Boletim

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

Publicação semestral *primavera de 2019*

Séries Temporais de Valor Inteiro

Informação Editorial

Endereço: Sociedade Portuguesa de Estatística.
Campo Grande. Bloco C6. Piso 4.
1749-016 Lisboa. Portugal.
Telefone: +351.217500120
e-mail: *spe*@spestatistica.pt
URL: http://www.spestatistica.pt
ISSN: 1646-5903
Depósito Legal: 249102/06
Tiragem: 350 exemplares
Execução Gráfica e Impressão: Gráfica Sobreirense
Editor: Fernando Rosado, *fernando.rosado@fc.ul.pt*

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

www.spestatistica.pt

Hotel Casa da Calçada

Amarante

6 a 9 de novembro de 2019

https://spe2019.estg.ipp.pt

Editorial

… o bem da Estatística, em época aniversária, também com a importância do *Congresso SPE* …

1. No próximo mês de novembro a SPE inicia o ano 40. Boa data, estimulante de uma reflexão sobre o dinamismo que lhe é devido. Em 2005, por ocasião dos 25 anos, foi editado o *Memorial da Sociedade Portuguesa de Estatística* que, com a preciosa colaboração de diversos colegas, fez um bom relato da vida científica dos estatísticos portugueses. Passados mais 15 anos a que devemos juntar todas as novas realidades e desafios é a hora de um novo olhar. Uma releitura do trabalho feito pela Antónia Amaral Turkman e publicado a páginas 93 e seguintes no *Memorial* acima referido pode ser um bom estímulo para criarmos esse novo olhar, por exemplo, sobre os Doutoramentos em Estatística em Portugal – uma boa variável de análise da vitalidade da Ciência Estatística. Fica o desafio editorial para "pequenos relatos" de colegas das diferentes universidades e aos quais se possam acrescentar "grandes desafios". As páginas do Boletim SPE decerto ficarão bem preenchidas com esse novo olhar.

2. Este número do Boletim (também) celebra 40 anos. De facto, o Boletim SPE, como se sabe, nasceu a partir do Boletim Informativo de Estatística e Investigação Operacional cujo Número 1 foi publicado em Fevereiro de 1979.


BOLETIM INFORMATIVO
DE
ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL
N.º 1 FEVEREIRO 1979

Há 40 anos!

O Boletim nasceu um pouco antes da SPE. No Editorial dessa publicação histórica é relatada a gestação de uma nova sociedade científica.

A SPEIO / SPE nasceu no ano seguinte. Assim, caminhando à frente, o Boletim foi (e ainda deve ser!) um arauto da Estatística. No Boletim primavera de 2009, como homenagem, na secção Episódios na História da Estatística, apresentámos o fac-símile das oito páginas daquela primeira edição histórica.

Pelo esforço e a generosidade de muitos autores, esta publicação tem conseguido uma atividade regular ao longo de quarenta anos e de que muito se deve orgulhar.

Aos editores meus antecessores presto homenagem e gratidão pelo empenho e o zelo com que contribuíram para que o Boletim sempre tivesse conseguido avançar - muitas vezes em situações bastante difíceis.

Como atual responsável pela edição renovo os meus agradecimentos a todos os autores que dedicadamente têm colaborado ajudando assim a criar novos exemplares. Através das suas páginas, o Boletim SPE, além de informações específicas, divulga muita investigação que se faz em Estatística - em torno de temas centrais e assuntos de índole mais abrangente. É também um incentivo à publicação científica dos mais jovens.

A sua leitura, acima de tudo, deve proporcionar conhecimento do panorama científico no campo da Estatística em Portugal. Então, cada exemplar é dedicado à Investigação e à Estatística. A presente edição festiva, celebra e realça esse desiderato:

3. Muitos colegas – sócios SPE e militantes desde há longa data – se têm manifestado sobre a importância dos Congressos e das Atas. E em ano de congresso sempre o tema emerge e coloca a exigência de "olhar para". De novo, deve manifestar-se que: a concretização de um Congresso SPE é o principal objetivo estatutário da SPE, como acontece com todas as associações científicas. É uma oportunidade de iniciação para muitos jovens estatísticos. É uma manifestação do valor científico da comunidade. É uma transmissão de saber! É uma manifestação de continuidade vital. É uma reunião magna; a todos os títulos meritória do carinho e do esforço que lhe é dedicado.

Estamos em ano de Congresso SPE. Neste Boletim já damos notícia desse grande acontecimento e, com agradecimento à Comissão Organizadora!

Um Congresso SPE é o resultado de uma soma de esforços: da Direção da SPE, da Comissão Científica e da Comissão Organizadora que realizam uma iniciativa Presidencial mandatada pelos sócios e que concretizam um dos grandes objetivos estatutários: transmissão do saber. É um desígnio de liderança! Quando realizamos congressos bienais "parece que cumprimos apenas metade desse desígnio". É sempre oportuno refletir sobre este assunto. E para que o sucesso fique em alta, façamos a nossa participação "render o dobro".

4. O *Boletim SPE primavera de 2019* teve a colaboração especial da Prof. Isabel Pereira, na eleição do Tema Central, bem como dos autores convidados. Agradecimentos são devidos a todos pela disponibilidade na partilha do seu saber com toda a comunidade científica.

O Tema Central do próximo Boletim SPE será *Estatística nas Ciências da Saúde.*

Fernando Rosa

Mensagem da Presidente

Caros Sócios da SPE,

Passado mais um ano do mandato da atual Direção é tempo de refletir sobre as atividades passadas e planear as atividades futuras. Este exercício foi feito na preparação da Assembleia Geral Ordinária de 2019 que teve lugar dia 22 de Fevereiro. Num resumo breve, as atividades nortearam-se segundo os dois grandes objectivos traçados por esta Direção:

Objetivo I- aumentar a sensibilização pública para a Estatística e aumentar a visibilidade da SPE na sociedade;

Objetivo II- aumentar a coesão interna da sociedade e apoiar o desenvolvimento da Estatística em Portugal.

Assim, no âmbito do **Objetivo I** destacam-se as atividades já usuais *Prémio Estatístico Júnior, AEVAE, Explorística* e presença na Feira da Matemática. De entre as atividades que contribuíram para o **Objetivo II** destaco o Prémio SPE 2018 atribuído a Soraia Pereira com o trabalho intitulado *Spatio-temporal models for georeferenced unemployment data*; e o Prémio Iniciação à Investigação atribuído a Pedro Nicolau com o trabalho intitulado: *Estimating the spatio-temporal variation of bird phenology using citizen science data*. Outros pontos altos do ano 2016 foram a realização do III Encontro Luso-Galaico de Biometria em Aveiro e a comemoração do 38º aniversário da SPE. A SPE tem apoiado diversos eventos científicos que se realizam Portugal e são de interesse para os sócios e tem-se assegurado a representação da SPE em diversas organizações nacionais. Os sócios podem consultar os detalhes das atividades no Relatório de atividades de 2018 que se encontra disponível, mais uma vez, na página da SPE. A Direção quer deixar aqui um agradecimento aos sócios que, certamente à custa de sacrifício pessoal, possibilitaram a realização das atividades no ano de 2018.

Relativamente a 2019, destaco a organização do Congresso sobre o qual podem os sócios consultar detalhes quer neste boletim quer na página web da SPE.

A Sociedade é dos sócios e para os sócios e é, essencialmente, o que os sócios fizerem dela. Assim, a Direção está aberta a apoiar todas as iniciativas em prol da Estatística em Portugal.

Porto, 13 de Março de 2019

Cordiais saudações

Maria Eduarda Silva

Bolsas para participação no XXIV Congresso da SPE, 2019

Pretendendo estimular o estudo e a investigação científica em Probabilidades e Estatística entre os jovens, a SPE atribui um número limitado de bolsas para participação no XXIV Congresso da SPE, que decorrerá entre 6 e 9 de Novembro de 2019 em Amarante, de acordo com o seguinte regulamento:

1. Os candidatos devem ser estudantes de mestrado ou de doutoramento inscritos no ano lectivo 2018/2019 em alguma instituição portuguesa.
2. São também admitidos candidatos que tenham completado o respectivo ciclo de estudos durante o ano de 2018.
3. Os candidatos não devem ter completado os 35 anos de idade até 31 de dezembro de 2019.
4. A bolsa é constituída pela inscrição no Congresso e por uma quantia de 100 euros.
5. A candidatura consta de um resumo alargado (documento em pdf com 2 a 4 páginas) para uma comunicação oral e de uma carta de apresentação onde deve constar uma breve biografia. O resumo deve ser escrito em português. A candidatura deve ser enviada à Direcção da SPE para spe@spestatistica.pt até 15 de Junho de 2019. Os candidatos devem fazer prova das condições de admissibilidade descritas em 1, 2 e 3.
6. A decisão será comunicada a 15 de Julho de 2019.

Notícias

• XXIV Congresso da Sociedade Portuguesa de Estatística

Entre 6 e 9 de novembro de 2019 realiza-se, no hotel Casa da Calçada – Amarante, o XXIV Congresso da Sociedade Portuguesa de Estatística. O evento conta com uma organização conjunta entre a Escola Superior de Tecnologia e Gestão (ESTG-PPorto), o Instituto Superior de Engenharia do Porto (ISEP-PPorto) e a Sociedade Portuguesa de Estatística (SPE) e tem como principal objetivo a partilha de novos desenvolvimentos na área da Estatística e respetivas implicações.

No dia 6 de novembro realiza-se o habitual minicurso pré-congresso, este ano, intitulado "Análise Estatística de Dados Financeiros" e apresentado pelos colegas Cláudia Nunes, Conceição Amado e Alberto Sardinha (todos do Instituto Superior Técnico, Universidade de Lisboa).

O programa científico conta com sessões conjuntas com outras Sociedades, várias sessões temáticas, comunicações livres selecionadas (orais e em poster) e quatro sessões plenárias proferidas pelos conferencistas convidados: Bruno Falissard (Universidade de Paris-Sud), Maria Manuela Neves (Instituto Superior de Agronomia, Universidade de Lisboa), Maria do Rosário Oliveira (Instituto Superior Técnico, Universidade de Lisboa) e Walter J. Radermacher (Presidente da FENStatS, Universidade de Roma).

A possibilidade de um convívio para estreitamento de relações, tanto ao nível científico como ao nível pessoal, não foi esquecida, havendo espaço e tempo pensados para tal.

As submissões de resumos deverão ser realizadas até **31 de maio**. Mais informações em *https://spe2019.estg.ipp.pt*.

Contamos com a presença de todos no congresso, pois foi para vós que ele foi pensado e organizado.

Até novembro, em Amarante!

A Comissão Organizadora Local

• Prémio SPE 2018

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

Júri do Premio SPE 2018:
Presidente: Raquel Menezes da Universidade do Minho
Vogais: Irene Oliveira da Universidade de Trás-os-Montes e Alto Douro
Russell- Alpizar-Jara da Universidade de Évora

O Prémio SPE 2018 foi atribuído a Soraia Alexandra Gonçalves Pereira.
No final desta edição do Boletim damos o devido destaque a esta notícia.

FR

• REVSTAT - Statistical Journal

REVSTAT-Statistical Journal é uma revista editada pelo INE, de acesso aberto e indexada em relevantes bases de dados de citações de artigos académicos, e que teve a liderança de Ivette Gomes no período 2003-2018.

REVSTAT lançou recentemente o Volume 17, Número 1 – janeiro de 2019 (*https://www.ine.pt/revstat/tables.html*), que inclui uma Nota Editorial da nova Editora-Chefe, Isabel Fraga Alves, para o próximo período de 5 anos. Neste período 2019-2023 a revista vai beneficiar de um Conselho Editorial Internacional constituído por conceituados especialistas em Estatística e Aplicações (*https://www.ine.pt/revstat/eboard.html*).

Os artigos originais, a serem publicados em três volumes por ano, podem abranger qualquer tópico de Probabilidade e Estatística e serem diversificados segundo os tipos Teoria, Métodos e Aplicações, Revisão, Estudo de casos e Comunicações breves.

A submissão de artigos à REVSTAT é feita de uma forma simples, de acordo com as instruções em *https://www.ine.pt/revstat/info.html* Os editores aguardam por submissão de manuscritos com resultados relevantes da investigação científica desenvolvida.

URL: *https://www.ine.pt/revstat/inicio.html*

Giovani Silva

• Sessão Comemorativa do 38º Aniversário da SPE

No passado dia 28 de Novembro, decorreu a Sessão Comemorativa do 38º Aniversário da SPE. O programa incluiu a Conferência Convidada "*Equações Diferenciais Estocásticas e Modelação e Inferência em Biologia*". Seguiu-se a apresentação e entrega do Prémio SPE 2018, de que adiante damos realce. Soraia Alexandra Gonçalves Pereira com o trabalho *Spatio-temporal models for georeferenced unemployment data* foi a vencedora. A Sessão terminou com um lanche convívio.

FR

• 11th European Conference on Mathematical and Theoretical Biology

A Biomatemática está em crescimento rápido na Europa e, para celebrar a importância das aplicações da matemática à biologia e ciências da vida, foi declarado o "Year of Mathematical Biology" 2018 (YMB) por iniciativa conjunta da European Mathematical Society (EMS) e da European Society for Mathematical and Theoretical Biology (ESMTB). Com a organização de múltiplos eventos, incluindo programas temáticos e conferências e workshops, o acontecimento principal do YMB foi a **11th European Conference on Mathematical and Theoretical Biology (ECMTB 2018)** (**http://www.ecmtb2018.org**).

Tradicionalmente organizado pela ESMTB, o ECMTB 2018 foi desta vez (e pela primeira vez) também organizado pela EMS, tendo a Sociedade Portuguesa de Matemática (SPM) como coorganizadora. A Conferência realizou-se na Faculdade de Ciências da Universidade de Lisboa (FCUL), sendo anfitrião o seu Centro de Matemática, Aplicações Fundamentais e Investigação Operacional (CMAFcIO). O ECMTB 2018 teve o alto patrocínio de Sua Excelência o Senhor Presidente da República e foi-lhe atribuído o selo da UNESCO pela Comissão Nacional da UNESCO. Para além do patrocínio das três sociedades organizadoras, foi apoiado por diversos centros de Investigação nacionais (CMAFcIO, CMA, CIMA, CEAUL) e a Fundação para a Ciência e a Tecnologia (FCT), pelo Instituto Gulbenkian de Ciência, por editoras internacionais (Springer, MDPI, PLOS One, Elsevier, EMS-PH, IOP Publishing. Oxford University Press, Wiley), pela Bernoulli Society, pela Sociedade Portuguesa de Estatística (SPE), pelo Centro Internacional de Matemática (CIM) e por outras organizações e empresas.

De 23 de julho (22 de julho para aqueles que participaram no registo antecipado e cocktail de boas vindas) a 27 de julho, a cidade de Lisboa acolheu os mais de setecentos participantes de 80 países com um clima agradável (muito mais fresco do que é habitual nesta usualmente cálida época do ano, uma bênção para os participantes da Europa Central e do Norte que vinham de temperadas inesperadamente elevadas para os seus países). O número record de participantes (só ultrapassado nas Conferências conjuntas ESMTB-SMB) é um sinal segura da crescente importância da Biologia Matemática.

Após a cerimónia de abertura, realizou-se um Tributo a Karl Peter Hadeler proferido por **Odo Diekmann**, a que seguiu imediatamente a conferência plenária de abertura e Bernoulli Society-European Mathematical Society Joint Lecture, proferida por **Samuel Kou** (Universidade de Harvard, EUA), sobre o excitante tema "Big data, Google and disease detection: A statistical adventure". As outras conferências plenárias focaram tópicos igualmente excitantes e foram proferidas pelos eminentes cientistas **Helen Byrne** (Universidade de Oxford, Reino Unido, "Mathematical approaches to modelling and remodelling biological tissues"), **Antonio DeSimone** (SISSA, Itália, "Biological and bio-inspired motility at microscopic scales: locomotion

by shape control"), **Eva Kisdi** (Universdade de Helsinlki, Finlândia, "Adaptive dynamics and the evolution of diversity"), **Mirjam Kretzschmar** (University Medical Centre Utrecht, Holanda, "Modelling the waning and boosting of immunity"), **Eva Löcherbach** (Universidade de Cergy-Pontoise, França, "Modeling interacting networks of neurons as processes with variable length"), **Andrea Pugliese** (Universidade de Trento, Itália, "Epidemic models structured by parasite load and immune level"), **Eörs Szathmáry** (Universidade Eötvös Loránd, Hungria, "Models of learning and evolution: what do they have in common?") e **Kees Weijer** (Universidade de Dundee, Reino Unido, "Analysis of collective cell behaviours underlying primitive streak formation in the chick embryo"). Um dos recentes vencedores do Prémio Reinhart Heinrich para a melhor tese de doutoramento na área, **Jochen Kursawe**, pôde comparecer e apresentar a tradicional palestra do vencedor sobre "Quantitative approaches to investigating epithelial morphogenesis".

