram propostos vários métodos multivariados aplicados a variáveis distribucionais. Análise Classificatória (Irpino & Verde, 2008; Dias & Brito, 2022); Análise de Componentes Principais (Oliveira, Subtil & Oliveira, 2021; Le-Rademacher & Bilard, 2017; Dias & Brito, 2022); Séries Temporais (Arroyo & Maté, 2009); Regressão Linear (Billard & Diday, 2007; Dias & Brito, 2015; Irpino & Verde, 2015B; Zhao, Wang, & Lu, 2022); Análise Discriminante (Dias, Brito & Amaral, 2021).

Usando abordagens diferentes, foram propostos alguns modelos de regressão linear para variáveis histograma (Billard & Diday, 2007; Dias & Brito, 2015; Irpino & Verde, 2015B; Zhao, Wang, & Lu,

2022). De entre esses modelos, o Modelo de Distribuição e Distribuição Simétrica (DSD) (Dias & Brito, 2015) foi proposto com o objetivo de generalizar o modelo de regressão linear clássico.

Muitos dos métodos multivariados assentam no conceito de combinação linear. Assim, um conceito fundamental para definir uma generalização do modelo de regressão linear clássico é a definição de combinação linear de histogramas. A combinação linear de histogramas/funções quantil deve permitir obter um novo histograma/função quantil. Para tal, e uma vez que vão ser realizadas operações de soma e multiplicação por um escalar entre funções quantil, todos os histogramas envolvidos na combinação linear têm que estar regularizados.

A simples generalização do conceito clássico de combinação linear aplicado a funções quantil, $\psi_Y(t) = a_1\psi_{X_1}(t) + a_2\psi_{X_2}(t) + \dots + a_p\psi_{X_p}(t)$, não permite obter sempre uma função quantil. Pelo facto de estarmos a trabalhar num espaço semi-vetorial, os parâmetros a_j , $j \in \{1, 2, \dots, p\}$ têm que ser sempre não negativos.

Assim, o modelo DSD foi definido de forma a: 1) permitir obter um qualquer histograma/função quantil a partir de outros histogramas/funções quantil; 2) ter sempre parâmetros não negativos; 3) permitir que a relação entre as variáveis preditiva e explicativas possa ser direta ou inversa.

Definição 6: Modelo de Distribuição e Distribuição Simétrica (DSD Model)

Considere as variáveis histograma X_j . As funções quantil que representam as distribuições destas variáveis histograma para cada unidade i são $\psi_{X_{ji}}(t)$ e as funções quantil que representam os respetivos simétricos dos histogramas X_j são $-\psi_{X_{ji}}(1-t)$.

Cada função quantil $\psi_{Y_i}(t)$, que representa a observação i da variável histograma Y pode ser obtida por

$$\psi_{Y_i}(t) = \psi_{\hat{Y}_i}(t) + e_i(t)$$

onde $\psi_{\hat{Y}_i}(t)$ é a função quantil estimada para a unidade i , obtida a partir de

$$\psi_{\hat{Y}_i}(t) = v + \sum_{j=1}^p a_j \psi_{X_{ji}}(t) - \sum_{j=1}^p b_j \psi_{X_{ji}}(1-t)$$

com $t \in [0,1]$, $a_j, b_j \geq 0$, $j \in \{1, 2, \dots, p\}$ e $v \in \mathbb{R}$.

Com esta definição, e apesar das restrições de não negatividade aos parâmetros, é então possível obter uma função quantil como combinação linear de outras funções quantil, sendo a relação linear entre Y e X_j direta quando $a_j > b_j$ e inversa quando $a_j < b_j$.

Na Figura 7 estão representados os diagramas de dispersão entre as variáveis explicativa X e preditiva Y , quando a relação entre estas variáveis é direta e inversa, respetivamente.

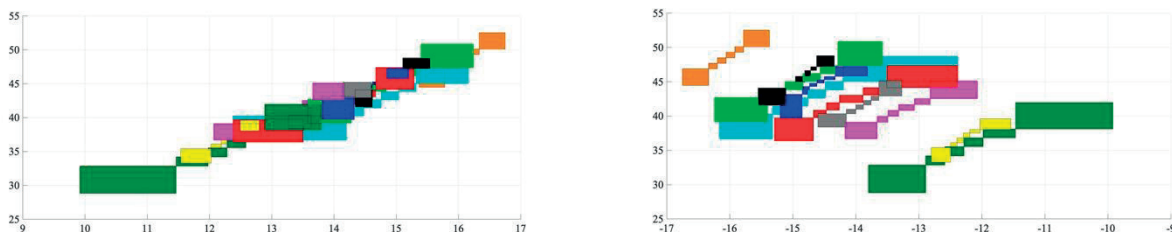


Figura 7: Projeção dos diagramas de dispersão das variáveis histograma X e Y quando os modelos DSD são $\psi_Y(t) = -1.95 + 0.41\psi_X(t) - 3.56\psi_X(1-t)$ e $\psi_Y(t) = -1.95 + 3.56\psi_X(t) - 0.41\psi_X(1-t)$, respetivamente.

No caso clássico a estimação dos parâmetros do modelo de regressão linear é obtida minimizando a soma dos quadrados dos erros, isto é, aplicando o método dos mínimos quadrados ao problema de minimização $\sum_{i=1}^m (y_i - \hat{y}_i)^2$, onde y_i e \hat{y}_i são os valores observados e estimados para a observação i , respetivamente.

No modelo DSD os parâmetros são estimados minimizando a soma dos quadrados dos erros entre as funções quantil observadas e estimadas, usando a distância de Mallows, obtendo-se assim um problema de otimização quadrática, sujeito a restrições de não negatividade aos parâmetros.

$$\begin{aligned} \text{Minimizar } SSE &= \sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}_i}(t)) \\ \text{sujeito a } & a_j, b_j \geq 0, j \in \{1, 2, \dots, p\} \\ & v \in IR. \end{aligned}$$

Também de modo análogo ao caso clássico, é possível definir uma medida para a qualidade do ajuste do modelo. Esta dedução é possível porque a soma do quadrado da distância de Mallows entre a distribuição observada $\psi_{Y_i}(t)$ e a média simbólica da variável Y pode ser decomposta por:

$$\sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \bar{Y}) = \sum_{i=1}^n D_M^2(\psi_{\hat{Y}_i}(t), \bar{Y}) + \sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \psi_{\hat{Y}_i}(t))$$

A partir desta decomposição é possível definir o coeficiente de determinação do modelo DSD, designado por Ω .

$$\Omega = \frac{\sum_{i=1}^n D_M^2(\psi_{\hat{Y}_i}(t), \bar{Y})}{\sum_{i=1}^n D_M^2(\psi_{Y_i}(t), \bar{Y})}$$

O valor de Ω varia entre 0 e 1.

O modelo de regressão linear DSD pode ser considerado um modelo de regressão linear global, porque ao ser particularizado para variáveis intervalares este modelo identifica-se com o modelo ID (Dias & Brito, 2017). Numa situação mais extrema, quando os histogramas se reduzem a valores reais, o modelo DSD identifica-se com o modelo de regressão linear clássico.

5.2. APLICAÇÃO DO MODELO DSD

Para ilustrar a aplicação a dados reais ao modelo DSD, vamos considerar uma base de dados que reúne registos de várias variáveis das comunidades no EUA (UCI Machine Learning Repository, 2013).

Tabela 5: Excerto da tabela com os registos das variáveis clássicas de comunidades dos EUA

<i>Comunidades</i>	<i>Estados</i>	<i>LEd</i>	<i>Emp</i>	<i>Div</i>	<i>Img</i>	<i>VC</i>
Selma	AL	16.59	46.94	13.35	73.86	2758.9
Bessemer	AL	16.97	46.83	14.46	18.39	1257.09
Dothan	AL	11.71	62.19	13.75	34.25	373.54
...
SanPablo	CA	14.03	55.94	16.57	62.3	374.07
Glendale	CA	11.54	60.04	11.12	60.4	644.75
...
Enfield	CT	6.55	68.24	8.38	27.01	78.65
Newingtont	CT	8.71	67.54	8.57	18.44	2127.02
NewHaven	CT	11.86	56.71	12.44	46.52	53.2
...

Os dados originais combinam condições socioeconómicas de dados do Censo de 90 e dados criminais de 1995. Para este estudo (Dias & Brito, 2015) foi selecionada como variável resposta a variável crimes violentos VC (número total de crimes violentos por 100 000 habitantes) e quatro variáveis explicativas: LEd (percentagem de pessoas com 25 ou mais anos e escolaridade inferior ao 9.º ano); Emp (percentagem de pessoas com 16 ou mais anos e empregadas); Div (percentagem da população divorciada); Img (percentagem de imigrantes que imigraram nos últimos 10 anos).