A ECMTB 2018 teve um programa igualmente rico com 36 Mini-simpósios em áreas de ponta, 60 sessões paralelas com Comunicações Orais e 2 sessões de Posters (numa amena confraternização em volta de um "coffee+cocktail break"), totalizando 455 comunicações orais e 119 posters cobrindo todas as áreas da Biologia Matemática e Teórica. Um júri atribuiu os 4 prémios para posters, que foram patrocinados por grandes editoras (Elsevier, MDPI, Springer). Programa em http://www.ecmtb2018.org/files/files/ecmtb_booklet_a4.pdf. Em http://www.ecmtb2018.org/ActMap o resumo. Em http://www.ecmtb2018.org/files/files/BookOfAbstracts_ECMTB2018_inclusions.pdf pode ver-se o Livro de Resumos com ISBN.

Além disso, foi organizado o "ECMTB Mentorship Programme", de forma a facilitar interações, em termos de Investigação e de carreira, entre cientistas juniores e cientistas mais seniores que participaram no evento.

A Assembleia Geral da ESMTB teve lugar a 26 de julho e estava aberta a membros e não-membros da Sociedade. Foram discutidos vários temas sobre o funcionamento da ESMTB e o desenvolvimento da Biologia Matemática e Teórica, que prosseguiram informalmente na prova de vinhos que se seguiu.


O programa social proporcionou amplas oportunidades para discussões científicas e contactos pessoais. Para além das atividades sociais já mencionadas e dos "coffee breaks" e "lunch breaks", realizaram-se excursões e o jantar da conferência, que começou com a atuação de uma tuna académica e proporcionou a seguir uns pés de dança ao som de uma banda.


Para celebrar o "Year of Mathematical Biology 2018" e desejando alargar o recrutamento a outros investigadores da área, a ESMTB convidou os participantes inscritos na ECMTB 2018 que ainda não fossem membros da ESMTB a associarem-se. A Sociedade dá as boas vindas aos que aceitarem este convite isentando-os da quota de sócio no primeiro ano. Note-se que este convite ainda se mantém válido (ver informação detalhada em http://dev.ecmtb2018.org/RegRules).

Em nome da Comissão Organizadora, vimos agradecer às sociedades organizadores pela confiança em nós depositada, à Comissão Científica, aos patrocinadores, aos oradores convidados, aos organizadores dos Mini-simpósios, aos presidentes das sessões, aos participantes do programa de mentorado, ao júri dos prémios dos posters, aos estudantes que ajudaram e aos eficientes e empenhados membros do Secretariado (Ana Rita Ferrer, Ana Isabel Figueiredo, Joana Guia).

Estamos especialmente gratos a todos os participantes, para os quais esta Conferência foi organizada, por terem feito dela um evento memorável e um marco importante no progresso da Biomatemática.

Maíra Aguiar, Carlos Braumann, Nico Stollenwerk (Conference Chairs)

• ECAS2019 on Statistical Analysis for Space-Time Data: 15-17 Julho 2019

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

O ECAS2019 on Statistical Analysis for Space-Time Data, organizado pela SPE (Sociedade Portuguesa de Estatística) e pelo SEIO (Spanish Society of Statistics and Operational Research), consiste num conjunto de **4 cursos** sobre o tema da **Análise de Dados Espaço-Temporais**, ministrados por peritos internacionais bem conceituados na área.

Vindo da Austrália, *Adrian Baddeley*, conjuntamente com *Ege Rukak*, falarão sobre a metodologia dos **padrões pontuais espaciais** e nas suas aplicações; Vindo do Canadá, *Patrick Brown* discorrerá sobre modelos estatísticos e inferência para **dados espaço-temporais em áreas**; Vindo da Arábia Saudita *Hävard Rue*, conjuntamente com *Haakon Bakka*, apresentarão **modelos espaciais e espaço-temporais** usando a **abordagem SPDE**, na moldura INLA; Vinda de França, *Liliane Bel* mostrará as novas tendências em **geoestatística espaço-temporal**; Todos os cursos se apoiarão no software R.

Os participantes (sobretudo alunos do primeiro ano de doutoramento) se desejarem podem ainda submeter um poster a uma única sessão de posters que irá acontecer, podendo naturalmente participar sem ter de o fazer.

Datas a reter são o prazo de submissão do resumo de poster, 26 de abril de 2019, o prazo de registo a preço reduzido, 31 de Maio de 2019, e o prazo final de registo, 21 de Junho. Mais informações em https://ecas2019.math.tecnico.ulisboa.pt/.

Isabel Natário

• Workshop em honra do Prof. Carlos Braumann

A 12th Workshop on Statistics, Mathematics and Computation, em Honra do Prof. Carlos Braumann, decorreu no passado mês de Novembro, na Universidade da Beira Interior e organizada em colaboração com diversos Centros de Investigação e Universidades; em dois dias e com um programa que incluiu Conferências Plenárias e Sessões com Comunicações Orais e Posters. Carlos Braumann, da Universidade de Évora; João Tiago Mexia, da Universidade Nova de Lisboa; Karl Moder da Áustria; Christos P. Kitsos da Grécia; Viktor Witkovsky da Eslováquia e James R. Bozeman de Malta foram os Convidados Especiais. Mais informação disponível em https://conferencesstat.wixsite.com/wsmc12

FR

• Provas de Agregação – José Gonçalves Dias

No mês de fevereiro de 2019, o nosso colega José Gonçalves Dias, com o maior sucesso, terminou as Provas de Agregação na área científica de Métodos Quantitativos Aplicados à Gestão no ISCTE do Instituto Universitário de Lisboa.
O título da lição foi: *Modelos de Mistura Finita, Uma revisão*.

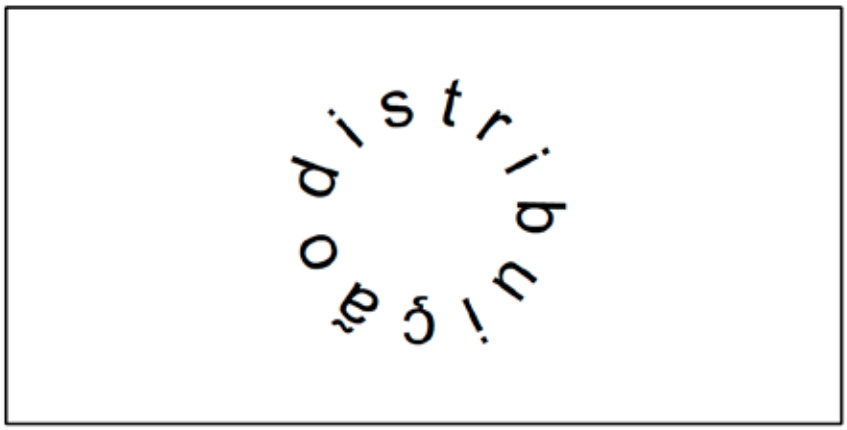
FR

• Provas de Agregação – Maria Paula Brito

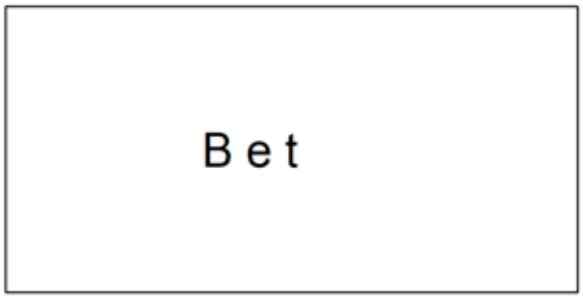
No passado mês de outubro de 2018, a nossa colega Maria Paula Brito, com o maior sucesso, terminou as Provas de Agregação na área científica de Matemática Aplicada da Universidade do Porto.
O título da lição foi: *Symbolic Data Analysis*.

FR

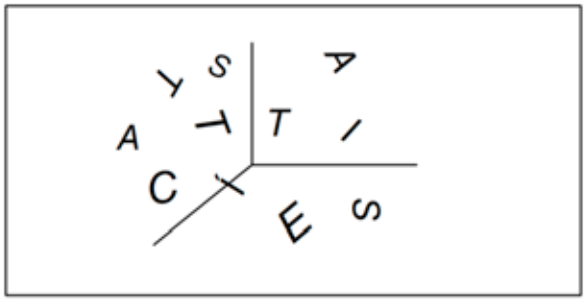
Enigmística de mefqa



No Boletim SPE outono de 2018 (p. 18):



Beta incompleta



Estatística espacial

Introdução à teoria dos operadores thinning na modelação de séries temporais de valores inteiros

Manuel G. Scotto, *manuel.scotto@tecnico.ulisboa.pt*

CEMAT e Departamento de Matemática,
Instituto Superior Técnico, Universidade de Lisboa

1 Introdução

A análise de séries temporais de valores inteiros tem suscitado um interesse crescente nas últimas décadas e, de forma mais acentuada, nos anos mais recentes. Este tipo de séries temporais surge associado, principalmente, a processos de contagem de acontecimentos, indivíduos ou objetos, sendo, portanto, de todo o interesse o desenvolvimento de métodos de modelação e análise adequados à natureza dos dados. Exemplos deste tipo de séries temporais podem ser encontrados em várias áreas de investigação, nomeadamente em economia e finanças (Costa et al., 2018; Quoreshi, 2014; Brännäs e Quoreshi, 2010), medicina e saúde pública (Sørensen, 2019; Fernández-Fontelo et al., 2016; Rao e McCabe, 2016), genética (Gouveia et al., 2017), ciências sociais (Monteiro et al., 2010; McCabe e Martin, 2005), turismo (Brännäs e Nordström, 2006) e na análise de processos geofísicos e ambientais (Santos et al., 2019; Livsey et al., 2018; Chen e Lee, 2017; Monteiro et al., 2015).

Até há relativamente pouco tempo, este tipo de séries temporais foram analisadas como se o seu suporte fosse o conjunto dos números reais. Importa referir que nos casos em que as séries apresentam contagens de valores elevados, este procedimento poderá, eventualmente, funcionar pela aplicação do teorema limite central; no entanto, quando as observações apresentam valores reduzidos, ignorar a natureza dos dados pode conduzir a resultados sem grande significado. Numa tentativa de ultrapassar esta limitação, nos últimos anos têm sido propostas várias classes de modelos para descrever e caraterizar, adequadamente, a estrutura de dependência nas séries temporais de valores inteiros. A maioria dos modelos que têm surgido na literatura podem ser classificados em duas grandes classes: a classe de modelos autorregressivos médias móveis de valores inteiros (INARMA, do inglês *INteger-valued AutoRegressive Moving Average*), e a classe de modelos GARCH de valores inteiros (do inglês *Generalized AutoRegressive Conditional Heteroscedastic*) com distribuição condicional na classe das leis discretas infinitamente divisíveis. De referir que a classe de modelos INARMA pode ser considerada como uma extensão, para o caso discreto, dos modelos ARMA convencionais. É importante salientar que os modelos ARMA convencionais não são, em princípio, de grande utilidade na modelação de séries de valores inteiros uma vez que o processo de multiplicação de um escalar real por um valor real, ou inteiro, conduz à obtenção de um valor real. No entanto, uma forma de ultrapassar esta limitação é substituir a operação multiplicação em cada um dos seu termos, por uma outra operação cujo resultado seja sempre um valor inteiro. Por outro lado, torna-se também necessário a adoção de uma distribuição discreta para a sucessão das inovações.

Este artigo visa proporcionar uma pesquisa abrangente, embora não exaustiva,[1] sobre os vários operadores propostos na literatura, que possibilitam a construção de modelos INARMA inspirados nos modelos ARMA. De entre os diversos operadores propostos destaca-se a família de operadores *thinning*. O

[1] Para uma revisão mais exaustiva ver Weiß (2018), Karlis (2016) e Scotto et al. (2015).

conceito de *thinning* surge naturalmente em variáveis de contagem, sempre que num conjunto de elementos cada um deles é selecionado ou eliminado com uma certa probabilidade. É importante referir que a família de modelos INARMA mantém muitas das propriedades dos modelos ARMA e, ao mesmo tempo, apresenta-se mais rica e, consequentemente, mais diversificada, devido à natureza aleatória do operador *thinning*.

O resto do artigo está organizado da seguinte forma: na secção 2 é introduzido o operador *thinning* binomial e as suas principais propriedades, assim como as várias modificações do mesmo que têm sido propostas para tornar a classe de modelos INARMA mais flexível. Na secção 3 são apresentados os principais operadores *thinning* usados em contexto bivariado e multivariado.

2 Operadores thinning: caso univariado

2.1 Operador thinning binomial

O operador *thinning* mais popular é o operador *thinning* binomial (também conhecido como subamostragem binomial), sugerido por Steutel e Van Harn (1979) e definido como $X_\alpha \equiv \alpha \circ X := Y_1 + \cdots + Y_X$, se $X > 0$, e 0 caso contrário, sendo X uma variável aleatória discreta com suporte no conjunto $\{0,1,\ldots,n\}$ ou $\mathbb{N}_0^+$, e os Y_i's uma sucessão i.i.d. de variáveis aleatórias de Bernoulli com parâmetro $\alpha \in (0,1)$, independentes de X. A variável aleatória X_α denomina-se α-*thinning* de X, enquanto que a sucessão Y_i é conhecida por sucessão de contagem. A razão pela qual este operador é binomial tem a ver com o facto de, fixado o valor de X, a variável aleatória $X_\alpha | X \sim \text{Bi}(X,\alpha)$, isto é, segue uma distribuição Binomial com parâmetros X e α. A interpretação do operador *thinning* binomial é bastante simples: considere-se uma população que tem X elementos, sendo que, a probabilidade de qualquer um dos elementos possuir uma determinada caraterística é α. Se os indivíduos dessa população possuem essa caraterística de forma independente uns dos outros então o número de elementos da população que possui essa caraterística é dado por X_α. Nestas condições, obviamente $X_\alpha \leq_{\text{st}} X$.[2] As principais propriedades do operador *thinning* binomial são apresentadas na Tabela 1.

1.	$0 \circ X = 0$; $1 \circ X = X$;
	Sejam, $\alpha, \beta \in [0,1]$;
2.	$\alpha \circ \beta \circ X \overset{d}{=} \beta \circ \alpha \circ X$;
3.	$\alpha \circ (\beta \circ X) \overset{d}{=} (\alpha\beta) \circ X$;
4.	$\alpha \circ (X+Z) \overset{d}{=} \alpha \circ X + \alpha \circ Z$, (com sucessões de contagem independentes);
5.	$E[\alpha \circ X] = \alpha E[X]$; $E[(\alpha \circ X)Z] = \alpha E[XZ]$;
6.	$E[(\alpha \circ X)^2] = \alpha^2 E[X^2] + \alpha(1-\alpha)E[X]$;
7.	$V[\alpha \circ X] = \alpha^2 V[X] + \alpha(1-\alpha)E[X]$;
8.	$Cov[\alpha \circ X, X] = \alpha V[X]$;
9.	$P_{\alpha \circ X}(s) := E[s^{\alpha \circ X}] = P_X(1-\alpha+\alpha s)$
10.	$\alpha \circ X + \beta \circ X \overset{d}{\neq} (\alpha+\beta) \circ X$;
11.	$\alpha \circ \max(X,Z) \overset{d}{\neq} \max(\alpha \circ X, \alpha \circ Z)$, (para *thinnings* independentes)

Tabela 1: Algumas propriedades do operador *thinning* binomial.

[2] O operador "$\leq_{\text{st}}$"define-se da seguinte forma: $Z \leq_{\text{st}} Y$ se e só se $P(Z > x) \leq P(Y > x)$, $\forall x \in \mathbb{N}_0$.