Para construir a tabela de dados simbólicos, a informação foi agregada por estado (agregação contemporânea). As unidades (de 2ª ordem) deste estudo são os estados dos EUA e observações para cada variável as distribuições dos registos das comunidades do respetivo estado (Tabela 5). Para o estudo foram apenas considerados os estados para os quais o número de registos das variáveis selecionadas foi superior a trinta. Usando esse critério, apenas vinte estados (assinalados na Figura 8) foram incluídos: AL, CA, CT, FL, GA, IN, MA, MO, NC, NJ, NY, OH, OK, OR, PA, TN, TX, VA, WA, WI.



Figura 8: Estados considerados para definir o modelo

Para construir os histogramas considerou-se em todas as observações, que os subintervalos de cada histograma teriam o mesmo peso de 0.2 (Tabela 6). Além disso, como a variável resposta, crimes violentos, apresenta apenas valores positivos e as distribuições desses valores são assimétricas, foi considerada como variável resposta, a variável LVC cujas observações são as distribuições do logaritmo do número de crimes violentos para cada estado dos EUA.

Tabela 6: Excerto da tabela de dados simbólicos, após agregação dos dados por estado dos EUA

<i>Estados</i>	<i>LEd</i>	...	<i>VC</i>
AL		...	
CA		...	
CT		...	
...

Nas condições apresentadas, o modelo DSD que permite prever a distribuição de LVC a partir das distribuições das variáveis explicativas LE*d*; Emp; Div e Im*g*, para cada estado dos EUA, é o seguinte:

$$\psi_{LVC(i)}(t) = 3.932 + 0.0009\psi_{LEd(i)}(t) - 0.012\psi_{Emp(i)}(1-t) + 0.207\psi_{Div(i)}(t) - 0.035\psi_{Div(i)}(1-t) + 0.019\psi_{Img(i)}(t)$$

O coeficiente de determinação associado a este modelo é $\Omega = 0.8680$.

Os valores dos parâmetros estimados o modelo permitem concluir que as variáveis LE*d*; Div e Im*g* têm influência direta no logaritmo do número de crimes violentos e que a percentagem de pessoas empregadas tem uma relação inversa com a variável LVC. Podemos também concluir que, para o conjunto de estados aos quais os dados se referem, quando a média simbólica da percentagem de população divorciada aumenta 1% e as restantes variáveis permanecem constantes, a média simbólica do LVC aumenta 0,1720. A percentagem da população divorciada é aquele que mais influencia a variável resposta.

Na Figura 9 estão representados os histogramas observados e estimados para a variável logaritmo do número de crimes violentos para os estados de Alabama, Califórnia e Connecticut.

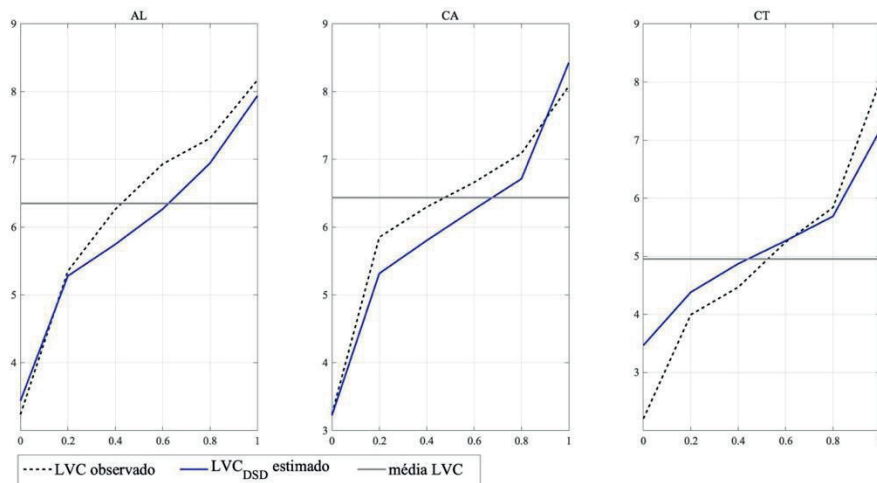


Figura 9: Representação das funções quantil observadas e estimadas para os estados AL;CA e CT.

6. CONCLUSÕES

As variáveis simbólicas permitem uma nova abordagem sobre o tratamento estatístico de bases de dados com grandes dimensões. Para além do desenvolvimento teórico dos conceitos e modelos, aplicações em diferentes áreas mostraram a utilidade dos métodos propostos pela SDA. A generalização dos métodos clássicos a dados distribucionais é um dos grandes desafios. A frase que Schweizer escreveu em 1984, “As distribuições são os números do futuro” – mostra a importância e relevância de se continuar a investigar e desenvolver métodos estatísticos para dados distribucionais.

BIBLIOGRAFIA

- Arroyo, J. (2008). *Métodos de Predicción para Series Temporales de Intervalos e Histogramas - PhD Thesis*. Universidad Pontificia Comillas, Madrid, Espanha.
- Arroyo, J., & Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting*, 25(1), 192-207.
- Billard, L., & Diday, E. (2000). Regression analysis for interval-valued data. *Data Analysis, Classification and Related Methods. Proceedings of the 7th Conference of the International Federation of Classification Societies* (pp. 369-374). Berlin Heidelberg: Springer.

- Billard, L., & Diday, E. (2002). Symbolic regression analysis. *Classification, Clustering, and Data Analysis. Proceedings of the 8th Conference of the International Federation of Classification Societies*, (pp. 281-288). Berlin Heidelberg: Springer.
- Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, 98(462), 470-487.
- Billard, L., & Diday, E. (2007). *Symbolic Data Analysis; Conceptual Statistics and Data Mining*. New York: John Wiley & Sons.
- Bock, H.-H., & Diday, E. (Edits.). (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Berlin: Springer-Verlag.
- Brito, P. (2014). Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs Data Mining and Knowledge Discovery*, 4 (4), 281-295.
- Brito, P., & Dias, S. (Edits.). (2022). *Analysis of Distributional Data*. Taylor and Francis Group.
- Dias, S. (2014). *Linear Regression with Empirical Distributions - PhD Thesis*. University of Porto, Porto, Portugal.
- Dias, S., & Brito, P. (2015). Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, 8(2), 75-113.
- Dias, S., & Brito, P. (2017). O the beaten track: A new linear model for interval data. *European Journal of Operational Research*, 258 (3), 1118-1130.
- Dias, S., Brito, P., & Amaral, P. (2021). Discriminant analysis of distributional data via fractional programming. *European Journal of Operational Research* 294(1), 206-2018.
- Diday, E. (1988). The symbolic approach in clustering and relating methods of data analysis: The basic choices. *Classification and Related Methods of Data Analysis. Proceedings of the 1st Conference of the International Federation of Classification Societies (IFCS 87)*, (pp. 673-684). Amsterdam, Holland.
- IPMA. (8 de Março de 2024). Obtido de IPMA: <https://www.ipma.pt/pt/otempo/obs.superficie/table-top-stations-all.jsp>
- Irpino, A., & Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Data Science and Classification. Proceedings of the 10th Conference of the International Federation of Classification Societies* (pp. 185-192). Berlin Heidelberg: Springer.
- Irpino, A., & Verde, R. (2008). Comparing Histogram Data Using a Mahalanobis- Wasserstein Distance. *Classification, Data Science and Classification, Proceeding of the 8th Conference of the International Federation of Classification Societies (IFCS08)* (pp. 77-89). Berlin: Springer.
- Irpino, A., & Verde, R. (2015A). Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification*, 9(2), 143-175.
- Irpino, A., & Verde, R. (2015B). Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. *Advances in Data Analysis and Classification*, 81-106.
- Le-Rademacher, J., & Billard, L. (2017). Principal component analysis for histogram-valued data. *Advances in Data Analysis and Classification*, 327-351.
- Oliveira, M. R., Subtil, A., & Oliveira, L. (2021). Detection Of Internet Attacks With Histogram Principal Component Analysis. *13th Scientific Meeting of the Classification and Data Analysis Group*. Florence, Italy.
- PORDATA. (8 de Março de 2024). Obtido de PORDATA: <https://www.pordata.pt/censos/resultados/emdestaque-portugal-1075>
- UCI Machine Learning Repository. (2013). Obtido de UCI Machine Learning Repository: In website: <http://www.ics.uci.edu/mllearn/MLRepository.html>
- Verde, R., & Irpino, A. (2008). Comparing histogram data using a Mahalanobis- Wasserstein distance. *Proceedings of the COMPOSTAT 2008, 18th International Conference on Computational Statistics* (pp. 77-89). Porto, Portugal: Physica-Verlag HD.
- Zhao, Q., Wang, H., & Lu, S. (2022). M-LDQ feature embedding and regression modeling for distribution-valued data. *Information Sciences-Elsevier*, 121-152.



Estudar Séries Temporais através de Redes Complexas

Vanessa Freitas Silva, vanessa.silva@fc.up.pt

Faculdade de Ciências, Universidade do Porto & INESC TEC – CRACS

Introdução

Quero começar por agradecer aos Editores deste Boletim Primavera 2024, cujo tema central é *Rising Stars*, o convite para escrever um texto que contribua para a partilha das novas rotas científicas que os jovens investigadores navegam atualmente. Assim, é com grande entusiasmo que compartilho uma perspetiva global da minha jornada científica.

Nos últimos anos, o meu foco tem sido a utilização de metodologias de análise de redes complexas para a análise de dados de séries temporais. Tenho me dedicado ao desenvolvimento de métodos eficientes para analisar séries temporais, tanto em contextos univariados como multivariados, utilizando abordagens da teoria de grafos e ciência de redes. Recentemente, conclui o meu doutoramento neste tema e embarquei em uma empolgante jornada numa carreira científica e académica que me permite continuar e consolidar o meu trabalho de doutoramento.

Estudar e analisar dados de séries temporais através de redes complexas é, para mim, um novo e complementar ramo da ciência atual, apesar da existência dos trabalhos de investigação de quase duas décadas. Este ramo engloba na sua essência duas grandes áreas científicas distintas, são elas a *Análise de Séries Temporais* e a *Ciência de Redes*. A seguir compartilharei uma visão muito geral deste tema, as minhas contribuições para o campo e a minhas linhas de investigação futuras.

Redes Complexas

Um grafo é uma estrutura matemática que, de forma surpreendente, permeiam muitos aspetos de nosso mundo. É composto por um conjunto de nós (ou vértices) e um conjunto de arestas (ou links) que conectam pares de nós. Esta simplicidade esconde um poder incrível, pois, como dizem, “*Everything is a Graph*”. Os grafos são possivelmente uma das estruturas de dados mais poderosas na Matemática e na Ciência da Computação, uma vez que qualquer sistema do mundo pode ser facilmente mapeado (ou transformado) em um grafo. De forma geral, os elementos que constituem um sistema qualquer podem ser representados pelo conjunto de nós e as conexões (ou relações) entre os elementos no sistema pelo conjunto de arestas [1].

Redes complexas são grafos especiais que possuem propriedades estruturais específicas, nomeadamente, propriedades topológicas não triviais, devido às características do sistema complexo subjacente, que não são nem aleatórios nem puramente regulares [1]. Estas estruturas normalmente apresentam uma distribuição não trivial das conexões entre os elementos (grau dos nós), existência de comunidades (subconjunto de nós com conexões mais fortes/densas entre si), efeito *small-world* (caminho médio entre quaisquer dois nós na rede é muito curto), entre outras propriedades.

A Ciência de Redes surge como a principal área que estuda redes complexas e tem origens na matemática discreta e na teoria dos grafos. Esta abordagem permite representar sistemas complexos de forma matemática unificada, o que possibilita o estudo de propriedades e características sistêmicas e

padrões emergentes em estruturas complexas de interconexões, indo além das conexões óbvias. Usarei os termos *grafo* e *rede* de forma intercambiável.

Séries Temporais como Redes Complexas

Viajando no universo fascinante dos grafos, imaginar a complexidade das correlações temporais e interações entre variáveis em séries temporais, damos vida a uma rede (ou conjunto de redes) que captura relações sutis e dinâmicas dos dados subjacentes. A minha pesquisa científica concentra-se exatamente nesse ponto de convergência. Ao transformar dados temporais em redes, abrimos uma janela para compreender estruturas subjacentes e padrões de evolução que podem estar ocultos nas abordagens tradicionais. Nas últimas quase duas décadas, vários métodos de mapeamento de dados de séries temporais em redes complexas que modelam as interações/dependências entre os dados têm emergido, com ênfase, nos anos mais recentes, nos dados multivariados.

Como todos sabemos, e dada a importância dos dados nos contextos atuais, encontrar as principais características identificativas de séries temporais univariadas ou multivariadas é um problema fundamental em vários domínios científicos, mas não é uma tarefa de todo trivial. As redes complexas apresentam um enquadramento para tentarmos responder a esse desafio. Séries temporais univariadas podem então ser mapeadas para redes de uma única camada através de métodos de mapeamento baseados em conceitos de visibilidade, transição e proximidade. E séries temporais multivariadas podem seguir abordagens de uma única camada ou múltiplas camadas. A Figura 1 apresenta a taxonomia de algoritmos de mapeamento de séries temporais em redes complexas existentes na literatura [2]. A vermelho destaco os dois novos métodos de mapeamento que, juntamente com a minha equipa de investigação, propus nos últimos anos.

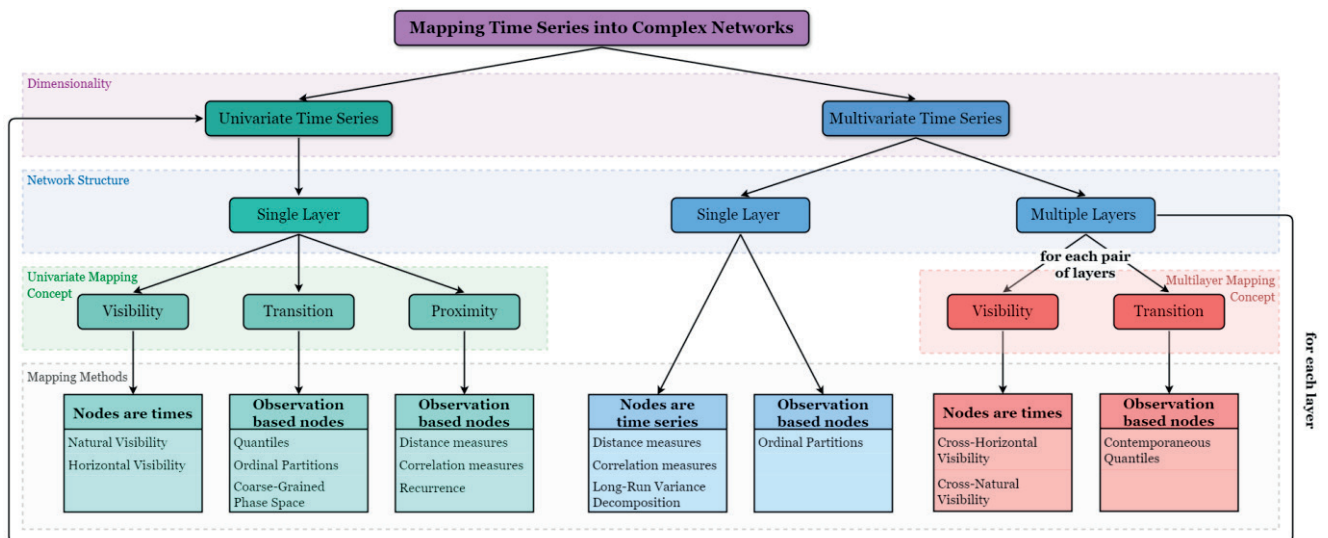


Figura 1 – Visão geral dos métodos de mapeamento. Taxonomia de algoritmos para mapeamento de séries temporais em redes complexas com base na dimensionalidade das séries temporais, estrutura da rede resultante, conceito de mapeamento e principais métodos de mapeamento [2].

Esta abordagem que visa representar séries temporais através de grafos oferece uma metodologia complementar aos métodos tradicionais e ao mesmo tempo oferece uma plataforma robusta para o desenvolvimento de métodos eficientes para a sua análise. Com base no método de mapeamento utilizado, podemos obter diferentes análises dos dados de séries temporais. Em geral, métodos baseados em conceitos de visibilidade mapeiam séries temporais em grafos com os nós representando cada *timestamp* e as arestas são definidas com base em uma relação geométrica entre valores de dados. Grafos de visibilidade refletem as propriedades locais e globais da série, especialmente por meio de máximos locais. Uma vantagem associada a estes métodos, é que não requerem pré-processamento dos dados da série temporal e são livres de parâmetros. Estes métodos tornaram-se muito populares na literatura, pois são intuitivos (devido aos critérios geométricos associados) e de fácil compreensão. Os métodos baseados em probabilidades de transição, intuitivamente, representam padrões de transição

(dinâmica) de uma série temporal com base em diferentes tipos de codificação simbólica dos dados. Nestes métodos, os dados são representados por quantis, padrões de ordem, ou em estados dinâmicos, reduzindo a significativamente sua dimensionalidade, os nós das redes resultantes são os diferentes símbolos de dados e as arestas são dirigidas e com pesos associados com base nas probabilidades de transição temporal entre os símbolos na séries temporal. Estes métodos exigem assim a escolha de parâmetros associados à codificação. Os grafos de proximidade representam a similaridade dos dados ao longo do tempo, refletindo como as propriedades locais das séries evoluem ao longo do tempo. Normalmente, os métodos envolvem a reconstrução dos dados da série temporal em vetores, como ciclos ou vetores no espaço de fase, exigindo a seleção de parâmetros. Os nós da rede representam esses vetores e as arestas representam medidas de similaridade ou distância entre os vetores. Os grafos de transição e de proximidade são bastante populares no estudo de sistemas dinâmicos.