O operador *thinning* binomial partilha algumas propriedades com a multiplicação usual, como é o caso da associatividade entre parâmetros *thinning*, em termos de igualdade em distribuição, e também das propriedades relativas a momentos de primeira ordem. No entanto, a multiplicação usual goza da propriedade distributiva da soma de escalares relativamente à multiplicação com uma variável aleatória X, em termos de igualdade em distribuição, propriedade esta que deixa de ser válida quando a multiplicação é substituída pelo operador *thinning* binomial. Note-se também que este operador introduz um termo acrescido na variância, dado por $\alpha(1-\alpha)E[X]$. Este termo corresponde à variância da variável aleatória $\text{Bi}(E[X],\alpha)$. É também importante salientar que, em geral, os momentos de ordem superior a um, que envolvem a operação *thinning* binomial também não são iguais aos respetivos momentos quando se usa a multiplicação usual em vez do referido operador.

Uma questão que habitualmente se coloca em relação à distribuição da variável aleatória X_α é saber em que casos as distribuições de X_α e X pertencem à mesma família. Nesses casos, diz-se que a distribuição de X é fechada sob subamostragem binomial. Puig e Valero (2007) mostraram que a condição necessária e suficiente para isto acontecer é que $P_X(s) = g(\mu_X(s-1))$, sendo $g(\cdot)$ uma função real analítica e $\mu_X := E[X]$. São exemplos de distribuições fechadas sob subamostragem binomial a Poisson, a Binomial Negativa e a família de distribuições de Hermite. A distribuição Poisson truncada em zero é um exemplo de uma distribuição que não é fechada sob subamostragem binomial.

Várias modificações do operador *thinning* binomial têm sido propostas, nos últimos anos, para torná-lo mais flexível na modelação de sucessões de contagem, por forma a caraterizar adequadamente, por exemplo, a sobredispersão associada a inúmeras séries temporais de valores inteiros observadas na prática, a inclusão de covariáveis nos modelos e na análise de séries temporais com suporte (finito ou infinito) em $\mathbb{Z}$. Nas próximas subsecções far-se-á uma revisão dos principais operadores *thinning* propostos na literatura de forma a abranger, também, estas situações.

2.2 Operador thinning generalizado

Latour (1998) introduziu o operador *thinning* generalizado cuja definição é idêntica à definição do operador *thinning* binomial, embora com a diferença das variáveis Y_i's não serem necessariamente do tipo 0-1, isto é Bernoulli. Casos particulares do operador *thinning* generalizado foram propostos e analisados, por exemplo, por Ristić et al. (2009) em que as variáveis da sucessão de contagem têm distribuição Geométrica de parâmatero $\alpha/(1+\alpha)$. Neste caso a distribuição de X_α dado X é Binomial Negativa. É importante referir também que este operador não satisfaz as mesmas propriedades do operador *thinning* binomial (ver tabela 1). Por exemplo, para este operador $1\circ X \neq_d X$ e $\alpha\circ(\beta\circ X) \neq_d (\alpha\beta)\circ X$. O caso em que as variáveis Y_i's têm distribuição de Poisson foi recentemente tratado por Kirchner (2016). Um outro caso particular do operador *thinning* generalizado é o operador *thinning* estendido proposto por Zhu e Joe (2003) em que os Y_i's formam uma sucessão i.i.d. de variáveis aleatórias com a mesma distribuição que uma variável aleatória Y, com função geradora de probabilidade do tipo

$$P_Y(s) = \frac{(1-\alpha)+(\alpha-\gamma)s}{(1-\alpha\gamma)-(1-\alpha)\gamma s},\ \gamma\in(0,1],$$

com média $E[Y]=\alpha$ e variância $V[Y]=\alpha(1-\alpha)(1+\gamma)/(1-\gamma)$. Através desta operação, os autores introduziram o conceito de distribuição autodecomponível para inteiros generalizada (do inglês, GDSD *Generalized Discrete Self-Decomposable*). O operador *thinning* binomial corresponde ao caso $\gamma=0$. Zhu e Joe (2010) propuseram o operador *thinning* esperado que inclui como casos particulares o operador *thinning* biniomial, o generalizado e o estendido. De referir que o termo "esperado" tem a ver com o facto de o operador ser definido por forma a garantir que $E[X_\alpha]\leq E[X]$. Neste operador os elementos da sucessão de contagem são variáveis aleatórias i.i.d. autogeneralizadas.[3] Mais recentemente,

[3]Uma variável aleatória $Y(\alpha)$ diz-se autogeneralizada, em relação ao parâmetro α, se $P_{Y(\alpha)}(P_{Y(\alpha)}(s;\alpha);\alpha') = P_{Y(\alpha)}(s;\alpha\alpha')$, para todo $\alpha,\alpha'\in[0,1]$. É importante salientar que a variável $Y(\alpha)$ satisfaz a propriedade $Y(\alpha)\otimes Y(\alpha') \overset{d}{=}$

Borges et al. (2016) introduziram o operador *thinning* ρ-binomial em que os Y_i's têm distribuição do tipo ρ-Bernoulli.[4] O caso $\rho = 0$ corresponde à distribuição de Bernoulli. Para este operador, $X_\alpha|X$ tem distribuição ρ-Binomial. O operador de Ristić et al. (2009) é um caso particular do operador *thinning* ρ-binomial para o caso $\alpha = \rho/(1+\rho)$. O operador de Borges e coautores é bastante flexível sendo adequado na modelação de séries temporais com sobredispersão e subdispersão. Outros operadores *thinning* que permitem modelar sobredispersão e subdispersão são o operador *thinning* BiNB introduzido por Bourguignon e Weiß (2017), que pode ser pensado como sendo a convolução do operador *thinning* binomial e do operador de Ristić et al. (2009), e o operador *thinning* esperado biparamétrico (Aly e Bouzar, 2018) em que a função geradora de probabilidades dos Y_i's é do tipo

$$P_Y(s) = 1 - m\frac{1-s}{1+r(1-s)},\ s \in [0,1],\ r \geq 0,\ 0 < m \leq r+1.$$

O operador *thinning* esperado biparamétrico é muito geral e inclui, como casos particulares, o operador *thinning* binomial ($r = 0, m = \alpha \in (0,1]$), o operador *thinning* estendido ($r = \gamma(1-\alpha)/(1-\gamma), m = \alpha$), o operador *thinning* de Ristić et al. (2009) em que $m = r = \alpha$, o operador *thinning* ρ-binomial ($r = \rho, m = \alpha(1+\rho)$) e o operador *thinning* BiNB ($r = \beta > 0, m = \alpha + \beta$), entre outros.

2.3 Operador thinning estocástico

Uma das limitações do operador *thinning* binomial é assumir que o parâmetro α é o mesmo para todas as variáveis aleatórias da sucessão de contagem. No entanto, em muitas situações práticas, a probabilidade dos elementos de uma população possuírem uma determinada caraterística pode variar de elemento para elemento, pode também depender de um conjunto de covariáveis ou pode ser aleatório. De forma a lidar com esta última situação, Gomes e Canto e Castro (2009) propuseram a seguinte extensão do operador *thinning* generalizado (denotada por $\bullet^G$) que opera sobre as variáveis aleatórias α e X, e é definido como $\alpha \bullet^G X|\alpha, X \sim G$, sendo α uma variável aleatória com suporte em $\mathbb{R}^+$ e G uma dada distribuição do tipo discreto de média $\mu = \alpha \cdot X$ e variância $\sigma^2 = \delta \cdot X$ finita, sendo $\delta \equiv \delta(\alpha, X)$ mensurável em $\mathbb{R}^+$. Escolhas possíveis para a distribuição G são:

- $G \equiv Bi(X, \alpha) \Longrightarrow \delta = \alpha(1-\alpha)$;
- $G \equiv BN(X, \frac{1}{1+\alpha}) \Longrightarrow \delta = \alpha(1+\alpha)$;
- $G \equiv Po(\alpha X) \Longrightarrow \delta = \alpha$;
- $G \equiv Ge(\frac{1}{\alpha X}) \Longrightarrow \delta = \alpha(\alpha X - 1)$.

O caso em que G é Binomial foi tratado também por Zheng et al. (2007). Weiß e Kim (2014) analisaram o caso em que a distribuição G é Beta-Binomial. Este último operador é útil na análise de séries de contagem com suporte finito. O caso em que o suporte da variável aleatória α é o intevalo $(0,1)$ foi tratado por Hall et al. (2010) e Roitershtein e Zhong (2013). Por outro lado, Leonenko et al. (2007) propuseram uma outra extensão do operador *thinning* binomial, denotado por operador *thinning* binomial misto, em que a probabilidade de sucesso da distribuição Bernoulli é definida como o produto de uma variável aleatória $Z \in (0,1)$ e uma constante $k \in (0,1)$, isto é $\alpha = Z \cdot k$. Os autores consideraram o caso em que Z tem distribuição Beta.

$Y(\alpha\alpha')$, para $0 \leq \alpha, \alpha' \leq 1$ (*closure property*), sendo "$\otimes$" o operador *thinning* esperado.

[4] Uma variável aleatória X tem distribuição ρ-Bernoulli se a função geradora de probabilidades associada for

$$P_X(s) = \frac{1-(1-s)[\alpha(1+\rho)-\rho]}{1+\rho(1-s)},\ |s| < 1,\ \alpha \in [0,1],\ \rho \in [0,1).$$

2.4 Operador thinning com estrutura de dependência

É importante salientar que em todos os operadores apresentados anteriormente assume-se que as variáveis da sucessão de contagem são independentes. No entanto, em muitas situações práticas tal imposição é muito restritiva. Para ultrapassar esta limitação, Brännäs e Hall (2001) propuseram uma extensão do operador *thinning* binomial em que as variáveis Y_i's apresentam várias estruturas de dependência. Ristić et al. (2013) consideraram a seguinte representação para as variáveis da sucessão de contagem $Y_i = (1-V_i)W_i + V_iZ$, sendo (W_i) e (V_i) sucessões i.i.d. de variáveis aleatórias de Bernoulli de parâmetros $\alpha \in [0,1]$ e $\theta \in [0,1]$ independentes entre si, e independentes da variável aleatória $Z \sim Be(\alpha)$. Esta representação implica que os Y_i's são variáveis aleatórias de Bernoulli dependentes com parâmetro $\alpha \in [0,1]$, sendo a $Corr(Y_r, Y_s) = \theta^2 \neq 0$ para $\theta \neq 0$ e $r \neq s$. O caso $\theta = 0$ corresponde ao operador *thinning* binomial. A representação para Y_i sugerida por Ristić e coautores guarda relação com o operador de Pegram a partir do qual Jacobs e Lewis (1983) e Biswas e Song (2009) propuseram uma metodologia unificada para construir modelos ARMA discretos. Mais recentemente, Shirozhan e Mohammadpour (2018) e Khoo et al. (2017) introduziram novas classes de modelos autorregressivos para séries de contagem, definidos a partir deste operador. O operador de Pegram define uma variável aleatória $Z = (\psi, U) \star (1-\psi, V)$ onde U e V são duas variáveis aleatórias discretas independentes e $\psi \in (0,1)$, sendo a função massa de probabilidade desta nova variável dada por

$$P(Z = x) = \psi P(U = x) + (1-\psi)P(V = x),\ x = 0, 1, 2, \ldots.$$

Uma outra forma de dependência é considerar que o parâmetro *thinning* depende de um conjunto de covariáveis, ou da própria variável X. Por exemplo, Rao e McCabe (2016), Monteiro et al. (2008), Brännäs e Hellström (2001) e Brännäs (1995) adotaram a seguinte especificação para α:

$$\alpha = 1/[1 + \exp(\boldsymbol{x}\boldsymbol{\omega})],$$

sendo $\boldsymbol{x}$ e $\boldsymbol{\omega}$ vetores de covariáveis fixas e parâmetros desconhecidos, respetivamente.

Operadores *thinning* com probabilidade de sucesso dependendo do número de elementos da população também têm sido propostos. Por exemplo, Monteiro et al. (2012) propuseram uma extensão do operador *thinning* binomial para análisar séries temporais por limiares. Os autores introduziram a seguinte definição para α, num modelo com dois regimes:

$$\alpha = \begin{cases} \alpha_1, & X \leq r \\ \alpha_2, & X > r \end{cases},$$

sendo r uma constante positiva. Möller et al. (2016) também consideram este operador, embora para a análise de séries temporais por limiares com suporte finito. Por outro lado, Gouveia et al. (2018) introduziram um novo operador *thinning* (operador de variação binomial) para modelar séries temporais com suporte finito que apresentam variação extra binomial. Neste operador *thinning*, a variável aleatória X toma valores no suporte $\{0, 1, \ldots, n\}$ sendo a distribuição do operador, dado X, Binomial de parâmetros n e X/n.

2.5 Operador thinning sinalizado

Uma das limitações do operador *thinning* binomial e das suas várias modificações acima referidas é o facto de poderem ser utilizados, unicamente, na modelação de séries de contagem de valores não negativos. No caso de ter que se lidar com séries de contagem que apresentem valores inteiros negativos, Kim e Park (2008) propuseram o operador *thinning* binomial sinalizado, em que $|\alpha| \in (0,1)$, X é uma variável aleatória com suporte em $\mathbb{Z}$ e $X_\alpha = \text{sgn}(\alpha)\text{sgn}(X)(Y_1 + \cdots + Y_{|X|})$, com

$$sgn(z) := \begin{cases} 1, & z \geq 0 \\ -1, & z < 0 \end{cases},$$

e sendo os Y_i's variáveis aleatórias i.i.d. de Bernoulli com parâmetro $|\alpha|$. O operador *thinning* binomial obtém-se no caso em que $P(X \geq 0) = 1$ e $\alpha \geq 0$. Zhang et al. (2010) propuseram a seguinte extensão do operador proposto por Kim e Park (2008) em que as variáveis da sucessão (Y_i) não têm, necessariamente, distribuição Bernoulli. Os autores consideraram o caso em que as variáveis da sucessão de contagem têm distribuição em série de potência generalizada. A esta classe pertencem, por exemplo, a distribuição Binomial, a Binomial Negativa e a Poisson, entre outras. Mais recentemente, Kachour e Truquet (2011) introduziram o operador *thinning* sinalizado generalizado, em que $F \circ X := \text{sgn}(X)(Y_1 + \cdots + Y_{|X|})$. Neste caso, $Y_i \sim F$. A vantagem deste operador em relação aos anteriores é não ser fixada uma distribuição específica para os Y_i's.

3 Operadores thinning: caso bivariado e multivariado

Embora no caso univariado já exista, como se viu, um vasto leque de operadores *thinning*, a literatura sobre extensões para os casos bivariado e multivariado é mais escassa. Nesta secção serão introduzidas três famílias de operadores *thinning* para os casos bivariado e multivariado.

3.1 Operador thinning multinomial

A primeira proposta para definir um operador *thinning* multivariado deve-se a McKenzie (1988). Este operador é definido em duas etapas: na primeira delas, o operador *thinning* binomial é generalizado da seguinte forma $X_{\boldsymbol{\alpha}} = \sum_{i=1}^{X} \boldsymbol{Y}_i$, sendo $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m)' \in (0,1)^m$ com $\sum_{i=1}^{m} \alpha_i < 1$. Os $\boldsymbol{Y}_i$'s são vetores aleatórios i.i.d. e independentes de X, com distribuição multinomial $\text{MULT}(1; \alpha_1, \ldots, \alpha_m)$. O operador *thinning* binomial corresponde ao caso $m = 1$. Numa segunda etapa, introduz-se primeiro a matriz $\mathbf{A} \in (0,1)^{m \times m}$, contendo as colunas os $\boldsymbol{\alpha}_j$'s definidos como $\boldsymbol{\alpha}_j := (\alpha_{j1}, \ldots, \alpha_{jm})'$, e o vetor aleatório $\boldsymbol{X} = (X_1, \ldots, X_m)$. O operador *thinning* multinomial define-se como $\boldsymbol{X}_{\boldsymbol{\alpha}} \equiv \mathbf{A} \star \boldsymbol{X} := \sum_{j=1}^{m} \sum_{i=1}^{X_j} \boldsymbol{Y}_{ji}$. Todos os operadores *thinning* envolvidos em $\boldsymbol{X}_{\boldsymbol{\alpha}}$ são independentes. Este operador partilha muitas das propriedades do operador *thinning* binomial, nomeadamente o facto de $\boldsymbol{\alpha} \star (\boldsymbol{X} + \boldsymbol{Z}) =_d \boldsymbol{\alpha} \star \boldsymbol{X} + \boldsymbol{\alpha} \star \boldsymbol{Z}$, sendo $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$.