Recentemente, as preocupações têm sido dirigidas para os dados de séries temporais multivariadas. Estas podem ser transformadas em redes de camada única que reduzem cada série temporal a um único nó na rede e os nós são conectados por arestas com base nas relações entre as séries temporais, utilizando por exemplo medidas de correção, modelos estatísticos, entre outros. Outra transformação possível e mais recente, é visa transformar séries temporais multivariadas em redes de múltiplas camadas, uma estrutura mais complexa de redes. Cada série temporal é mapeada em uma camada de redes individual com base em mapeamentos usados em contextos univariados, e estabelecem as relações entre pares de séries temporais conectando arestas entre nós das camadas de redes correspondentes na rede de múltiplas camadas. A Figura 2 ilustra duas redes de múltiplas camadas resultantes de uma série temporal multivariadas real de exemplo com base em dois métodos de mapeamento que desenvolvi durante o meu doutoramento. Estas redes de múltiplas camadas foram desenhadas para capturar as dependências cruzadas entre pares de séries temporais e as probabilidades de transição da dinâmica (variações de dados) entre os dados, respetivamente, através das novas arestas entre camadas de redes incorporadas nos métodos de mapeamento.

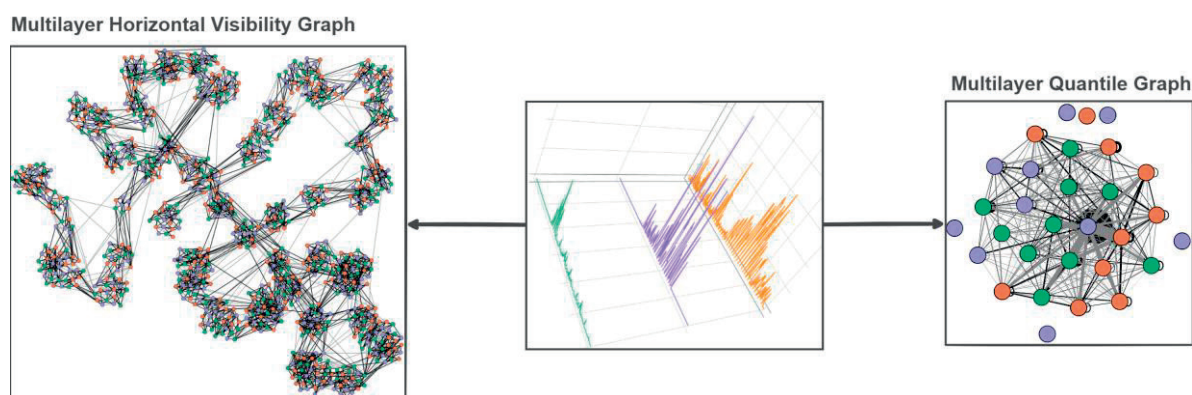


Figura 2 – Ilustração de duas redes de múltiplas camadas que resultam de dois métodos de mapeamento de séries temporais multivariadas diferentes. À esquerda o método é baseado no conceito de visibilidade [3] e à direita baseado no conceito de probabilidade de transição [4]. Nós de cores diferentes correspondem às diferentes variáveis de séries temporais, as linhas (arestas) pretas representam as conexões internas e as linhas (arestas) a cinza as conexões entre dimensões.

Novas Medidas para Análise de Séries Temporais: Medidas Topológicas

Um tópico quente na análise de séries temporais são as abordagens baseadas em features, que resultam de medidas estatísticas descritivas dos dados de séries temporais. Muitas vezes, os métodos baseados em features de séries temporais são uma tarefa preliminar essencial em muitas aplicações práticas, tais como problemas envolvendo classificação, clustering, deteção de anomalias ou outliers, geração de dados sintéticos e até mesmo em problema de previsão de séries temporais. Esta abordagem permite simplificar a representação dos dados temporais, reduzindo a sua complexidade e dimensionalidade uma vez que caracteriza e descreve os conjuntos de dados através de um vetor muito mais reduzido de características específicas extraídas dos dados. Várias técnicas de extração de features de séries

temporais univariadas têm sido propostas, mas para contextos multivariados não existem conjunto de features típicas.

NetF [5] é uma nova metodologia que desenvolvi para extrair novos conjuntos de features descritivas para séries temporais univariadas. Esta metodologia combina features topológicas de redes obtidas de dados de séries temporais univariadas resultantes de diferentes métodos de mapeamento, nomeadamente, grafos de visibilidade e grafos de probabilidade de transição, dos quais são extraídas 5 medidas topológicas de redes de cada tipo de grafo. Esta combinação de features enriquece o conjunto de estatísticas comumente usadas na análise de séries temporais, introduzindo novas propriedades discriminatórias dos dados que são capturadas pelas diferentes redes que capturam propriedades distintas dependendo do método de mapeamento usado. O *NetF* resulta em um conjunto de 15 features que podem ser usadas nos diversos problemas de análise de séries temporais e está disponível em software R para a comunidade científica através da seguinte página web: em <https://www.dcc.fc.up.pt/~vanessa.silva/publications.html>. Na JOCLAD2023, este trabalho foi destacado com o *Prémio Fernando Nicolau*.

Tendo por base o *NetF* e os métodos de mapeamento para dados de séries temporais multivariadas, a expansão da metodologia acima para o contexto multivariado é o quanto óbvia. No entanto, é uma tarefa mais complexa devido às questões de dimensionalidade dos dados. O *MNetF* surge assim como uma proposta a essa expansão onde combinamos features topológicas de redes de múltiplas camadas que resultam dos dois métodos de mapeamento de séries temporais multivariadas desenvolvidos. Dados os desenvolvimentos da ciência de redes e dada a estrutura de alto-nível das redes de múltiplas camadas, são muitas as features topológica que podemos extrair de tais redes e que podem servir de avanço para a introdução de novas features descritivas dos dados originais. Resultados preliminares mostram que estas features topológicas de alto-nível podem classificar dados reais de séries temporais multivariadas, em problemas de classificação de dados, por exemplo. No entanto, algumas questões precisam ser ainda estudadas, nomeadamente, a extração eficiente dessas características continua a ser um desafio em termos de complexidade temporal e espacial, e a seleção/escolha adequadas das features mais representativas para conjunto de dados específicos. Este é um dos tópicos em andamento no meu trabalho atual.

Durante o percurso desenvolvido nos últimos anos de doutoramento, tenho desenvolvido uma framework em linguagem de programação C++, à qual chamei de *tsmnet*, que inclui diversas metodologias e métodos computacionais para analisar séries temporais de alta dimensão através de metodologias de redes complexas, em particular, redes de múltiplas camadas. Esta framework estará brevemente disponível para ser utilizada de forma prática pela comunidade científica.

Perspetivas Futuras

Analisar dados de séries temporais através das metodologias da ciência de redes pode e tem-se mostrado uma abordagem bastante promissora para abordar questões fundamentais em aberto na análise de séries temporais. Entre as inúmeras possibilidades, destaco as seguintes: séries temporais irregulares, dados com valores em falta, visualização de séries temporais multivariadas, dados espaço temporal e privacidade de dados de séries temporais.

Este novo ramo científico tem, a meu ver, um potencial enorme em termos de aplicação. Espero que o meu trabalho de investigação contribua para a exploração de novos horizontes e inspire também outros a explorar as infinitas possibilidades que os grafos oferecem na compreensão dos dados temporais.

Referências

- [1] Albert, R., & Barabási, A. L. (2002). *Statistical mechanics of complex networks*. *Reviews of modern physics*, 74(1), 47.
- [2] Silva, V. F., Silva, M. E., Ribeiro, P., & Silva, F. (2021). *Time series analysis via network science: Concepts and algorithms*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(3), e1404.

- [3] Silva, V. F., Silva, M. E., Ribeiro, P., & Silva, F. (2023). *MHVG2MTS: Multilayer Horizontal Visibility Graphs for Multivariate Time Series Analysis*. arXiv preprint arXiv:2301.02333.
- [4] Silva, V. F., Silva, M. E., Ribeiro, P., & Silva, F. (2023). *Multilayer Quantile Graph for Multivariate Time Series Analysis and Dimensionality Reduction*. arXiv preprint arXiv:2311.11849.
- [5] Silva, V. F., Silva, M. E., Ribeiro, P., & Silva, F. (2022). *Novel features for time series analysis: a complex networks approach*. *Data Mining and Knowledge Discovery*, 36(3), 1062-1101.



O *software* de fonte aberta na inteligência artificial e na sociedade

José Pombal, jose.pombal@unbabel.com

Unbabel

Em meados do século XVII, Peter Chamberlen terá sido contratado, pela rainha Henriqueta e pelo seu marido, o rei Carlos I de Inglaterra, para servir de cirurgião obstetra da família real. Seguir-se-ia um período de mais de 100 anos de partos bem sucedidos — ou com uma taxa de sucesso mais alta — em que algum membro Chamberlen serviria a família real enquanto obstetra. Durante este período, a razão do sucesso era desconhecida pelo o povo em geral; corriam rumores de que sempre que uma caixa ornamentada entrava para o quarto onde se estava a dar o parto, saía um bebé e uma mãe saudáveis. 500 anos mais tarde, o fórceps é um instrumento usado, quando indicado, para salvaguardar a saúde do bebé e da mãe em partos mais complicados, mas permaneceu um segredo da família real inglesa durante 150 anos. O número de vidas que poderiam ter sido salvas se a ferramenta tivesse sido divulgada aquando da sua invenção é incontável.