3.2 Operador thinning matricial binomial

Franke e Subba Rao (1993) introduziram o modelo INAR multivariado de primeira ordem com base numa matriz de operadores *thinning* binomiais independentes. O operador *thinning* matricial binomial define-se da seguinte forma: seja $\boldsymbol{X} = (X_1, \ldots, X_m)'$ e $\mathbf{A}$ a matriz de coeficientes

$$\mathbf{A} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mm} \end{bmatrix},$$

sendo $\mathbf{A} \in [0,1]^{m \times m}$. Então o operador *thinning* matricial binomial é definido como sendo

$$\mathbf{A} \circ \boldsymbol{X} = \begin{bmatrix} \alpha_{11} \circ X_1 + \alpha_{12} \circ X_2 + \cdots + \alpha_{1m} \circ X_m \\ \alpha_{21} \circ X_1 + \alpha_{22} \circ X_2 + \cdots + \alpha_{2m} \circ X_m \\ \vdots \\ \alpha_{m1} \circ X_1 + \alpha_{m2} \circ X_2 + \cdots + \alpha_{mm} \circ X_m \end{bmatrix},$$

onde é assumido que todas as operações *thinning* são independentes. Pedeli e Karlis (2013a, 2011) analisaram o caso diagonal, isto é, o caso em que $\alpha_{ij} = 0$ se $i \neq j$, para $i, j = 1, \ldots, m$. Em tal caso

$[\mathbf{A} \circ \boldsymbol{X}]_i =_d \alpha_{ii} \circ X_i$, ou seja, as distribuições marginais são as mesmas que no caso univariado. No entanto, esta propriedade implica que a correlação cruzada ente $[\mathbf{A} \circ \boldsymbol{X}]_i$ e $[\mathbf{A} \circ \boldsymbol{X}]_j$ para $i \neq j$, seja nula, o que, em muitos casos, é muito restritivo. O caso em que a correlação cruzada não é nula foi estudado em pormenor por Franke e Subba Rao (1993), Boudreault e Charpentier (2011) e Pedeli e Karlis (2013b). Neste caso, as distribuições marginais já não são as mesmas que no caso univariado.

3.3 Operador thinning binomial bivariado

A generalização do operador *thinning* binomial para o caso bivariado foi proposta por Scotto et al. (2014). Estes autores introduziram o operador *thinning* binomial bivariado cuja definição é a seguinte: para o vetor aleatório $\boldsymbol{X} = (X_1, X_2)'$ e $\boldsymbol{\alpha}$ o vetor de parâmetros $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \phi_\alpha)$ com $0 < \alpha_1, \alpha_2 < 1$, e $|\phi_\alpha| \leq 1$, o operador *thinning* binomial bivariado entre $\boldsymbol{X}$ e $\boldsymbol{\alpha}$ define-se como

$$\boldsymbol{\alpha} \otimes \boldsymbol{X} \mid \boldsymbol{X} \sim \text{BVB}_{\text{II}}\left(X_1, X_2, \min\{X_1, X_2\}; \alpha_1, \alpha_2, \phi_\alpha\right),$$

isto é, $\boldsymbol{\alpha} \otimes \boldsymbol{X} \mid \boldsymbol{X}$ segue uma distribuição binomial bivariada de tipo II. Este operador apresenta um conjunto de caraterísticas importantes, nomeadamente o facto de as distribuições condicionais serem binomiais, e a dependência entre as duas componentes de $\boldsymbol{\alpha} \otimes \boldsymbol{X} \mid \boldsymbol{X}$ poder ser positiva ($\phi_\alpha > 0$) ou negativa ($\phi_\alpha < 0$).

Referências

[1] Aly, E-E.A.A., Bouzar, N. (2018). Expectation thinning operators based on linear fractional probability generating functions. *Journal of the Indian Society for Probability and Statistics* (no prelo).

[2] Biswas, A., Song, P.X.K. (2009). Discrete-valued ARMA processes. *Statistics and Probability Letters* **79**, 1884–1889.

[3] Borges, P., Fajardo Milinares, F., Bourguignon, M. (2016). A geometric time series model with inflated-parameter Bernoulli counting series. *Statistics and Probability Letters* **119**, 264–272.

[4] Boudreault, M., Charpentier, A. (2011). Multivariate integer-valued autoregressive models applied to earthquake counts. *arXiv:1112.0929v1 [stat.AP]*.

[5] Bourguignon, M., Weiß, C.H. (2017). An INAR(1) process for modeling count time series with equidispersion, underdispersion and overdispersion. *Test* **26**, 847–868.

[6] Brännäs, K. (1995). Explanatory variables in the AR(1) count data model. *Umeå Economic Studies* 381.

[7] Brännäs, K., Hall, A. (2001). Estimation in integer-valued moving average models. *Applied Stochastic Models in Business and Industry* **17**, 277–291.

[8] Brännäs, K., Hellström, J. (2001). Generalized integer-valued autoregression. *Econometric Reviews* **20**, 425–443.

[9] Brännäs, K., Nordström, J. (2006). Tourist accommodation effects of festivals. *Tourism Economics* **12**, 291–302.

[10] Brännäs, K., Quoreshi, A.M.M.S. (2010). Integer-valued moving average modelling of the number of transactions in stocks. *Applied Financial Economics* **20**, 1429–1440.

[11] Chen, C.W.S., Lee, S. (2017). Bayesian causality test for integer-valued time series models with applications to climate and crime data. *Journal of the Royal Statistical Society, Series C* **66**, 797–814.

[12] Costa, C., Pereira, I., Scotto, M.G. (2018). Surveillance in discrete time series. In: Oliveira TA, Kitsos CP, Oliveira A, Grilo L (eds) *Recent Studies on Risk Analysis and Statistical Modeling*. Springer International Publishing, pp 197-212.

[13] Fernández-Fontelo, A., Cabaña, A., Puig, P., Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine* **35**, 4875–4890.

[14] Franke, J., Subba Rao, T. (1993). Multivariate first-order integer-valued autoregression. *Technical report*, Universität Kaiserslautern.

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

[16] Gouveia, S., Möller, T.A., Weiß, C.H., Scotto, M.G. (2018). A full ARMA model for counts with bounded support and its application to rainy-days time series. *Stochastic Environmental Research and Risk Assessment* **32**, 2495–2514.

[17] Gouveia, S., Scotto, M.G., Weiß, C.H., Ferreira, P.J.S.G. (2017). Binary auto-regressive geometric modelling in a DNA context. *Journal of the Royal Statistical Society, Series C* **66**, 253–271.

[18] Hall, A., Scotto, M.G., Cruz, J.P. (2010). Extremes of integer-valued moving average sequences. *Test* **19**, 359–374.

[19] Jacobs, P.A., Lewis, P.A.W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis* **4**, 19–36.

[20] Kachour, M., Truquet, L. (2011). A p-order signed integer-valued autoregressive (SINAR(1)) model. *Journal of Time Series Analysis* **32**, 223–236.

[21] Karlis, D. (2016). Modelling multivariate times series for counts. In: Davis RA, Holan SH, Lund R, Ravishanker N (eds) *Handbook of Discrete-valued Time Series.* CRC Press, Boca Raton, pp 407-424.

[22] Khoo, W.C., Ong, S.H., Biswas, A. (2017). Modeling time series of counts with a new class of INAR(1) model. *Statistical Papers* **58**, 393–416.

[23] Kim, H.Y., Park, Y. (2008). A non-stationary integer-valued autoregressive model. *Statistical Papers* **49**, 485–502.

[24] Kirchner, M. (2016). Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications* **126**, 2494–2525.

[25] Latour, A. (1998). Existence and stochastic structure of a non-negative integer-valued autoregressive processes. *Journal of Time Series Analysis* **4**, 439–455.

[26] Leonenko, N.N., Savani, V., Zhigljavsky, A.A. (2007). Autoregressive negative binomial processes. *Annales de l'I.S.U.P* **LI**, 25–47.

[27] Livsey, J., Lund, R., Kechagias, S., Pipiras, V. (2018). Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts. *The Annals of Applied Statistics* **12**, 408–431.

[28] McCabe, B.P.M., Martin, G.M. (2005). Bayesian prediction of low count time series. *International Journal of Forecasting* **21**, 315–330.

[29] McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Advances in Applied Probability* **20**, 822–835.

[30] Möller, T.A., Silva, M.E., Weiß, C.H., Scotto, M.G., Pereira, I. (2016). Self-exciting threshold binomial autoregressive processes. *AStA Advances in Statistical Analysis* **100**, 369–400.

[31] Monteiro, M., Pereira, I., Scotto, M.G. (2008). Optimal alarm systems for count processes. *Communications in Statistics - Theory and Methods* **37**, 3054–3076.

[32] Monteiro, M., Scotto, M.G., Pereira, I. (2010). Integer-valued autoregressive processes with periodic structure. *Journal of Statistical Planning and Inference* **140**, 1529–1541.

[33] Monteiro, M., Scotto, M.G., Pereira, I. (2011). Integer-valued self-exciting threshold autoregressive processes. *Communications in Statistics - Theory and Methods* **41**, 2717–2737.

[34] Monteiro, M., Scotto, M.G., Pereira, I. (2015). A periodic bivariate integer-valued autoregressive model. In: Bourguignon JP, Jelstch R, Pinto A, Viana M (eds) *Dynamics, Games and Science - International Conference. Advanced School Planet Earth DGS II.* Springer, Switzerland, pp 455-477.

[35] Pedeli, X., Karlis, D. (2011). A bivariate INAR(1) process with application. *Statistical Modelling* **11**, 325–349.

[36] Pedeli, X., Karlis, D. (2013a). On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis* **34**, 206–220.

[37] Pedeli, X., Karlis, D. (2013b). Some properties of multivariate INAR(1) processes. *Computational Statistics and Data Analysis* **67**, 213–225.

[38] Puig, P., Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli* **13**, 544–555.

[39] Quoreshi, A.M.M.S. (2014). A long-memory integer-valued time series model, INARFIMA, for financial application. *Quantitative Finance* **14**, 225–2235.

[40] Rao, Y., McCabe, B.P.M. (2016). Real-time surveillance for abnormal events: the case of influenza outbreaks. *Statistics in Medicine* **35**, 2206–2220.

[41] Ristić, M.M., Bakouch, H.S., Nastić, A.S. (2009). A new geometric first-order integer-valued autoregressive (NGINAR(1)) process. *Journal of Statistical Planning and Inference* **139**, 2218–2226.

[42] Ristić, M.M., Nastić, A.S., Miletić Ilić, A.V. (2013). A geometric time series model with dependent Bernoulli counting series. *Journal of Time Series Analysis* **34**, 466–476.

[43] Roitershtein, A., Zhong, Z. (2013). On random coefficient INAR(1) processes. *Science China Mathematics* **56**, 177–200.

[44] Santos, C., Pereira, I., Scotto, M.G. (2019). On the theory of periodic multivariate INAR processes. (submetido para publicação)

[45] Scotto, M.G., Weiß, C.H., Gouveia, S. (2015) Thinning-based models in the analysis of integer-valued time series: a review. *Statistical Modelling* **15**, 590–618.

[46] Scotto, M.G., Weiß, C.H., Silva, M.E., Pereira, I. (2014). Bivariate binomial autoregressive models. *Journal of Multivariate Analysis* **125**, 233–251.

[47] Shirozhan, M., Mohammadpour, M. (2018). A new class of INAR(1) model for count time series. *Journal of Statistical Computation and Simulation* **88**, 1348–1368.

[48] Sørensen, H. (2019). Independence, successive and conditional likelihood for time series of counts. *Journal of Statistical Planning and Inference* **200**, 20–31.

[49] Steutel, F.W., van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability* **7**, 893–899.

[50] Weiß, C.H. (2018). *An Introduction to Discrete-valued Time Series.* John Wiley & Sons, Inc., Chichester.

[51] Weiß, C.H., Kim, H.-Y. (2014). Diagnosing and modeling extra-binomial variation for time-dependent counts. *Applied Stochastic Models in Business and Industry* **30**, 588–608.

[52] Zhang, H. Wang, D., Zhu, F. (2010). Inference for INAR(p) processes with signed generalized power series thinning operator. *Journal of Statistical Planning and Inference* **140**, 667–683.

[53] Zheng, H.T., Basawa, I.V., Datta, S. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference* **173**, 212–229.

[54] Zhu, R., Joe, H. (2003). A new type of discrete self-decomposability and its applications to continuous-time Markov processes for modeling count data time series. *Stochastic Models* **19**, 235–254.

[55] Zhu, R., Joe, H. (2010). Negative binomial time series models based on expectation thinning operators. *Journal of Statistical Planning and Inference* **140**, 1874–1888.

Métodos de deteção de *outliers* baseados em *wavelets*: o caso dos modelos INAR(1) de Poisson

Isabel Silva, *ims@fe.up.pt*
Faculdade de Engenharia, Universidade do Porto & CIDMA

Maria Eduarda Silva, *mesilva@fep.up.pt*
Faculdade de Economia, Universidade do Porto & CIDMA

1 Introdução

A presença de *outliers* ou observações discordantes nas séries temporais, assim como noutros tipos de dados, pode provocar efeitos adversos na identificação do modelo e na estimação dos parâmetros. Ignorar estas observações díspares pode mascarar a presença de importantes fenómenos subjacentes, impedindo a análise de risco. Várias metodologias para detetar e estimar *outliers* e outros efeitos de intervenção têm sido estabelecidas para modelos ARMA e alguns modelos de séries temporais não lineares, ver por exemplo Chang *et al*. [6], Chen e Liu [7] e Tsay [18]. As abordagens propostas baseiam-se principalmente em métodos iterativos, estimadores robustos e estatísticas baseadas na razão de verosimilhanças.

Segundo Fox [8], no contexto das séries temporais, podem ser considerados dois tipos de *outliers*: os aditivos e os inovacionais. Os *outliers* aditivos são erros externos ou mudanças exógenas que ocorrem num determinado instante e por isso só afetam a observação correspondente ao tempo onde a perturbação acontece. Os *outliers* inovacionais estão associados a mudanças internas ou efeitos endógenos no processo do ruído, afetando as observações subsequentes.

Nos últimos 20 anos, métodos baseados em *wavelets* foram propostos para resolver o problema de deteção da localização de *outliers* em modelos lineares e não lineares (Bilen e Huzurbazar [4], Grané e Veiga [9]). *Wavelets* são funções que combinam propriedades como localização no tempo e em escala, ortonormalidade, diferentes graus de suavidade, suporte compacto e implementação rápida (ver Percival e Walden [12]). A Transformada *Wavelet* Discreta (TWD) é uma poderosa ferramenta para a análise multi-resolução em tempo-escala e consiste na a aplicação de filtros de diferentes frequências de corte usados para analisar um sinal em diferentes escalas. Os coeficientes da TWD são capazes de revelar mudanças na variância e de nível, assim como outros tipos de descontinuidades nos dados.

Apesar destes desenvolvimentos, a deteção e estimação de *outliers* no contexto das séries temporais de contagem tem recebido menos atenção. A análise de séries temporais de contagem tem-se tornado uma área de investigação ativa nos últimos anos. Estas séries caracterizam-se por apresentar valores baixos ou nulos, assimetrias e sobre-dispersão e podem ser encontradas nas mais variadas áreas de aplicação. Uma das abordagens mais populares propostas para analisar estes tipos de dados consiste na utilização de modelos baseados numa operação aleatória chamada *thinning*, que conjugada com inovações discretas, permite preservar a natureza discreta das contagens. Assim surge a família de modelos auto-regressivos e de médias móveis de valor inteiro, INARMA, que tem sido amplamente estudada na literatura (ver Scotto *et al*. [13]). Barczy *et al*. [2, 3] propuseram um método dos Mínimos Quadrados Condicionais para a estimação dos parâmetros do modelo INAR(1) contaminado com *outliers* aditivos ou inovacionais, assumindo que os tempos de ocorrência do *outlier* são conhecidos mas os seus tamanhos desconhecidos. Adicionalmente, Silva e Pereira [16] sugeriram uma abordagem Bayesiana para detetar *outliers* aditivos em modelos PoINAR(1), isto é, INAR(1) de Poisson. Recentemente, Bourguignon e Vasconcellos [5]

usaram métodos baseados em características e sinais para obter estimadores robustos dos parâmetros do modelo INAR(1) na presença de *outliers* aditivos.

Neste trabalho são apresentados dois métodos baseados em *wavelets*, propostos por Silva e Silva [14], que permitem a identificação dos tempos de ocorrência dos *outliers*, aditivos ou inovacionais, isolados, múltiplos ou em grupo (*patches*), em modelos PoINAR(1). No primeiro método, adaptado de Grané e Veiga [9], os chamados coeficientes de detalhe obtidos pela aplicação da TWD, usando a *wavelet* de Haar, são comparados com um limiar. No segundo, o método de reamostragem paramétrica de Tsay [19] é usado para obter a distribuição empírica dos tais coeficientes de detalhe.