Nos dias de hoje, a Ciência é consideravelmente mais aberta. Desde a erradicação da varíola nos anos 60, aos esforços de investigação relacionados com a COVID-19 mais recentemente, não é difícil encontrar exemplos de colaboração ao nível mundial, em que investigação é partilhada para atingir um bem comum. E de mão dada com esta abertura, vem o salvamento de inúmeras vidas, e um desenvolvimento de sociedades mais sustentáveis à escala mundial. Os benefícios da ciência aberta são impossíveis de negar.

No caso do *software* também existe uma noção de abertura. Por um lado, chama-se *software* de *fonte aberta* aos programas que qualquer utilizador tem a liberdade de utilizar, copiar, distribuir, estudar, alterar, e melhorar o código fonte.¹ *Software* de *fonte fechada* (ou proprietário) é tudo o resto. Todos os dias utilizamos ambos, e os programas de fonte aberta são cada vez mais prevalentes: a maior parte de nós já contactou com um destes programas, o *Android*, o sistema operativo de telemóveis mais utilizado no mundo. Mesmo quando são de fonte fechada, os programas muitas vezes assentam em *software* de fonte aberta. Por exemplo, o núcleo do sistema operativo MacOS da Apple baseia-se no XNU.

Dito isto, o *software* não é como a medicina. A sua abertura, ou a falta dela, não põe vidas em jogo — ou, quando põe, a ligação não é tão clara. Por isso, o dilema colocado perante a família Chamberlen mantém-se de forma mais persistente: para quê divulgar como reproduzir a minha invenção, quando posso mantê-la em segredo e beneficiar dela sem concorrência? Neste caso, a ideia não é manter o *software* em segredo total — como o fórceps era —, mas sim vendê-lo sem permitir a sua reprodução, ocultando o seu código fonte. Muitos dirão que a possibilidade de venda com pouca concorrência é o incentivo que leva pessoas e empresas a inventarem produtos. Eliminado esse incentivo, não haverá inovação.

No entanto, a crescente prevalência de *software* de fonte aberta põe esta crença em causa. Sei que não estou sozinho quando escrevo que este tipo de *software* tem um impacto brutal na minha vida, área profissional, e na empresa onde trabalho, a Unbabel. Assim, o objetivo deste texto é que o leitor o termine

¹ De acordo com a Open Source Initiative, uma das entidades que mais promove o desenvolvimento de *software* de *fonte aberta*.

questionando-se se a produção deste software é suficientemente incentivada, e se os benefícios da fonte aberta são devidamente promovidos na nossa sociedade.

Acordo, e olho brevemente para o meu telemóvel, cujo sistema operativo é Android. Navego um pouco o *X* para consultar alguns dos novos artigos da minha área — inteligência artificial (IA) —, que foram publicados nos últimos dias. A vasta maioria dos quais se baseia em, e contribui para, *software* de fonte aberta. No trabalho, quase todas as ferramentas que utilizo são de fonte aberta; desde o *software* que uso para programar (VSCode), ao *browser* que utilizo (Firefox), à linguagem de programação que escrevo (Python), e ao sistema operativo das máquinas onde realizo experiências (Linux). Uma das únicas exceções é o meu portátil de trabalho, Apple, cujo sistema operativo é o MacOS; este é de fonte fechada, embora, tal como referi acima, se baseie em *software* de fonte aberta. Convido o leitor a refletir acerca dos programas que usa, e a consultar se são de fonte aberta.

Se der um passo atrás até aos meus tempos de universidade, vejo que o padrão se repete. Quase tudo o que aprendi e usei se baseava em programas de fonte aberta. Isso levou a que, tendo já tido a oportunidade de trabalhar com vários profissionais de diferentes nacionalidades, me apercebesse que tivéramos bases muito parecidas na universidade. Tivemos acesso a recursos semelhantes e pudemos trabalhar com base no mesmo *software*. Por isso, hoje, conseguimos colaborar facilmente numa grande variedade de projetos. A existência e a proliferação de *software* de fonte aberta permite que alunos de países como Portugal estejam ao nível de estudantes de países com significativamente mais meios. O mesmo não seria possível com *software* fechado; este estaria provavelmente barrado por trás de licenças que não seriam acessíveis nem a todas as universidades, nem a todos os alunos. Também seria bastante mais difícil de ensinar através dele, dada a sua natureza mais opaca e menos manipulável. O *software* de fonte aberta é um verdadeiro motor da igualdade de oportunidades.

A minha experiência pessoal não evidencia todas as vantagens de *software* de fonte aberta. Uma das mais importantes é que este tende a ser mais robusto e seguro que alternativas proprietárias. Um bom exemplo disto é o facto de Linux ser o sistema operativo da grande maioria de supercomputadores, servidores, e infraestrutura computacional de todo o tipo de empresas e instituições.² Esta vantagem assenta num dos princípios base da fonte aberta: qualquer pessoa pode rever e melhorar o código fonte. Se existir uma vulnerabilidade, há potencialmente milhares de olhos a tentar encontrar a solução correspondente. No caso de *software* proprietário, há apenas uma equipa de engenharia. É claro que também pode haver muitos olhos a procurar vulnerabilidades para fins ilícitos, mas a perseverança dos inúmeros projetos de fonte aberta de grande sucesso (Linux, Mozilla Firefox, Android, etc...), assim como estudos empíricos [1], demonstram que o risco compensa. Para conhecer melhor as vantagens deste tipo de *software*, aponto o leitor para os sítios da [Open Source Initiative](#) e da [Free Software Foundation](#).

Então, se o *software* de fonte aberta é assim tão bom, e se todos beneficiamos muito da sua existência, porque é que existe *software* proprietário? A resposta parece simples: o *software* de fonte aberta, enquadrado num modelo de negócio tradicional, dificilmente gera lucro. Este *software* depende de uma comunidade, e a maior parte dos seus contribuidores não recebe qualquer tipo de compensação; não é um produto, mas sim uma ferramenta pública e transparente. Olhando para a questão da perspetiva económica, como será possível que algo completamente copiável gere lucro? Na verdade, é possível; não só com modelos híbridos, mas também com modelos que dependem exclusivamente de *software* de fonte aberta, como mostrarei adiante.

Nos próximos parágrafos, darei exemplos da área da inteligência artificial e é necessário estabelecer algumas definições. O produto das empresas desta área é muitas vezes um modelo que é usado para um determinado fim, como conversar, prever a probabilidade de fraude de uma transação, ou gerar uma imagem. O modelo é caracterizado pelos seus pesos que são, na prática, uma série de matrizes com números. Se estes números não são conhecidos, o modelo é fechado. Adicionalmente, é necessário *treinar* o modelo em *dados*, o que requer código. Este código pode ser de fonte aberta, e os dados

² Por exemplo, desde 2017, todos os supercomputadores na lista *Top 500*, uma lista renomeada dos supercomputadores mais potentes do mundo, correm em Linux (https://en.wikipedia.org/wiki/TOP500#Architecture_and_operating_systems).

públicos, mas o modelo pode ainda assim ser fechado. De modo geral, para um modelo de inteligência artificial ser completamente de fonte aberta, é necessário que 1) o código utilizado para o treinar seja de fonte aberta e inteiramente detalhado, 2) os dados de treino sejam de fonte aberta, e 3) o modelo seja aberto (os seus pesos sejam conhecidos). Um modelo que apenas cumpra o último requisito é chamado “aberto”, ou “de pesos abertos”.

As empresas que procuram fazer dinheiro com *software* procuram mantê-lo fechado, e a área de inteligência artificial não é exceção. Um caso muito mediático foi o da OpenAI, a empresa que lançou o ChatGPT. Esta empresa foi fundada com o propósito de contribuir para a investigação de inteligência artificial de forma inteiramente transparente, produzindo *software* (e modelos) de fonte aberta para acelerar o progresso científico. No entanto, por volta do lançamento do ChatGPT, anunciou que iria alterar a sua política: a produção de *software* de fonte aberta já não seria uma prioridade. Hoje em dia, uma subscrição ao modelo ChatGPT mais avançado custa cerca de 20\$ por mês, e o modelo não cumpre nenhum dos requisitos de fonte aberta acima delineados. Este golpe deixou a comunidade ressentida, especialmente considerando que os modelos GPT se baseiam numa série de inovações de fonte aberta, a principal das quais — o *Transformer* — sendo de uma empresa competidora, a Google.

A OpenAI deparou-se com um produto com o potencial de alterar fundamentalmente a sociedade, e estava na posição perfeita para o explorar. Por isso, escolheu o caminho mais direto para o lucro, desistindo da sua missão inicial. Apesar de os incentivos económicos serem claros, este tipo de decisão tem vários riscos. No caso de modelos de geração de texto como o ChatGPT, fala-se frequentemente de como podem falhar e gerar texto perigoso e tóxico. Ora, a forma mais eficaz de nos protegermos contra estas falhas e de promover a sua resolução é abrir os modelos, permitindo o acesso ao seu código de treino, aos dados em que foram treinados, e aos seus pesos. No mínimo, a abertura dos pesos já permitiria algum grau de segurança. No caso do ChatGPT, nenhum destes requisitos é cumprido, e temos de confiar na OpenAI para salvaguardar os interesses da sociedade. É uma posição precária, no mínimo, e perigosa, no pior dos casos.

Será, então, que, apesar de o *software* de fonte aberta constituir uma externalidade francamente positiva para a sociedade, não há um grande incentivo para o produzir? Não; não só o *software* de fonte aberta pode coexistir com o de fonte fechada, como também pode ser lucrativo por si só. Tal como foi referido anteriormente, tanto um como o outro tipo de *software* têm papéis preponderantes nas nossas vidas. E, na realidade, o *software* de fonte aberta tem um papel fundamental no funcionamento das empresas produtoras de *software* de fonte fechada. Além dos vários exemplos já dados, a *Meta* é outro importante. Não só é a empresa detentora do Facebook e Instagram, como também se posiciona atualmente como uma das maiores empresas defensoras da fonte aberta. Entre vários outros projetos, a *Meta* coordena o desenvolvimento do *PyTorch*, a biblioteca de programação da qual depende a maior parte dos projetos de inteligência artificial, incluindo, provavelmente, o ChatGPT.³

A Unbabel, a empresa onde trabalho, é outro exemplo onde o desenvolvimento de *software* de fonte aberta coexiste e sinergiza com a produção de *software* de fonte fechada. A Unbabel vende uma plataforma cujo componente principal é a tradução automática de conteúdo escrito. No entanto, as traduções nem sempre são perfeitas à primeira; podem ser necessários melhoramentos antes de serem entregues ao cliente. A forma como traduções menos boas são detetadas na Unbabel é através de um modelo de inteligência artificial chamado COMET [2].⁴

A primeira versão do COMET foi lançada em 2020 de forma totalmente aberta. A sua utilização e reprodução livre foi permitida pela Unbabel, e possibilitada pelo seu código de treino ser de fonte aberta, os seus dados de treino também, e os seus pesos serem abertos — cumprindo os 3 requisitos. É

³ Sendo de fonte fechada, é impossível saber exatamente como é que o ChatGPT está a ser desenvolvido. No entanto, a vasta maioria dos projetos do mesmo tipo depende de alguma forma de *PyTorch*, por isso é razoável assumir que o ChatGPT também.

⁴ De forma muito simplificada, o COMET recebe uma tradução, e devolve uma pontuação de qualidade, assim como os excertos do texto que contêm erros (se os houver).

atualmente a melhor e mais utilizada ferramenta do seu tipo.⁵ Apesar de a base da tecnologia ser completamente pública, é, ainda assim, uma vantagem competitiva considerável para a Unbabel. A versão mais recente do COMET, o xCOMET [3], é também de fonte aberta, mas os seus termos de utilização não permitem o seu uso para fins comerciais que possam competir com os da Unbabel. Apesar de não cumprir totalmente a definição da Open Source Initiative, esta solução apresenta um equilíbrio interessante, conjugando alguns benefícios da fonte aberta com as necessidades de mercado da empresa. Por um lado, esta continua a ter um incentivo para desenvolver o produto; por outro, qualquer pessoa pode utilizar e manipular o *software* para fins de aprendizagem, investigação, e avanço científico.

Mais recentemente, a Unbabel lançou o TOWER [4], um modelo do mesmo tipo do ChatGPT, mas orientado para tradução e tarefas relacionadas, como pós-edição automática de traduções. Os seus pesos são abertos, e cumpre parcialmente os requisitos de abertura do código e dados de treino. É o melhor modelo aberto de tradução para as línguas em que foi treinado, e também tem uma licença que permite o seu uso para investigação, mas não para fins comerciais. Não é totalmente de fonte aberta, mas atinge um equilíbrio entre beneficiar a comunidade científica, e proteger os incentivos à inovação da empresa. Além disso, a comunidade ajuda a resolver alguns problemas com o modelo, algo que é benéfico para Unbabel. Uma semana depois de lançarmos o TOWER, utilizadores do modelo, externos à Unbabel, encontraram *bugs* que nos tinham escapado até então. Sem estas contribuições, seria possível que futuros clientes que utilizassem o TOWER viessem a sofrer destes problemas, prejudicando a competitividade da Unbabel.

Construir um modelo de negócio lucrativo baseado apenas em *software* de fonte aberta também é possível. A empresa mais conhecida que segue este modelo é a Red Hat, que desenvolve *software* de fonte aberta, e vende serviços de apoio e treino a consumidores dos programas. Além disso, já fez aquisições de *software* proprietário, que depois voltou a lançar como fonte aberta. É uma empresa altamente lucrativa, e foi adquirida em 2018 pela IBM por 34 mil milhões de dólares. Outros exemplos são a Igalia, e a Canonical; esta última desenvolve o Ubuntu, a distribuição de Linux que usei no meu computador durante a minha licenciatura e mestrado. Produzir *software* de fonte aberta pode, de facto, ser altamente lucrativo.

Os governos também podem incentivar o desenvolvimento de *software* de fonte aberta. Alguns dos fundos que empresas como a Unbabel receberam para investigação têm requisitos em relação à abertura do *software* desenvolvido. Mais iniciativas com esta visão devem ser fomentadas. Da mesma forma, deve ser promovido o diálogo sobre os benefícios do *software* de fonte aberta e o seu desenvolvimento. Também ao nível Europeu, a abertura do *software* deve ser tomada em conta. O recente AI Act prevê algumas condições em relação a este tipo de *software*. O balanço parece geralmente positivo, mas não é claro que todas as medidas sejam benéficas para todas as empresas, especialmente as pequenas.⁶

Por fim, a educação também tem um papel importante — talvez o mais importante — no que diz respeito à divulgação dos benefícios do *software* de fonte aberta, e, por consequência, ao seu desenvolvimento. Eu, enquanto autor deste texto que tem um viés claro a favor da fonte aberta, só ouvi falar deste dilema no mundo do *software* durante a universidade. Num mundo cada vez mais digitalizado, em que já se ensina crianças a programar na escola, é necessário que este conhecimento seja ensinado muito antes na vida. Os cidadãos precisam de saber os prós e contras da natureza do *software* que usam, e o impacto que isso pode ter nas suas vidas a nível da privacidade, segurança, entre outros. Quanto mais cedo aprendermos e começarmos a refletir sobre estas questões, melhor.

Mais do que uma característica de um produto, a fonte aberta é uma atitude. Pode ser aplicada ao *software*, à ginecologia-obstetrícia, no caso do fórceps, à epidemiologia, no caso da COVID-19, ou a

⁵ O COMET ganhou várias vezes, incluindo a deste ano, as competições de estimativa de qualidade de traduções da WMT, a conferência internacional mais prestigiada da área, onde empresas como a Google também competem.

⁶ Uma coleta breve de parte dos prós e contras do AI Act em relação a *software* de fonte aberta foi publicado pela Creative Commons, uma das organizações sem fins lucrativos mais conhecida na área do licenciamento de *software* (<https://creativecommons.org/2023/12/11/on-openness-copyright-eu-ai-act-final-version-appears-to-include-promising-changes/>).

qualquer outra área. Os benefícios são claros, e os riscos da atitude contrária também. Esta atitude permitiu que eu tivesse uma educação universitária ao nível das melhores do mundo, apesar de viver num país pequeno. Permitiu que a Unbabel produzisse modelos de inteligência artificial melhores que os de empresas muito maiores. E permitiu que a OpenAI criasse o ChatGPT. É a atitude que levará ao desenvolvimento mais rápido, eficaz, sustentável, e igualitário da ciência e da sociedade, e, por isso, deve ser incentivada, divulgada, e protegida com todos os meios disponíveis.

Referências

- [1] Wang, J., & Carroll, J. M. (2011, May). Behind Linus's law: A preliminary analysis of open source software peer review practices in Mozilla and Python. In 2011 International Conference on Collaboration Technologies and Systems (CTS) (pp. 117-124). IEEE.
- [2] Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. arXiv preprint arXiv:2009.09025.
- [3] Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. (2023). xcomet: Transparent machine translation evaluation through fine-grained error detection. arXiv preprint arXiv:2310.10482.
- [4] Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, A., ... & Martins, A. F. (2024). Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. arXiv preprint arXiv:2402.17733.



Há mar e mar, há ir e voltar

Pedro Miranda Afonso, p.mirandaafonso@erasmusmc.nl

Erasmus Medical Center em Roterdão, Países Baixos

Vivemos num tempo em que a definição de emigração no debate público parece reencontrar as palavras de Eça, “em Portugal a emigração não é, como em toda a parte, a transbordação de uma população que sobra; mas a fuga de uma população que sofre.” É importante recordar, sem demagogias, que a mobilidade internacional sempre esteve presente no nosso sistema científico. Olhando hoje para os nomes que se destacam no sistema científico português, nas mais diversas áreas do conhecimento, facilmente se encontra um denominador comum: uma experiência científica internacional. Assim, no contexto da investigação científica, a mobilidade é uma condição natural e habitual.