O resto do trabalho está organizado da seguinte forma: a Transformada *Wavelet* Discreta é brevemente descrita na Secção 2. A Secção 3 apresenta os modelos PoINAR(1) contaminados com *outliers* aditivos e inovacionais. Os procedimentos para detetar o tempo de ocorrência de *outliers* nestes modelos são descritos na Secção 4 e ilustrados num conjunto de dados reais na Secção 5. A Secção 6 contém algumas observações finais.

2 Transformada *wavelet* discreta

Uma *wavelet* é uma função que pode ser vista como uma pequena onda que cresce e decresce num período de tempo limitado (mais detalhes em Percival e Walden [12]). A análise *wavelet* usa versões com diferentes escalas e deslocamentos da chamada *wavelet* mãe, de modo a fornecer a localização no tempo de cada componente espectral de modo semelhante à análise de Fourier, onde funções sinusoidais são usadas para encontrar as componentes em frequência que compõem um sinal.

Seguindo o trabalho de Percival e Walden [12], seja $\mathbf{X} = \{X_t, t = 0, \ldots, N-1\}$ uma série temporal, com $N = 2^J$, $J \in \mathbb{N}$. Os coeficientes da TWD $\mathbf{W} = \{W_n, n = 0, \ldots, N-1\}$ definem-se por:

$$\mathbf{W} = \mathcal{W}\,\mathbf{X} \quad \Leftrightarrow \quad [\mathbf{W}_1 \ \ldots \mathbf{W}_J \ \mathbf{V}_J]^T = [\mathcal{W}_1 \ \ldots \mathcal{W}_J \ \mathcal{V}_J]^T \mathbf{X},$$

onde $\mathcal{W}$ é uma matriz ortonormal, de dimensão $N \times N$, de versões dilatadas e deslocadas da *wavelet* mãe $\psi(\cdot)$, definidas através de $\frac{1}{\sqrt{d}}\psi\left(\frac{u-t}{d}\right)$ sendo d o parâmetro de dilatação e t o parâmetro de deslocamento, em que $d = 2^j$ e $t = k2^j$, para $j,k \in \mathbb{Z}$.

É possível reconstruir a série temporal à custa da TWD inversa, através de $\mathbf{X} = \mathcal{W}^T\mathbf{W} = \sum_{j=1}^{J} \mathcal{W}_j^T \mathbf{W}_j + \mathcal{V}_J^T \mathbf{V}_J$.

Na prática, a matriz dos coeficientes da TWD, $\mathbf{W}$, calcula-se através do chamado algoritmo piramidal proposto por Mallat [10], baseado em filtragens. Especificamente, para uma largura par L, consideram-se o filtro *wavelet* $\{h_l : l = 0, \ldots, L-1\}$ (filtro passa-alta) e o filtro de escala $g_l = (-1)^{l+1} h_{L-1-l}$ (filtro passa-baixas). No primeiro passo do algoritmo, dois conjuntos de coeficientes são produzidos pela convolução de $\mathbf{X}$ com $\{g_l\}$ (coeficientes de aproximação do primeiro nível $c\mathbf{A}_1$) e com $\{h_l\}$ (coeficientes de detalhe do primeiro nível $c\mathbf{D}_1$); posteriormente é realizada uma decimação (*downsample*), retendo unicamente as observações filtradas intercaladas. No próximo passo, repete-se o procedimento anterior substituindo $\mathbf{X}$ por $c\mathbf{A}_1$ de modo a obter $c\mathbf{A}_2$ e $c\mathbf{D}_2$. Desta forma, no nível j, a decomposição de $\mathbf{X}$ tem a seguinte estrutura $[c\mathbf{A}_j, c\mathbf{D}_j, c\mathbf{D}_{j-1}, \ldots, c\mathbf{D}_1]$.

Os coeficientes de detalhe capturam certas características da série temporal, como mudanças repentinas ou picos, apresentando valores elevados na presença dessas singularidades, podendo então ser usados para detetar *outliers*.

A *wavelet* mãe escolhida neste trabalho foi a *wavelet* de Haar, que pode ser considerada como uma onda quadrada, sendo assim apropriada para séries de contagem. Está definida por:

$$\psi(t) = \begin{cases} -1/\sqrt{2}, & -1 \le t \le 0 \\ 1/\sqrt{2}, & 0 < t \le 1 \\ 0, & \text{caso contrário,} \end{cases}$$

e neste contexto, os filtros passa-baixa correspondem a médias móveis das observações enquanto que os filtros passa-alta correspondem a diferenças móveis das observações.

3 Modelos INAR(1) de Poisson contaminados com *outliers*

Séries temporais de contagem aparecem nos mais variados domínios de investigação, nomeadamente nas ciências sociais, biologia, economia e finanças, telecomunicações, seguros, entre outros. Uma das abordagens propostas na literatura para analisar este tipo de dados é o modelo auto-regressivo de valor inteiro de primeira ordem, INAR(1), proposto independentemente por Al-Osh e Alzaid [1] e McKenzie [11]. Este modelo usa a operação *thinning* binomial proposta por Steutel e Van Harn [17], definida por $\alpha \circ X = \sum_{k=1}^{X} Y_k$, onde X é uma variável aleatória de valor inteiro não negativo, $\alpha \in [0,1]$ e $\{Y_k\}$, $k = 1,\ldots,X$, é uma sequência de variáveis aleatórias de Bernoulli independentes e identicamente distribuídas, independente de X, chamada série de contagem. Note-se que $\alpha \circ X|X \sim \text{Bi}(X,\alpha)$ (mais detalhes e propriedades desta operação podem ser encontrados em Silva e Oliveira [15]).

O processo estocástico em tempo discreto de valor inteiro não negativo, $\{X_t\}$, é um processo INAR(1) de Poisson, denotado abreviadamente por PoINAR(1), se satisfaz a seguinte equação:

$$X_t = \alpha \circ X_{t-1} + e_t, \tag{1}$$

onde $e_t \sim \text{Poisson}(\lambda)$, é o processo de chegada, $0 < \alpha < 1$, e para cada t, todas as séries de contagem $\alpha \circ X_{t-1}$ são mutuamente independentes e independentes de $\{e_t\}$. Nestas condições, o processo é estritamente estacionário e $X_t \sim \text{Poisson}(\frac{\lambda}{1-\alpha})$ quando $X_0 \sim \text{Poisson}(\frac{\lambda}{1-\alpha})$.

Diz-se que o processo PoINAR(1) está contaminado com $I \in \mathbb{N}$ *outliers* aditivos (AO), que ocorrem nos tempos $s_i \in \mathbf{N}$, e têm magnitude $\omega_i \in \mathbb{N}$ para $i = 1,\ldots,I$, se

$$Y_t = X_t + \sum_{i=1}^{I} \delta_{i,s_i}\omega_i,$$

onde X_t é um modelo PoINAR(1), que satisfaz (1), e $\delta_{k,m}$ é uma função indicadora ($\delta_{k,m} = 1$, se $k = m$; $\delta_{k,m} = 0$, se $k \neq m$). Por outro lado, o processo PoINAR(1) diz-se contaminado com $I \in \mathbb{N}$ *outliers* inovacionais (IO), com magnitude ω_i e ocorrendo nos tempos s_i, $i = 1,\ldots,I$, se

$$Y_t = \alpha \circ Y_{t-1} + \eta_t,$$

com $\eta_t = e_t + \sum_{i=1}^{I} \delta_{i,s_i}\omega_i$, onde $e_t \sim \text{Poisson}(\lambda)$ e $\delta_{k,m}$ está definida como anteriormente.

Note-se que em ambos os casos, o processo sem *outliers*, X_t, não é observado.

4 Procedimentos para detetar o tempo de ocorrência de *outliers*

Dois procedimentos, baseados em *wavelets*, para a deteção do tempo de ocorrência de *outliers* em processos PoINAR(1) podem ser descritos através dos seguintes passos:

Passo 1 Dada uma série temporal de contagem, $\mathbf{Y} = \{Y_t, t = 0,\ldots,N\}$, ajusta-se um modelo PoINAR(1) e estimam-se os resíduos de Pearson[1] $\mathbf{Z} = \{\hat{z}_t, t = 1,\ldots,N-1\}$, onde $\hat{z}_t = \dfrac{Y_t - (\hat{\alpha}Y_{t-1} + \hat{\lambda})}{\sqrt{\hat{\alpha}(1-\hat{\alpha})Y_{t-1} + \hat{\lambda}}}$.

[1] $Z_t = \frac{Y_t - \text{E}[Y_t|Y_{t-1}]}{\sqrt{\text{Var}(Y_t|Y_{t-1})}}$

Passo 2 Aplica-se a TWD aos resíduos de Pearson de modo a obter os coeficientes de detalhe do primeiro nível, $c\mathbf{D}_1 = (d_1, d_2, \ldots, d_{N/2})$.

Passo 3a Abordagem pelo limiar:

(i) Define-se o limiar k_1^a (discutido na Subsecção 4.1).

(ii) Determina-se o conjunto de índices (ordenados) $\mathbf{S} = \{s_1, \ldots, s_I\}$, das posições onde os coeficientes de detalhe estão acima do limiar k_1^a. Como sugerido por Grané e Veiga [9], os *outliers* são detetados de forma recursiva. Assim, $\mathbf{Z}$ é reconstruído a partir da TWD inversa aplicada aos coeficientes de detalhe modificados de forma a que o maior coeficiente de detalhe (em valor absoluto) seja substituído por zero. O procedimento termina quando não são detetados mais *outliers*.

Passo 3b Abordagem pela reamostragem paramétrica:

(i) Calcula-se o envelope de aceitação (discutido na Subsecção 4.2).

(ii) Determina-se o conjunto de índices (ordenados) $\mathbf{S} = \{s_1, \ldots, s_I\}$, das posições onde os coeficientes de detalhe estão fora do envelope de aceitação.

Passo 4 Para obter a posição exata do *outlier* na série dos resíduos, seja s um elemento genérico de $\mathbf{S}$, calcula-se a média amostral de $\mathbf{Z}$ sem as observações $2s$ e $2s-1$, isto é, $\bar{z}_{N-2} = \dfrac{1}{N-2} \sum_{i \neq 2s, 2s-1} \hat{z}_i$. Então, o tempo de ocorrência do *outlier* na série dos resíduos é $2s$ se $|\hat{z}_{2s} - \bar{z}_{N-2}| > |\hat{z}_{2s-1} - \bar{z}_{N-2}|$, ou $2s-1$ caso contrário.

Note-se que no caso de *outliers* em grupo (*patches*), os coeficientes do primeiro nível só permitem detetar o tempo em que começa, mas não a sua duração, sendo necessário recorrer aos coeficientes de detalhe do segundo nível $c\mathbf{D}_2$ (Bilen e Huzurbazar [4] e Grané e Veiga [9]). Assim, no **Passo 3a** há um limiar para cada nível, $k_1^{a_1}$ e $k_2^{a_2}$, respetivamente. Analogamente, tem que ser determinado um envelope de aceitação para cada nível no **Passo 3b**.

4.1 Definição do limiar

É sabido que os coeficientes *wavelet* de dados Gaussianos ou de um ruído branco são eles próprios Gaussianos ou ruído branco. Adicionalmente, como referido por Bilen e Huzurbazar [4] e Percival e Walden [12], os coeficientes *wavelet* em $\mathbf{W}_j$ são aproximadamente não correlacionados mesmo quando os dados são altamente correlacionados. Contudo, no contexto não Gaussiano das séries de contagem, não há resultados disponíveis para a distribuição dos coeficientes de detalhe da análise *wavelet*. Por isso, são usadas simulações de Monte Carlo para obter a distribuição empírica do máximo dos coeficientes de detalhe (em valor absoluto) para os resíduos de Pearson dos modelos PoINAR(1). O limiar é então definido da seguinte forma. Para cada par de parâmetros (α, λ) no conjunto $\{(\alpha, \lambda) : \alpha = (2k+1) \times 10^{-1}, k = 0, \ldots, 4; \lambda = 2k+1, k = 0, \ldots, 14\}$, 20000 repetições do processo PoINAR(1) correspondente são geradas para cada tamanho de amostra $N = 2^J + 1$, para $J = 7, \ldots, 10$. O modelo é ajustado, os resíduos de Pearson, $\hat{z}_i$, para $i = 1, \ldots, N-1$, são estimados e os máximos dos coeficientes de detalhe do primeiro e segundo nível são obtidos. Os limiares $k_1^{\mathrm{a}_1}$ e $k_2^{\mathrm{a}_2}$ são definidos como os $100(1-\mathrm{a})^\circ$ percentis das distribuições empíricas correspondentes, para $\mathrm{a} = \mathrm{a}_1$ ou $\mathrm{a} = \mathrm{a}_2$. Os resultados obtidos indicam que os limiares variam não apenas com o tamanho da amostra N mas também com a combinação específica dos parâmetros α e λ. Portanto, adotando uma estratégia conservadora, para cada tamanho de amostra N os limiares são definidos como o mínimo obtido para todos as combinações de parâmetros em cada nível de decomposição. Os limiares obtidos são apresentados na Tabela 1.

Tabela 1: Limiares correspondentes aos 90° e 95° percentis da distribuição empírica do máximo dos coeficientes de detalhe (de primeiro e de segundo nível), em valor absoluto, para resíduos de Pearson de modelos PoINAR(1).

N	128	256	512	1024
$k_1^{0.05}$	3.469	3.694	3.886	4.118
$k_1^{0.1}$	3.182	3.450	3.657	3.840
$k_2^{0.05}$	3.157	3.347	3.518	3.691
$k_2^{0.1}$	2.936	3.138	3.320	3.504

4.2 Determinação do envelope de aceitação

Tsay [19] propôs um método para obter a distribuição empírica de um funcional escolhido usando amostras *bootstrap* geradas a partir de um modelo ajustado, e então comparar o valor observado para a série com a distribuição empírica obtida. Para este propósito, um envelope de aceitação é calculado a partir dos $100(1-\mathtt{a}/2)^\circ$ e $100\mathtt{a}/2^\circ$ percentis desta distribuição empírica. Se o modelo ajustado for adequado, o funcional de interesse dos dados originais deve estar dentro do envelope. Neste trabalho, os funcionais de interesse são os coeficientes de detalhe de primeiro e segundo nível dos resíduos de Pearson do modelo PoINAR(1). Assim, para vários tamanhos de amostra $N = 2^J + 1, J = 7,8,9$, e valores dos parâmetros $\{(\alpha,\lambda) : \alpha \in \{0.1,0.5,0.9\}; \lambda \in \{1,5,9,13\}\}$, 20000 realizações do processo PoINAR(1) são geradas e os correspondentes resíduos de Pearson são estimados. Para cada série de resíduos de Pearson, a TWD é aplicada para obter os coeficientes de detalhe de primeiro e segundo nível, $c\mathbf{D}_1$ e $c\mathbf{D}_2$, e os envelopes de aceitação são construídos a partir dos 0.01° e $99{,}99^\circ$ percentis da distribuição empírica de $c\mathbf{D}_1$ e $c\mathbf{D}_2$, respetivamente. Mais uma vez, os resultados mostram que os envelopes de aceitação variam não apenas com o tamanho da amostra N, mas também com a combinação dos valores dos parâmetros (α,λ). Portanto, assumindo uma estratégia conservadora, para cada tamanho de amostra, é escolhido um envelope de aceitação com a amplitude mínima.

5 Ilustração

Nesta secção ilustram-se com dados reais os procedimentos de deteção do tempo de ocorrência de *outliers* descritos anteriormente. Para isso, consideramos uma série temporal com 241 observações relativas ao número de diferentes endereços IP, registados em períodos de 2 minutos, que acedem ao servidor do Departamento de Estatística da Universidade de Würzburg, entre as 10h e as 18h do dia 29 de novembro de 2005, representada na Figura 1. O valor da média amostral ($\bar{x} = 1.32$), da variância amostral ($\hat{\sigma}^2 = 1.39$) e a análise das funções de autocorrelação e autocorrelação parcial amostrais indicam que pode ser ajustado um modelo PoINAR(1) a este conjunto de dados. A aplicação dos procedimentos descritos anteriormente permitem detetar a ocorrência de um *outlier* na observação $t = 224$ (correspondendo a $S = \{112\}$). Na Figura 2 estão representados o limiar e o envelope de aceitação para este conjunto de dados. A deteção de *outlier* em $t = 224$ está de acordo com os resultados de Weiß[20] e Silva e Pereira [16]. A estimativa CLS proposto por Barczy *et al.* [3] para o tamanho do *outlier* é $\hat{\omega} = 6.79$, o que significa que o valor verdadeiro da observação 224 é aproximadamente igual a 1. Note-se que o trabalho de Weiß[20] indica que o valor verdadeiro é $X_{224} = 1$ enquanto que Silva e Pereira [16] detetam um *outlier* em $t = 224$ com probabilidade 0.99 e estimativas dos parâmetros dadas por $\hat{\alpha} = 0.27$, $\hat{\lambda} = 0.89$ e $\omega = 7$.