Em ciência sabemos que a pergunta (de investigação) certa é geralmente mais importante do que a resposta certa à pergunta errada. Recorrentemente ouvimos “O que podemos fazer para reter este talento em Portugal?” Entendo que esta não é a pergunta certa. Fazer esta pergunta é limitar o potencial dos nossos investigadores, privando-os de sistemas de investigação mais maduros. Estas experiências internacionais são salutares. A mobilidade em ciência potencia a diversidade e, essa, enriquece não só a produção científica mas também os seus agentes. Problemas globais requerem respostas globais. O problema não está na partida, mas sim na atratividade para o regresso. A pergunta deve focar-se não na retenção de talento, mas sim na sua captação. Como podemos tornar Portugal um polo de atracção científica? Atrativo não só para investigadores portugueses, mas para investigadores de qualquer nacionalidade. Se o for para os outros, também será para nós. Integro hoje um departamento onde a ciência se faz com colegas das mais variadas nacionalidades. Não faz, por isso, sentido para mim pensar que o desenvolvimento científico em Portugal se fará apenas em português.

Entendo que a SPE pode ter aqui um papel preponderante. Promovendo o seu encontro para uma audiência internacional e tornando-o também um ponto de encontro para a diáspora científica portuguesa a trabalhar nas áreas de Probabilidades e Estatística. Um encontro para troca de experiências, conhecimentos e o começo de novas colaborações. Apesar da distância, estou certo de que, tal como eu, muitos outros investigadores portugueses no estrangeiro não são alheios ao Portugal contemporâneo.

Até já.

P.S.: Aos estudantes que estão neste momento na dúvida em fazer a tese de mestrado, uma visita científica, ou até um (pos-)doutoramento no estrangeiro. Não sigam caminhos fáceis. Venham. Estas experiências fazem-nos entender melhor quem somos e quem queremos ser. Porque, como um dia nos confidenciou Saramago no conto da ilha desconhecida, “é necessário sair da ilha para se ver a ilha.”

Pedro Miranda Afonso
(estudante de Doutoramento)

• Artigos em Revistas

- Aknouche, A., Scotto, M.G. (2024). A multiplicative thinning-based integer-valued GARCH model. *Journal of Time Series Analysis* 45, 4-26.
- Brilhante, M.F.; Pestana, D.; Pestana, P.; Rocha, M.L.; Sequeira, F. (2023). A Metrologia e a Transição Digital: Medição da Severidade de Vulnerabilidades e Risco de Exploração, *Revista Medições e Ensaios*, 17, p. 3-34.
- Brilhante, M.F.; Pestana, D.; Pestana, P.; Rocha, M.L. (2023). Measuring the Risk of Vulnerabilities Exploitation, *Applied Maths*, 4(1), p. 20-54.
- Brilhante, M.F., Gomes, M.I., Mendonça, S., Pestana, D. & Pestana, P. (2023). Generalized Beta Models and Population Growth: So Many Routes to Chaos. *Fractal and Fractional* 7:2, 194, 40 pages.
<https://doi.org/10.3390/fractalfract7020194>
- Gomes, M.I. (2023). Recollection 3, In: Hsing, T & Rootzén, H. (2023). Remembering Ross Leadbetter: some personal recollections. *Extremes* 26, 399-412
<https://doi.org/10.1007/s10687-023-00464-y>
- Jamba, Nelson T.; Filipe, Patrícia A.; Jacinto, Gonçalo; Braumann, Carlos A. (2023). Stochastic differential equations mixed model for individual growth with the inclusion of genetic characteristics. *Statistics, Optimization & Information Computing*,
<https://doi.org/10.19139/soic-2310-5070-1829>
(Published online on December 19, 2023, <http://www.iapress.org/index.php/soic/article/view/1829>).
- Martins, A, Scotto, M.G., Weiss, C.H. Gouveia, S. (2023). Space-time integer-valued ARMA modelling for time series of counts. *Electronic Journal of Statistics* 17, 3472-3511.

• Revista

Título: *REVSTAT* – Statistical Journal
Ano: 2023 / 24. Volume 21 – Número 4 e Volume 22 – Número 1
<https://revstat.ine.pt/index.php/REVSTAT/issue/view/74>
<https://revstat.ine.pt/index.php/REVSTAT/issue/view/75>
Listagem dos artigos nas páginas da secção Notícias deste Boletim SPE.

• Livros

Título: *Reliable Alternative Ways to Manage the Risk of Extreme Events*.
Autores: Gomes, M.I., Figueiredo, F. e Henriques-Rodrigues, L.
In: C.P. Kitsos, T.A. Oliveira, F. Pierrri & M. Restaino (eds.), *Statistical Modelling and Risk Analysis*, Springer Proceedings in Mathematics & Statistics 430,
Ano: 2023
ISBN: 978-3-031-39863-6,
ISSN: 2194-1009,
Editora: Springer Nature Switzerland AG, Cham, pp. 91-105.
https://doi.org/10.1007/978-3-031-39864-3_8

Título: A Escola de Extremos em Portugal—PORTSEA, do Inglês ‘Portuguese School of Extremes and Applications’.
Autora: Maria Ivette Gomes
Ano: 2023
Em: *Memórias da Academia das Ciências de Lisboa, Classe de Ciências*, Tomo XLIX, pp 65-78.

Título: *A Escola de Extremos e Aplicações em Portugal*—PORTSEA.

Autora: Maria Ivette Gomes

Ano: 2023

Book Preprint under submission to ACL (viii+247 pp.).

Available at: <https://doi.org/10.13140/RG.2.2.19277.41447>

Título: 2023 *Institute of Mathematical Statistics (IMS) International Conference on Statistics and Data Science (ICSDS): Program Book*.

Editores: Gomes, M.I., Oliveira, T., Oliveira, A., Pestana, P. & Xu, M.

Ano: 2023

Editora: IMS editions.

ISBN: 978-0-940600-86-7 (xlii + 673 pages).

<https://sites.google.com/view/icsds2023/full-program-book>

• Tese de Mestrado

Título: *Crédito Renovável*

Autora: Joana Miranda Oliveira, joana99mo@gmail.com

Orientadoras: Margarida Brito e Maria do Carmo Miranda Guedes



Doutoramento

Título: *Sobre a teoria de modelos espaciotemporais para séries temporais de contagens e sua aplicação a eventos de saúde*

Autora: Ana Martins (a.r.martins@ua.pt)

Orientadores: Sónia Gouveia (Universidade de Aveiro), Manuel Scotto (IST, Universidade de Lisboa) e Christian Weiss (Helmut-Schmidt-Universität, Alemanha)

A minha tese focou-se no desenvolvimento de modelos estatísticos para séries temporais de contagens, isto é, valores inteiros não negativos. Estes modelos têm particular relevância quando as séries temporais observadas contêm contagens baixas, uma vez que os modelos contínuos assentam no pressuposto da distribuição Normal e, sendo esta uma distribuição simétrica não é capaz de acomodar o comportamento das contagens baixas, que terão uma distribuição assimétrica. A investigação em modelos de séries temporais para contagens tomou impulso no início da década de 80 com a introdução do operador de filtragem binomial (em inglês, *binomial thinning operator*). Rapidamente este operador foi utilizado como uma alternativa à multiplicação nos modelos autorregressivos (AR), garantindo assim, juntamente com um processo de inovação discreto, que o processo seria de valor inteiro, daí a sua designação de modelos inteiros autorregressivos (INAR). A introdução dos modelos INAR com um processo de inovação de Poisson iniciaram o desenvolvimento de uma classe de modelos bastante rica, uma vez que ao contrário dos modelos contínuos estes não têm uma especificação única. Esta classe designa-se por modelos INARMA - inteiros autorregressivos e de médias móveis.

Há já diversas contribuições de modelos univariados na classe de modelos INARMA. Em contraste, o desenvolvimento de modelos multivariados tem progredido mais lentamente. Isto resulta da dificuldade no desenvolvimento teórico destes modelos, mas também da implementação de estratégias de estimação eficientes. Assim, o meu trabalho de doutoramento pretendeu dar uma contribuição dentro desta classe, olhando também para a dependência espacial de processos de valores inteiros. Esta nova classe designa-se por modelos inteiros espaciotemporais ARMA (STINARMA) e é inspirada nos modelos contínuos STARMA. A informação espaciotemporal é introduzida no modelo através de uma matriz de vizinhança que é incorporada no operador de filtragem binomial.

Os modelos STINARMA foram inicialmente estudados detalhadamente nas suas componentes puramente de médias móveis (STINMA) e puramente autorregressiva (STINAR), posteriormente resultados para a classe, combinando as duas componentes AR e MA, foram obtidos. Para cada um destes modelos os momentos até à segunda ordem, incluindo as funções de autocovariância e autocorrelação espaciotemporais, bem como propriedades probabilísticas foram estudadas. A estimação de parâmetros foi também abordada neste trabalho, nomeadamente através do Método dos Momentos, dos Mínimos Quadrados Condicionais e da Máxima Verosimilhança Condicional. O desempenho dos métodos de estimação em amostras finitas foi avaliado através de estudos de simulação. Finalmente, estes modelos foram utilizados para estudar a associação entre contagens diárias de admissões hospitalares respiratórias e fatores ambientais no distrito de Aveiro, recorrendo a modelos STINAR não estacionários com parâmetros variáveis no tempo.