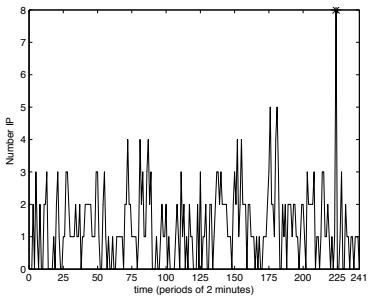


Figura 1: Cronograma da série temporal IP.

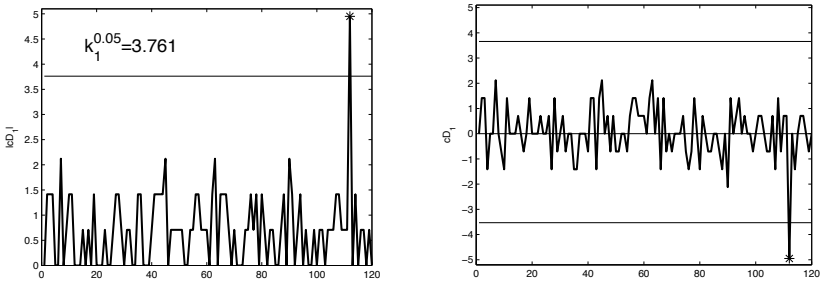


Figura 2: Resultados da deteção do tempo de ocorrência do *outlier* na série temporal IP, com a abordagem do limiar (lado esquerdo) e com a abordagem da reamostragem paramétrica (lado direito).

6 Observações finais

As metodologias apresentadas para a deteção do tempo de ocorrência de *outliers* em modelos PoI-NAR(1) não requerem o conhecimento prévio do número de *outliers* e adequam-se aos casos de um único ou múltiplos *outliers*, do tipo aditivo ou inovacional, assim como em grupo (*patches*). Contudo, a discriminação do tipo de *outlier* é ainda um tema em aberto.

Adicionalmente, os procedimentos propostos podem ser aplicados noutros contextos e também podem ser estendidos para detetar mudanças na estrutura e dinâmica dos processos. Nestes casos, será necessário calibrar os percentis das distribuições empíricas usadas para detetar o tempo de ocorrência de *outliers*, seja na abordagem do limiar ou na abordagem da reamostragem paramétrica. É ainda possível que diferentes aplicações precisem de diferentes níveis de decomposição na TWD.

Agradecimentos Este trabalho foi parcialmente financiado pela FCT- Fundação para a Ciência e a Tecnologia, através do Centro de Investigação e Desenvolvimento em Matemática e Aplicações (CIDMA), projeto UID/MAT/04106/2019.

Referências

[1] Al-Osh, M. A. e Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Ser. Anal.* **8**, 261–275.

[2] Barczy, M., Ispány, M., Pap, G., Scotto, M. e Silva, M. E. (2010). Innovational Outliers in INAR(1) Models. *Commun. Stat. - Theor. M.* **39**, 3343–3362.

[3] Barczy, M., Ispány, M., Pap, G., Scotto, M. e Silva, M. E. (2011). Additive outliers in INAR(1) models. *Stat. Pap.* **53**, 935–949.

[4] Bilen, C. e Huzurbazar, S. (2002). Wavelet-Based Detection of Outliers in Time Series. *J. Comp. Graph. Stat.* **11**, 311–327.

[5] Bourguignon, M. e Vasconcellos, K. L. P. (2018). The effects of additive outliers in INAR(1) process and robust estimation, *Stat. Theory Relat. Fields* **2**, 206–214.

[6] Chang, I., Tiao, G. C. e Chen, C. (1988). Estimation of time series parameters in the presence of outliers. *Technometrics* **30**, 193–204.

[7] Chen, C. e Liu, L. M. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Am. Stat. Assoc.* **88**, 284–297.

[8] Fox, A. J. (1972). Outliers in Time Series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **34**, 350–363.

[9] Grané, A. e Veiga, H. (2010). Wavelet-based detection of outliers in financial time series. *Comput. Stat. Data An.* **54**, 2580–2593.

[10] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**, 674–693.

[11] McKenzie, E. (1985). Some simple models for discrete variate time series. *Water Resources Bull.* **21**, 645–650.

[12] Percival, D. e Walden, A. (2006). *Wavelet methods for time series analysis*, Cambridge Series in Statistical and Probabilistic Mathematics, New York, Cambridge University Press.

[13] Scotto, M. G., Weiß, C. H. e Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: a review. *Stat. Modelling* **15**, 590–618.

[14] Silva, I. e Silva, M. E. (2018). Wavelet-based detection of outliers in Poisson INAR(1) time series, In Oliveira, T. A., Kitsos, C. P., Oliveira, A., Grilo, L. (Eds), *Contributions to Statistics - Recent Studies on Risk Analysis and Statistical Modeling*, Springer, 183–195.

[15] Silva, M. E. e Oliveira, V. L. (2004). Difference equations for the higher-order moments and cumulants of the INAR(1) model. *J. Time Ser. Anal.* **25**, 317–333.

[16] Silva, M. E. e Pereira, I. (2015). Detection of additive outliers in Poisson INAR(1) time series. In Bourguignon, J. P. *et al.* (Eds.) *CIM Series in Mathematical Sciences - Mathematics of Energy and Climate Change*, Springer, 377–388.

[17] Steutel, F. W. e Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Ann. Probab.* **7**, 893–899.

[18] Tsay, R. S. (1986). Time series model specification in the presence of outliers. *J. Am. Stat. Assoc.* **81**, 132–141.

[19] Tsay, R. S. (1992). Model checking via parametric bootstraps in time series analysis. *J. R. Stat. Soc. Ser. C Appl. Stat.* **41**, 1–15.

[20] Weiß, C. H. (2007). Controlling correlated processes of Poisson counts. *Qual. Reliab. Eng. Int.* **23**, 741–754.

Modelos de contagem com estrutura periódica

Isabel Pereira, *isabel.pereira@ua.pt*
CIDMA e Departamento de Matemática,
Universidade de Aveiro

Magda Monteiro, *msvm@ua.pt*
CIDMA e ESTGA,
Universidade de Aveiro

Cláudia Santos, *csps@ua.pt*
CIDMA-Univ. Aveiro e ESAC-Instituto Politécnico de Coimbra

1 Introdução

Os modelos correlacionados periodicamente (ou cicloestacionários) introduzidos por Bennett (1958) e Gladyshev (1961, 1963) têm recebido muita atenção e desenvolvimento na literatura, parcialmente motivado pela sua enorme aplicabilidade a diversas áreas, das quais se destacam a hidrologia, economia, meteorologia e processamento de sinal. Grande parte da literatura neste tópico contempla os denominados modelos autorregressivos de médias móveis periódicos (PARMA), sendo extensões dos modelos ARMA por forma a incluírem parâmetros variando periodicamente no tempo. Podem ser destacados os trabalhos de Lund *et al.* (2006), Shao (2006) e Basawa e Lund (2001), que estudam as respetivas propriedades probabilísticas do modelo assim como técnicas de inferência e previsão. Em contrapartida, a análise de séries temporais de valores inteiros não se encontra ainda tão desenvolvida, apesar de haver uma grande aplicabilidade para modelar, entre outros, fenómenos associados à procura turística, à ocorrência de incêndios, ou ainda em ciências sociais. Com este objetivo, Monteiro *et al.* (2010) introduziram a classe dos modelos univariados INAR baseados em parâmetros de filtragem variando periodicamente no tempo. Posteriormente Monteiro *et al.* (2015) generalizaram esta classe por forma a contemplar o caso bivariado. Nesse trabalho, foram considerados os seguintes processos de inovações: com distribuição bivariada de Poisson - usualmente considerado devido à sua maior facilidade de tratamento e boas propriedades e com distribuição bivariada binomial negativa - mais flexível e adequada para modelar dados reais, pois estes habitualmente apresentam sobredispersão (variância exceder o valor médio do processo). Mais recentemente, Santos (2017) estendeu os resultados anteriores ao caso multivariado, mantendo como orientação a proposta de utilização de uma matriz diagonal feita por Pedeli e Karlis (2011), matriz esta necessária para se fazer a generalização do processo para uma dimensão superior à unidade.

Este artigo pretende proporcionar uma visão global sobre estes processos de contagem com estrutura periódica e mostrar a sua grande aplicabilidade, tendo como objetivo principal apresentar a informação necessária para a sua formulação no caso univariado e a sua extensão para uma dimensão superior a um. O estudo das suas propriedades básicas probabilísticas e estatísticas e a análise do comportamento de estimadores obtidos por diferentes abordagens, podem ser encontrados na literatura indicada a esse propósito.
O resto do artigo está organizado da seguinte forma: na secção 2 é introduzido o modelo periódico

INAR(1) e são referidas algumas propriedades básicas probabilísticas; na secção 3 é feita a extensão para o caso multivariado e apresenta-se uma aplicação a um conjunto de dados reais; finaliza-se tecendo algumas observações.

2 Modelo periódico INAR(1)- PINAR(1)

2.1 Motivação

Para motivar e ilustrar este tipo de modelos periódicos de estrutura autorregressiva, apresenta-se na Figura 1 o número mensal de desempregados de curta duração no concelho de Penamacor, no período de janeiro de 1997 a dezembro de 2007 e a correspondente função de autocorrelação amostral, onde está bem patente a existência de periodicidade de período 12.

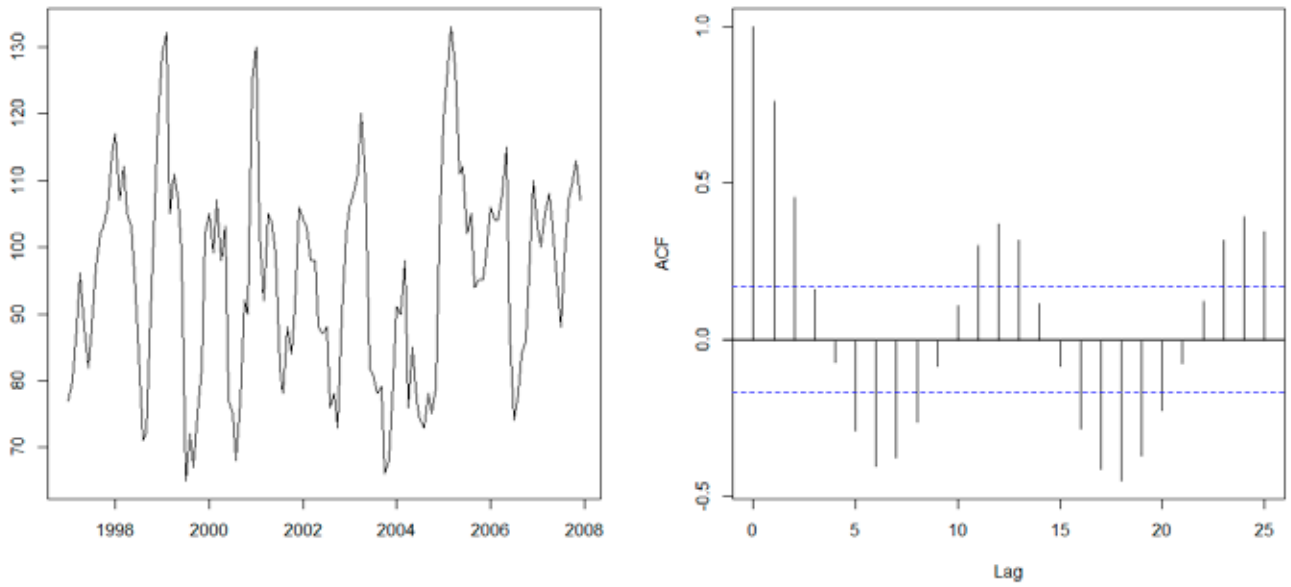


Figura 1: Número mensal de desempregados de curta duração no concelho de Penamacor, de janeiro de 1997 a dezembro de 2007 e função de autocorrelação amostral.

2.2 Definição e propriedades

O modelo de valores inteiros autorregressivo com estrutura periódica de ordem 1 e período T (designado por PINAR$(1)_T$) é definido pela equação recursiva:

$$X_t = \phi_t \circ X_{t-1} + Z_t, t \in \mathbb{N} \tag{1}$$

onde $\phi_t = \alpha_j \in (0,1)$ para $t = j + kT, (j = 1, \ldots, T; k \in \mathbb{N}_0)$ e o operador *thinning* $\circ$ é definido por

$$\phi_t \circ X_{t-1} \stackrel{d}{=} \sum_{i=1}^{X_{t-1}} U_{i,t}(\phi_t),$$

sendo $(U_{i,t}(\phi_t))$ uma sucessão periódica de v.a.'s independentes com probabilidade de sucesso $P(U_{i,t}(\phi_t) = 1) = \phi_t$, para $i = 1, 2, \ldots$.

Adicionalmente assume-se que (Z_t) é uma sucessão periódica de v.a.'s de distribuição de Poisson com valor médio υ_t, i.e, $Z_t \sim P(\upsilon_t))$ com $\upsilon_t = \lambda_j$ para $t = j+kT, (j = 1,\ldots,T; k \in \mathbb{N}_0)$, que se assumem serem independentes de X_{t-1} e $\phi_t \circ X_{t-1}$. Para evitar qualquer ambiguidade, considera-se T como sendo o menor valor inteiro positivo que satisfaz (1).

Convém reforçar que a natureza discreta do processo (X_t) é assegurada pela introdução do operador $\circ$, assumindo o papel análogo ao da multiplicação nos processos contínuos ARMA.

- **Distribuição cicloestacionária e momentos**

  No resultado que se segue introduz-se a distribuição ciclo-estacionária de X_t para cada um dos T períodos. Por uma questão de simplicidade define-se a sucessão periódica $\beta_{t,i}$ de período T

$$\beta_{t,i} = \begin{cases} \prod_{j=0}^{i-1} \phi_{t-j} & i > 0 \\ 1 & i = 0 \end{cases},$$

  que pode ser reescrita como

$$\beta_{t,i} = \begin{cases} \beta_{t,j}\beta_{T,T}^{k}, & i = j+kT; j = 1,2,\ldots,T \\ 1, & i = 0 \end{cases}.$$

  Prova-se a seguinte proposição:

  Proposição: O processo (X_t) com $t = j+kT$ (para um valor fixo de $j = 1,\ldots,T$ e $k \in \mathbb{N}$) é uma cadeia de Markov irredutível, aperiódica e recorrente positiva; considerando o ciclo fixo j, a distribuição estacionária de (X_t) é a mesma de

$$V_j = \sum_{m=1}^{+\infty} \sum_{a=0}^{T-1} \left(\beta_{j,j}\beta_{T,a}\beta_{T,T}^{m-1}\right) \circ Z_{T(m+1)-a} + \sum_{m=0}^{j-1} \beta_{j,m} \circ Z_{j-m},$$

  com a série a convergir quase certamente e em média quadrática.

  A partir do resultado anterior obtêm-se a média periódica e a função de autocovariância de (X_t).

  Lema: Para um valor fixo $j = 1,\ldots,T$, com $T \in \mathbb{N}$, $t = j+kT$ e $k \in \mathbb{N}_0$

$$\mu_j = \mu_t = E(X_t) = V(X_t) = \frac{\sum_{k=0}^{j-1} \beta_{j,k}\lambda_{j-k} + \beta_{j,j} \sum_{i=0}^{T-j-1} \beta_{T,i}\lambda_{T-i}}{1-\beta_{T,T}} \tag{2}$$

  sob a convenção $\sum_{i=0}^{-1} = 0$. Adicionalmente, para $j = 1,\ldots,T$ e $h \geq 0$, $\gamma_{j+kT}(h) = \gamma_j(h) = \beta_{j+h,h}\mu_j$ e $\gamma_{j+kT}(-h) = \gamma_j(-h) = \beta_{j+kT,h}\mu_{j+kT-h}$.

  É de referir que o valor médio μ_j pode ser calculado de forma recursiva através da expressão

$$\mu_j = \beta_{j,j}(\mu_T + \frac{1}{\beta_{j,j}} \sum_{k=0}^{j-1} \beta_{j,k}\lambda_{j-k}),\ j = 1,\ldots,T.$$

  É de realçar que a função de autocovariância $\gamma_j(\cdot)$ já não é simétrica em h, no entanto $\gamma_t(-h) = \gamma_{t-h}(h)$ e $\gamma_t(h) = \gamma_{t+h}(-h)$. Por outro lado, e uma vez que h pode ser reescrita na forma $h = i+mT$, para algum $i \in \{1,\ldots,T\}$ e $m \in \mathbb{N}$, a função de autocovariância toma a forma $\gamma_j(h) = \beta_{T,T}^{m}\beta_{j+i,i}\mu_j$ e $\gamma_j(-h) = \beta_{T,T}^{m}\beta_{j+T,i}\mu_{j+T-i}$.

- **Distribuição marginal de** (X_t)

  Prova-se que a distribuição marginal de (X_t), com $t = j+kT$ e considerando um valor fixo de $j = 1,\ldots,T$, com $T \in \mathbb{N}$ e $k \in \mathbb{N}_0$, é Poisson com valor médio μ_j sse o processo periódico das inovações (Z_t) formar uma sucessão de variáveis aleatórias independentes também com distribuição de Poisson com valor médio λ_j.

2.3 Estimação dos parâmetros

Para se estimar o vetor parâmetro $\boldsymbol{\theta} = (\alpha_1, \lambda_1, \ldots, \alpha_T, \lambda_T)$, considera-se o conjunto de observações $(X_1, \ldots, X_{NT})$ satisfazendo o modelo (1), assumindo-se que N é o número completo de ciclos. Em Monteiro *et al.* (2010) mostra-se que os estimadores de Yule-Walker, de mínimos quadrados condicionais e de máxima verosimilhança condicionais são centrados e consistentes, sendo estes últimos também eficientes.

3 Generalização para modelos periódicos multivariados INAR(1)- PMINAR(1)

Monteiro *et al.* (2015) generalizaram a classe dos modelos univariados INAR(1), definidos na secção anterior, ao caso bivariado e Santos (2017) estendeu-o ao caso multivariado mais geral. Por uma questão de simplicidade, neste artigo apresenta-se a metodologia usada para a inclusão do caso bivariado.

3.1 Definição e probabilidades de transição do modelo PBINAR(1)

Generalizando o modelo proposto por Pedeli e Karlis (2011) para o caso periódico e considerando uma sucessão bivariada periódica de inovações obtém-se o modelo periódico bivariado definido por

$$\boldsymbol{X}_t = \boldsymbol{A}_t \circ \boldsymbol{X}_{t-1} + \boldsymbol{Z}_t = \begin{bmatrix} \phi_{1,t} & 0 \\ 0 & \phi_{2,t} \end{bmatrix} \circ \begin{bmatrix} X_{1,t-1} \\ X_{2,t-1} \end{bmatrix} + \begin{bmatrix} Z_{1,t} \\ Z_{2,t-1} \end{bmatrix},$$

onde $\phi_{j,t} = \alpha_{j,i}$, for $t = i+kT$ $(i = 1,\ldots,T,\, k \in \mathbb{N}_0)$. Neste contexto o operador *thinning* $\circ$ é definido por

$$\phi_{j,t} \circ X_{t-1} \stackrel{d}{=} \sum_{m=1}^{X_{t-1}} U_{m,t}(\phi_{j,t}),$$

sendo $(U_{m,t}(\phi_{j,t}))$, $m = 1,2,\ldots$, uma sucessão periódica de variáveis aleatórias de Bernoulli com probabilidade de sucesso $P(U_{m,t}(\phi_{j,t}) = 1) = \phi_{j,t}$. No caso bivariado a operação binomial *thinning* é uma operação matricial que atua como a multiplicação usual de matrizes, mantendo as propriedades da operação binomial *thinning*. Assim, pela definição da operação *thinning* matricial tem-se

$$X_{j,t} = \phi_{j,t} \circ X_{j,t-1} + Z_{j,t},\, j = 1,2.$$

Assume-se que $(Z_{1,t}, Z_{2,t})_{t\in\mathbb{N}}$ é uma sucessão periódica de vetores aleatórios independentes com a mesma distribuição de média $\boldsymbol{\delta}_t = \begin{bmatrix} \delta_{1,t} \\ \delta_{2,t} \end{bmatrix}$, com $\delta_{j,t} = \lambda_{j,i}$ e variância $\boldsymbol{\theta_t} = \begin{bmatrix} \theta_{1,t} \\ \theta_{2,t} \end{bmatrix}$, $\theta_{j,t} = \upsilon_{j,i}\lambda_{j,i}$ para $t = i+kT (i = 1,\ldots,T; k \in \mathbb{N}_0)$, onde para cada t, $Z_{j,t}$ é independente de $X_{j,t-1}$ e de $\phi_{j,t} \circ X_{j,t-1}$. Tal como anteriormente T é o menor inteiro positivo que satisfaz a equação inicial.

Inovações de Poisson:
Considerando que também $\mathbf{Z}_{i+kT}(i=1,\ldots,T)$ segue a distribuição bivariada de Poisson (Johnson *et al.* 1997, p.125) dada por:

$$\begin{aligned} &P(Z_{1,i+kT}=a, Z_{2,i+kT}=b) = \\ &= e^{-(\lambda_{1,i}+\lambda_{2,i}-\varphi_i)} \sum_{m=0}^{\min(a,b)} \frac{(\lambda_{1,i}-\varphi_i)^{a-m}}{(a-m)!} \frac{(\lambda_{2,i}-\varphi_i)^{b-m}}{(b-m)!} \frac{\varphi_i^m}{m!}, \end{aligned}$$

onde $\lambda_{1,i}, \lambda_{2,i} > 0$ e $\varphi_i \in [0, \min(\lambda_{1,i}, \lambda_{2,i})]$, prova-se que a distribuição marginal de (X_{i+kT}), para $i = 1,\ldots,T$ e $k \in \mathbb{N}_0$ segue também distribuição de Poisson. Neste caso as probabilidades de transição são dadas por

$$\begin{aligned} p_i(\boldsymbol{y}|\boldsymbol{x}) = P(\boldsymbol{X}_{i+kT} = \boldsymbol{y} | \boldsymbol{X}_{i-1+kT} = \boldsymbol{x}) = \\ &= \sum_{m_1=0}^{M_1} \sum_{m_2=0}^{M_2} P(\alpha_{1,i} \circ X_{1,i-1+kT} = m_1, \alpha_{2,i} \circ X_{2,i-1+kT} = m_2 | \boldsymbol{X}_{i-1+kT} = \boldsymbol{x}) \times \\ &\quad \times P(Z_{1,i+kT} = y_1 - m_1, Z_{2,i+kT} = y_2 - m_2) \\ &= \sum_{m_1=0}^{M_1} \sum_{m_2=0}^{M_2} \prod_{j=1}^{2} C_{m_j}^{x_j} \alpha_{j,i}^{m_j} (1-\alpha_{j,i})^{x_j - m_j} \times e^{-(\lambda_{1,i}+\lambda_{2,i}-\varphi_i)} \times \\ &\quad \times \sum_{l=0}^{L} \frac{(\lambda_{1,i}-\varphi_i)^{y_1-m_1-l}}{(y_1-m_1-l)!} \frac{(\lambda_{2,i}-\varphi_i)^{y_2-m_2-l}}{(y_2-m_2-l)!} \frac{\varphi_i^l}{l!} \\ &= e^{-(\lambda_{1,i}+\lambda_{2,i}-\varphi_i)} \sum_{m_1=0}^{M_1} \sum_{m_2=0}^{M_2} \sum_{l=0}^{L} \frac{\varphi_i^l}{l!} \prod_{j=1}^{2} C_{m_j}^{x_j} \alpha_{j,i}^{m_j} (1-\alpha_{j,i})^{x_j - m_j} \frac{(\lambda_{j,i}-\varphi_i)^{y_j-m_j-l}}{(y_j-m_j-l)!} \end{aligned}$$

com $M_1 = \min(x_1, y_1)$, $M_2 = \min(x_2, y_2)$ e $L = \min(y_1 - m_1, y_2 - m_2)$.

O vetor parâmetro a estimar de dimensão $5T$ é

$$\boldsymbol{\theta} = (\alpha_{1,1},\ldots,\alpha_{1,T},\alpha_{2,1},\ldots,\alpha_{2,T},\lambda_{1,1},\ldots,\lambda_{1,T},\lambda_{2,1}\ldots,\lambda_{2,T},\varphi_1,\ldots,\varphi_T).$$

Inovações com distribuição binomial negativa:
O processo PBINAR(1) de inovações binomiais negativas torna-se mais flexível que o modelo com processo de inovações de Poisson. Neste caso as probabilidades de transição, associadas ao i-ésimo período, são dadas por:

$$p_i(\boldsymbol{y}|\boldsymbol{x}) = \left(\frac{\beta_i^{-1}}{\eta_i}\right)^{\beta_i^{-1}} \sum_{m_1=0}^{M_1} \sum_{m_2=0}^{M_2} \prod_{j=1}^{2} \left[\frac{\Gamma(\beta_i^{-1} + x_1 - m_1 + x_2 - m_2)}{\Gamma(\beta_i^{-1})\Gamma(x_j - m_j + 1)} C_{m_j}^{x_j} \alpha_{i,j}^{m_j} (1-\alpha_{i,j})^{x_j - m_j} \left(\frac{\lambda_{i,j}}{\eta_i}\right)^{y_j - m_j} \right],$$

com $M_1 = \min(x_1, y_1)$, $M_2 = \min(x_2, y_2)$ e $L = \min(y_1 - m_1, y_2 - m_2)$, $\lambda_{i,j}, \beta_i > 0$, $j = 1,2$ e $\eta_i = \lambda_{i,1} + \lambda_{i,2} + \beta_i^{-1}$, $i = 1,\ldots,T$.

Por conseguinte os parâmetros associados à distribuição binomial negativa bivariada são: $\lambda_{1,i}, \lambda_{2,i}, \beta_i > 0$, para todo $i \in \{1,2,\ldots,T\}$. Além disso, $\lambda_{1,i}$ e $\lambda_{2,i}$ representam as médias de cada componente na estação

i e β_i é o parâmetro associado à sobredispersão para cada estação i. De facto, a variância $\sigma^2_{j,i}$ é igual a $\lambda_{j,i}(1+\beta_i\lambda_{j,i})$. A covariância entre duas componentes, em cada estação é dada por $\varphi_i = \lambda_{1,i}\lambda_{2,i}\beta_i$ com $i = 1,\ldots,T$, permitindo apenas correlação positiva.

Considerando o processo mais geral PMINAR(1), prova-se a existência de distribuição única e estacionária, assim como as propriedades assintóticas de não enviesamento, consistência e eficiência de estimadores baseados na verosimilhança. No caso de dimensão superior a dois, e com vista a reduzir o tempo envolvido com o cálculo computacional, sugere-se o uso de estimadores que maximizam a verosimilhança composta (Pedeli e Karlis 2013, Santos 2017).

3.2 Aplicação a séries de contagens de incêndios

Apresenta-se um exemplo de aplicação do modelo PMINAR(1) a três conjuntos de dados de contagens de incêndios. Os dados representam o número de incêndios mensais em três concelhos do distrito de Aveiro, nomeadamente de Anadia, Oliveira do Bairro e Vagos, durante 32 anos consecutivos de 1986 a 2017. As contagens mensais de fogos diários nesses concelhos estão representadas na Figura 2. Os fo-

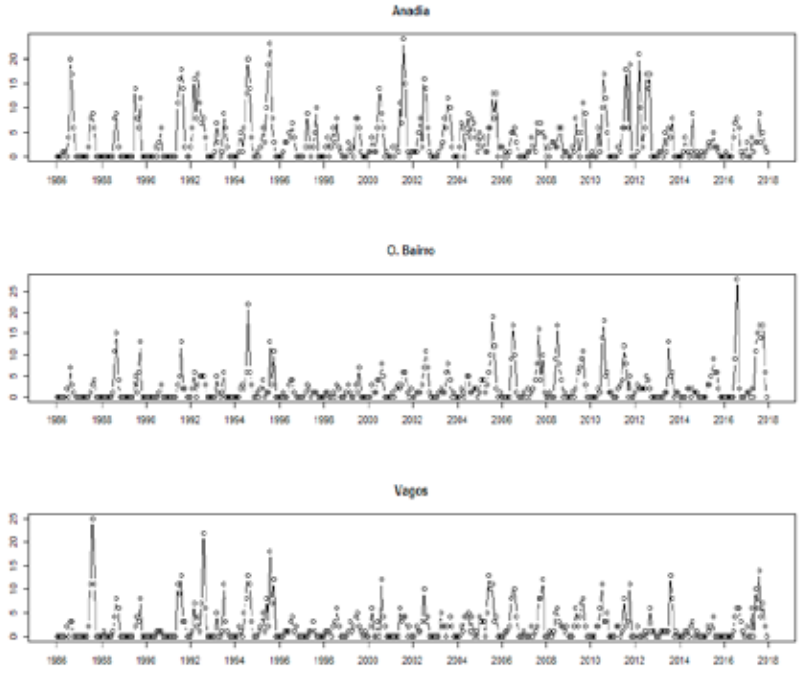


Figura 2: Número de incêndios mensais nos concelhos de Anadia, Oliv. Bairro e Vagos.

gos florestais são um dos grandes problemas da atualidade sobretudo em muitos países europeus a norte do Mar Mediterrâneo, nomeadamente Portugal, Espanha, Itália e Grécia, representando uma ameaça não apenas a nível ambiental mas também para pessoas e bens. Na Europa, Portugal é o país com um maior número de fogos por unidade de área ardida e por habitante (San Miguel-Ayanz e Camia 2009). A frequência dos incêndios é nitidamente diferente de norte a sul e de este a oeste (Nunes *et al.* 2016; Nunes 2012). A distribuição dos incêndios durante o ano segue um padrão regular, fortemente influenciado por variações sazonais de temperatura e ocorrência de chuva. Pelo que é expectável haver um maior número de ocorrências de fogos no verão, com um pico em julho/agosto e um menor número de ocorrências na estação chuvosa. Pode-se encontrar mais informação em Tonini *et al.* (2017) e Scotto *et al.* (2014).

A função de autocorrelação amostral ilustrada na Figura 3 mostra um padrão periódico de 12 meses. A

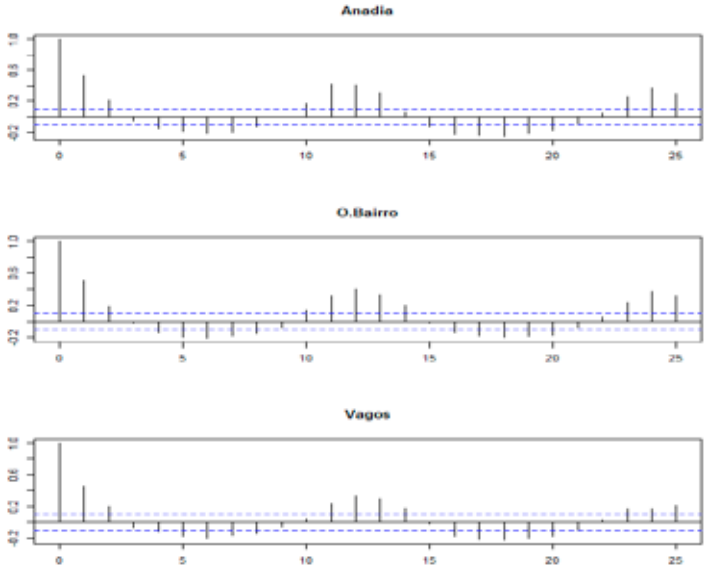


Figura 3: Número de incêndios mensais nos concelhos de Anadia, Oliv. Bairro e Vagos: função de autocorrelação amostral.