O desenvolvimento deste trabalho foi financiado por uma Bolsa Individual de Doutoramento (SFRH/BD/143973/2019) da Fundação para a Ciência e a Tecnologia, atribuída na área científica da Matemática.

Ana Martins





SPE

Sociedade Portuguesa
de Estatística

PRÉMIOS “ESTATÍSTICO JÚNIOR 2024”

REGULAMENTO

Está aberto **até 31 de maio de 2024** o concurso para atribuição dos prémios “Estatístico Júnior 2024”, de acordo com o seguinte regulamento:

1. A atribuição dos prémios “Estatístico Júnior 2024” é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio do Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL), e tem como objetivo estimular e desenvolver o interesse dos alunos dos Ensinos Básico e Secundário pelas áreas de Probabilidade e Estatística.
2. Podem candidatar-se aos prémios “Estatístico Júnior 2024” os alunos inscritos, no ano letivo 2023/24, no 3.º Ciclo do Ensino Básico, no Ensino Secundário, nos Cursos de Educação e Formação (CEF) ou nos Cursos de Educação e Formação de Adultos (CEFA).
3. As candidaturas podem ser **individuais** ou em **grupo com um máximo de 3 alunos**. De cada candidatura pode ainda fazer parte um professor, do grau de ensino em que o trabalho se insere, ao qual cabe o papel de orientador.
4. Os candidatos devem apresentar um trabalho com temática envolvendo as áreas de **Probabilidade ou Estatística**.
5. O **trabalho** deverá ser constituído por um texto escrito em português com um máximo de 10 páginas A4, com tipo, tamanho e estilo de letra padrão. O texto deverá ser acompanhado de um *vídeo* (com uma duração máxima de 5 minutos) ou de uma apresentação em *PowerPoint/Google Slides/Canvas* (máximo 15/20 slides) ou de um *poster* formato A2, que resuma os principais aspetos do trabalho.
6. Poderão ser atribuídos prémios “Estatístico Júnior 2024” a sete trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3.º Ciclo do Ensino Básico, aos três primeiros classificados de entre os trabalhos candidatos do Ensino Secundário e ao primeiro classificado de entre os trabalhos candidatos dos Cursos CEF ou CEFA. Os prémios são constituídos por livros juvenis de divulgação científica e serão entregues numa cerimónia de premiação com a comunidade académica.
7. Ao professor orientador do trabalho classificado em 1.º lugar, em cada grau de ensino, é atribuída uma anuidade grátis como sócio da SPE e um livro de divulgação científica.
8. A candidatura é composta pelo **Boletim de Candidatura**, devidamente preenchido, e pelo **trabalho** (texto e vídeo/apresentação/poster). Para incluir o vídeo, caso seja a sua escolha, por favor, faça *upload* do vídeo para uma plataforma de alojamento de vídeos (como *YouTube, Vimeo*, ou outra plataforma da sua escolha) e forneça o *link* diretamente no Boletim de Candidatura. A candidatura, dirigida ao Presidente da SPE, deverá ser enviada
 - a) em formato digital (pdf) por *e-mail* para spe@spestatistica.pt
 - b) impressa em papel (juntamente com o link de alojamento do vídeo, caso pretenda submeter um vídeo) para Sociedade Portuguesa de Estatística – Bloco C6, Piso 4 – Campo Grande – 1749-016 Lisboa.
9. O carimbo do correio validará a data de entrega do trabalho, sendo os autores notificados por *e-mail* sobre a sua receção no prazo de uma semana.
10. A admissibilidade e apreciação dos trabalhos submetidos a concurso é da competência de um júri, constituído e nomeado pela Direção da SPE. O júri é soberano nas decisões, não havendo lugar a impugnação ou recurso.
11. A atribuição dos prémios “Estatístico Júnior 2024” será anunciada logo que seja conhecida a decisão do júri e a sua entrega formal será realizada numa sessão pública expressamente dedicada a essa entrega.
12. Os prémios “Estatístico Júnior 2024” poderão não ser atribuídos.
13. O boletim de candidatura e este regulamento podem ser obtidos em www.spestatistica.pt

Apoios



Centro de Estatística e Aplicações
Universidade de Lisboa



Fundação
para a Ciência
e a Tecnologia

DOI: 10.54499/UIDB/00006/2020

Edições SPE - Mini Cursos

Título: *Análise Estatística de Dados Financeiros*

Autores: C. Amado, C. Nunes, A. Sardinha

Ano: 2019.

Título: *Uma introdução à Meta-Análise*

Autora: Maria de Fátima Brilhante

Ano: 2017.

Título: *Estatística Bayesiana*

Computacional – uma introdução

Autores: M. Antónia Amaral Turkman e Carlos Daniel Paulino

Ano: 2015.

Título: *Análise de Valores Extremos: Uma Introdução*

Autoras: M. Ivette Gomes, M. Isabel Fraga Alves e Claudia Neves

Ano: 2013.

Título: *Modelos com Equações*

Estruturais

Autora: Maria de Fátima Salgueiro

Ano: 2012.

Título: *Análise de Dados Longitudinais*

Autoras: Maria Salomé Cabral e

Maria Helena Gonçalves

Ano: 2011

Título: *Uma Introdução à Estimação*

Não-Paramétrica da Densidade

Autor: Carlos Tenreiro

Ano: 2010

Título: *Análise de Sobrevivência*

Autoras: Cristina Rocha e

Ana Luísa Papoila

Ano: 2009

Título: *Análise de Dados Espaciais*

Autoras: M. Lucília de Carvalho e

Isabel C. Natário

Ano: 2008

Título: *Introdução aos Métodos*

Estatísticos Robustos

Autores: Ana M. Pires e

João A. Branco

Ano: 2007

Título: *Outliers em Dados Estatísticos*

Autor: Fernando Rosado

Ano: 2006

Título: *Introdução às Equações*

Diferenciais Estocásticas e

Aplicações

Autor: Carlos Braumann

Ano: 2005

Título: *Uma Introdução à Análise de Clusters*

Autor: João A. Branco

Ano: 2004

Título: *Séries Temporais – Modelações lineares e não lineares*

Autoras: Esmeralda Gonçalves e

Nazaré Mendes Lopes

Ano: 2003 (2ª Edição em 2008)

Título: *Modelos Heterocedásticos.*

Aplicações com o software Eviews

Autor: Daniel Muller

Ano: 2002

Título: *Inferência sobre Localização e Escala*

Autores: Fátima Brilhante, Dinis

Pestana, José Rocha e

Sílvio Velosa

Ano: 2001

Título: *Modelos Lineares*

Generalizados – da teoria à prática

Autores: M. Antónia Amaral

Turkman e Giovanni Silva

Ano: 2000

Título: *Controlo Estatístico de Qualidade*

Autoras: M. Ivette Gomes e

M. Isabel Barão

Ano: 1999

Título: *Tópicos de Sondagens*

Autor: Paulo Gomes

Ano: 1998

Retrospectiva do Boletim SPE - Tema Central

Disponíveis em <https://www.spestatistica.pt/publicacoes/categoria/boletim-da-spe>

- Outono de 2023** – Educação (e) Estatística
- Primavera de 2023** – PORTSEA – um mar de Extremos em Portugal
- Outono de 2022** – Prémios na *Sociedade Portuguesa de Estatística*
- Primavera de 2022** – Liderança Estatística
- Outono de 2021** – *Machine Learning* e Inteligência Artificial
- 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 – perspectiva 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 Económicos
- 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



O MUNDO DA ESTATÍSTICA

ORGANIZAÇÃO PARTICIPANTE



Federation of European National Statistical Societies



Índice

Editorial	2
Mensagem do Presidente	4
Notícias	5
<i>Enigmística</i>	14
<i>SPE e a Comunidade</i>	
Literacia para a Qualidade em Estatísticas Oficiais	
<i>Pedro Campos e J. Pinto Martins</i>	15
<i>Rising Stars</i>	
Explorando a lei de formação das estrelas por seleção de variáveis	
<i>Ana Helena Tavares</i>	19
Misturas de modelos de regressão com efeitos aleatórios	
<i>Luísa Novais</i>	22
Dados simbólicos – Os dados do futuro!	
<i>Sónia Dias</i>	25
Estudar Séries Temporais através de Redes Complexas	
<i>Vanessa Silva</i>	38
O software de fonte aberta na inteligência artificial e na sociedade	
<i>José Pombal</i>	43
Há mar e mar, há ir e voltar	
<i>Pedro Miranda Afonso</i>	48
<i>Ciência Estatística</i>	49
<i>Doutoramento</i>	51
<i>Prémios Estatístico Júnior 2024</i>	52
<i>Edições SPE - Minicursos</i>	53
<i>Boletim através do Tema Central</i>	54