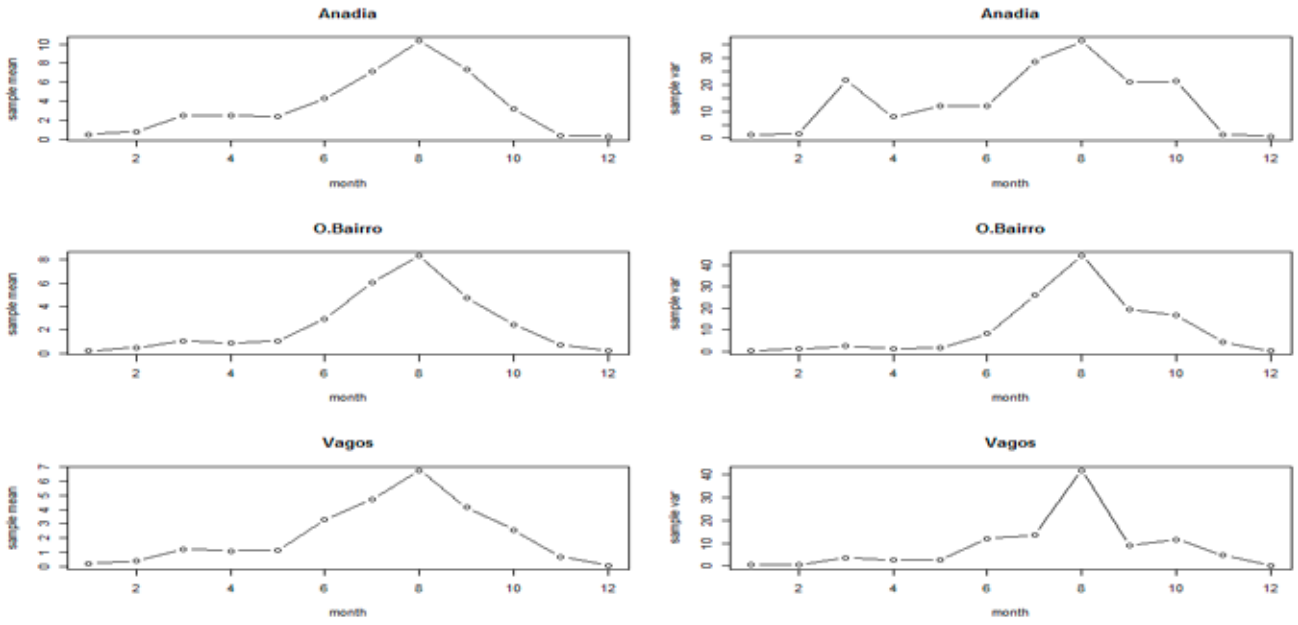


Figura 4: Número de incêndios mensais nos concelhos de Anadia, Oliv. Bairro e Vagos: média e variância amostrais.

Figura 4 apresenta as médias e variâncias amostrais do número de fogos por mês. Também se pode ver a representação das correspondentes correlações cruzadas na Figura 5.

Nestes três concelhos, constata-se que em muitos meses a variância é superior à média, indicando sobredispersão dos dados; por esse motivo assumiu-se um processo de inovações trivariado periódico de distribuição binomial negativa. A Tabela 1 apresenta as estimativas de máxima verosimilhança condicional (CML) e usando a verosimilhança composta (CL) obtidas pelo ajustamento do modelo trivariado periódico INAR(1) de período 12. Os erros padrão (SE) foram calculados numericamente a partir da matriz Hessiana durante o processo de otimização incluído no R.

Da análise dos valores registados na tabela verifica-se que em muitos casos as estimativas (CL ou CML) são muito próximas. Apesar de haver alguma perda de eficiência quando se usa o método baseado na verosimilhança composta (CL), pode ser uma alternativa razoável para a estimação de parâmetros deste

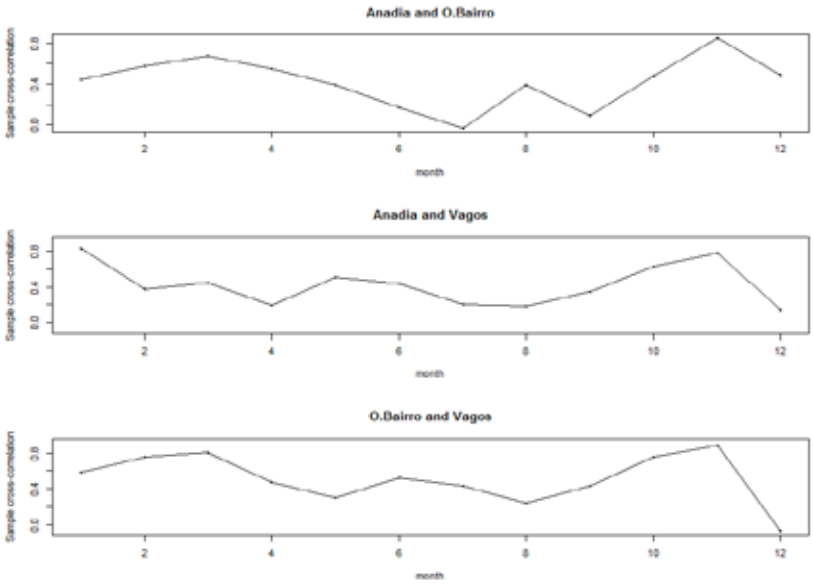


Figura 5: Número de incêndios mensais nos concelhos de Anadia, Oliv. Bairro e Vagos: correlações cruzadas.

complexo processo perante a falha de outros métodos. Acresce ainda referir que as estimativas CL foram usadas para inicializar o processo de otimização conducente a estimativas de máxima verosimilhança (condicionais). Relativamente às estimativas dos parâmetros de autocorrelação, algumas delas não são estatisticamente significativas, sugerindo que nesses meses o número de fogos é sobretudo modelado pelo processo de inovações.

4 Observações finais

Apesar da escolha de uma matriz diagonal poder restringir a aplicabilidade deste modelo em situações reais, no exemplo de aplicação tratado o pressuposto de independência entre as contagens dos fogos nos três concelhos parece ser defensável. No entanto, uma das limitações deste tipo de modelos, construidos com base nos modelos de Pedeli e Karlis (2011) é o facto de a matriz de autocorrelação ser diagonal, não permitindo a existência de correlação cruzada entre as contagens das séries e implicando que as marginais se comportem como o modelo univariado (1). Pelo que fará todo o sentido desenvolver outro tipo de extensões por forma a tornar estes modelos mais flexíveis, nomeadamente para permitir correlações negativas, correlações cruzadas entre as séries e eventualmente a introdução de covariáveis para acomodar alguma dependência dos parâmetros *thinning* em relação a certos fatores considerados relevantes no contexto do problema. No entanto, um dos grandes condicionalismos deste tipo de modelos, ou outros adequados para fenómenos com periodicidade, será o elevado número de parâmetros a estimar, implicando que se tenha de ter um especial cuidado com a dimensão da amostra dos dados a analisar.

Tabela 1: Estimativas e erros padrão usando os métodos CL e CML, obtidas pelo ajustamento do modelo trivariado periódico INAR(1) de período 12.

	Composite Likelihood (CL)							Conditonal Maximum Likelihood (CML)						
	Anadia		O.Bairro		Vagos			Anadia		O.Bairro		Vagos		
	α_1	λ_1	α_2	λ_2	α_3	λ_3	β	α_1	λ_1	α_2	λ_2	α_3	λ_3	β
January	0.3138	0.4284	0.0004	0.2003	0.0018	0.2331	4.0031	0.3296	0.4098	0.0003	0.1936	0.0002	0.2258	4.1623
	(0.1533)	(0.1360)	(0.0079)	(0.0738)	(0.0079)	(0.0820)	(1.4481)	(0.1321)	(0.0815)	(0.0063)	(0.0658)	(0.0006)	(0.1034)	(0.1751
February	2.4×10^{-04}	0.8423	0.5238	0.3813	0.1598	0.3504	1.9601	0.0008	0.8125	0.5360	0.3683	0.1041	0.3522	2.0677
	(1.3×10^{-05})	(0.1799)	(0.1954)	(0.1021)	(0.1942)	(0.1007)	(0.5981)	(0.0028)	(0.2727)	(0.2689)	(0.1489)	(0.2816)	(0.1462)	(0.8986)
March	0.0137	2.5256	0.3015	0.9517	0.3617	1.0902	1.8017	0.0185	2.4692	0.0544	1.0368	0.2816	1.0818	1.9611
	(0.1358)	(0.4403)	(0.1547)	(0.1958)	(0.2165)	(0.2200)	(0.3567)	(0.3261)	(0.7041)	(0.2527)	(0.3328)	(0.3218)	(0.3427)	(0.6019)
April	0.3775	1.6122	0.1198	0.7392	0.0918	1.0211	1.4199	0.3826	1.5867	0.1242	0.7118	0.1031	0.9713	1.4330
	(0.0552)	(0.2875)	(0.0638)	(0.1563)	(0.0735)	(0.2034)	(0.4152)	(0.0767)	(0.0854)	(0.0981)	(0.4241)	(0.2226)	(0.2877)	(0.6154)
May	0.1608	2.0588	0.4898	0.6386	0.1343	1.0460	1.5132	0.1638	2.0228	0.5005	0.6090	0.0830	1.0654	1.4160
	(0.0602)	(0.3645)	(0.0930)	(0.1442)	(0.0881)	(0.2121)	(0.3824)	(0.0872)	(0.1262)	(0.1201)	(0.5312)	(0.2004)	(0.3133)	(0.5206)
June	0.1651	4.0059	0.4291	2.5060	0.0546	3.2998	0.6851	0.1565	3.8998	0.3723	2.5536	0.0164	3.2935	0.6456
	(0.0877)	(0.4894)	(0.1541)	(0.3380)	(0.1673)	(0.4270)	(0.1436)	(0.1164)	(0.2238)	(0.2262)	(0.7065)	(0.5070)	(0.6242)	(0.2072)
July	0.3829	5.5467	0.7075	4.1229	0.3899	3.5249	0.4738	0.3702	5.5399	0.6336	4.1700	0.3204	3.6262	0.3761
	(0.0844)	(0.6256)	(0.0792)	(0.4577)	(0.0788)	(0.4292)	(0.0972)	(0.1097)	(0.1127)	(0.1169)	(0.8504)	(0.6504)	(0.6287)	(0.1186)
August	0.4761	6.5477	0.3116	6.4485	0.0539	6.6102	0.5053	0.3966	7.4866	0.2949	6.5650	0.1043	6.2613	0.4062
	(0.0614)	(0.7256)	(0.0648)	(0.7007)	(0.0783)	(0.7115)	(0.0992)	(0.0911)	(0.0881)	(0.0954)	(1.1461)	(0.9976)	(0.9379)	(0.1270)
September	0.3359	3.6745	0.0725	4.2058	0.2783	2.2748	0.5909	0.3592	3.6391	0.0955	3.9529	0.2757	2.2953	0.5274
	(0.0491)	(0.6011)	(0.0496)	(0.6019)	(0.0389)	(0.3500)	(0.1443)	(0.0647)	(0.0720)	(0.0546)	(0.8382)	(0.8483)	(0.5024)	(0.1976)
October	0.0242	2.9897	0.0408	2.2900	0.0620	2.3881	2.5848	0.0287	2.9770	0.0330	2.2808	0.0851	2.2089	2.6112
	(0.0194)	(0.5716)	(0.0351)	(0.4632)	(0.0510)	(0.4961)	(0.5708)	(0.0385)	(0.9424)	(0.0448)	(0.7316)	(0.0623)	(0.7235)	(0.9358)
November	5.1×10^{-04}	0.4189	0.0116	0.6807	0.0034	0.7104	6.0760	3.6×10^{-04}	0.4063	0.0152	0.6506	0.0004	0.6875	6.1412
	(0.0003)	(0.1378)	(0.0178)	(0.2127)	(0.0025)	(0.2174)	(1.6077)	(1.4×10^{-04})	(0.2195)	(0.0247)	(0.3224)	(0.0295)	(0.3304)	(2.7802)
December	0.5025	0.1442	0.1072	0.1185	0.0187	0.0638	6.3424	0.4937	0.1434	0.1175	0.1067	0.0293	0.0625	6.9802
	(0.1053)	(0.0663)	(0.0545)	(0.0595)	(0.0388)	(0.0369)	(2.8320)	(0.1506)	(0.1002)	(0.0761)	(0.0817)	(0.1322)	(0.0651)	(2.6052)

Referências

[1] Bennett, W.R. (1958). Statistics of regenerative digital transmission. *Bell System Technol. J.* **37**, 1501–1542.

[2] Basawa, I.V., Lund, R.B. (2001). Large sample properties of parameter estimates for periodic ARMA models. *J. Time Ser. Anal.* **22**, 651–666.

[3] Gladyshev, E.G. (1961). Periodically correlated random sequences, *Soviet Math.* **2**, 385–388.

[4] Gladyshev, E.G. (1963). Periodically and almost-periodically correlated random processes with a continuous time parameter. *Theory Prob. Appl.* **8**, 173–177.

[5] Johnson, N., Kotz, S., Balakrishnan, N. (1997). *Discrete multivariate distributions*. New York: Wiley.

[6] Lund, R.B., Shao, Q., Basawa, I.V. (2006). Parsimonious periodic time series modeling. *Aust. N. Z. J. Statist.* **48**, 33–47.

[7] Monteiro, M., Scotto, M.G., Pereira, I. (2010). Integer-valued autoregressive processes with periodic structure. *J. Stat. Plann. Inference* **140**, 1529–1541.

[8] Monteiro, M., Scotto, M.G., Pereira, I. (2015). A periodic bivariate integer-valued autoregressive model. In: Bourguignon JP, Jelstch R, Pinto A, Viana M (eds) *Dynamics, Games and Science - International Conference. Advanced School Planet Earth DGS II.* Springer, Switzerland, pp 455-477.

[9] Nunes, A.N. (2012). Regional variability and driving forces behind forest fires in Portugal an overview of the last three decades (1980–2009). *Applied Geog.* **34**, 576–586.

[10] Nunes, A.N., Lourenço, L., Castro Meira A.C. (2016). Exploring spatial patterns and drivers of forest fires in Portugal (1980–2014). *Sci. Total Environ.* **573**, 1190–1202.

[11] Pedeli, X., Karlis, D. (2011). A bivariate INAR(1) process with application. *Stat. Modelling* **11**, 325–349.

[12] Pedeli, X., Karlis, D. (2013). On composite likelihood estimation of a multivariate INAR(1) model. *J. Time Ser. Anal.* **34**, 206–220.

[13] San Miguel-Ayanz J, Camia A (2009). Forest fires at a glance: facts, figures and trends in the EU. Living with wildfires: what science can tell us. A Contribution to the Science-Policy Dialogue, Joensuu: European Forest Institute.

[14] Santos, C. (2017). Statistical Analysis of Count Time Series with Periodic Structure, PhD thesis, Universidade de Aveiro.

[15] Scotto, M.G., Gouveia, S., Carvalho, A., Monteiro, A., Martins. V., Flannigan, M., San Miguel-Ayanz, J., Miranda, A.I., Borrego, C. (2014). Area burned in Portugal over recent decades: an extreme value analysis. *Int. J. Wildland Fire* **23**, 812–824.

[16] Shao, Q. (2006). Mixture periodic autoregressive time series models. *Statist. Probab. Lett.* **76**, 609–618.

[17] Tonini, M., Pereira, M.G., Parente, J., Orozco, C.V. (2017). Evolution of forest fires in Portugal: from spatio-temporal point events to smoothed density maps. *Nat. Hazards* **85**, 1489–1510.

Uso de distribuições geométricas autorregressivas na análise de sequências de ADN

Sónia Gouveia, *sonia.gouveia@ua.pt*

Instituto de Engenharia Electrónica e Informática de Aveiro (IEETA),
Centro de I&D em Matemática e Aplicações (CIDMA), Universidade de Aveiro

1 Introdução

Muitos dos desenvolvimentos em modelação estatística são inspirados pela necessidade de analisar e interpretar dados reais. Este trabalho apresenta um desses casos, onde o desenvolvimento da teoria é motivado e contextualizado na análise de dados genéticos, mais concretamente na análise de sequências de ADN (o acrónimo para ácido desoxirribonucleico). Aqui desenvolve-se a forma teórica da distribuição de probabilidade geométrica assumindo que a variável binária que a gera exibe uma estrutura de autocorrelação do tipo autorregressiva. Ilustra-se também como é que esta distribuição poderá ser utilizada no contexto da análise de sequências de ADN mitocondrial, não sendo, seguramente, a única aplicação possível para este tipo de distribuição.

Em genética, uma estratégia habitual na análise de ADN consiste em comparar a informação extraída da sequência com um padrão de referência, para posteriormente interpretar as diferenças obtidas. Estas diferenças poderão realçar por exemplo, diferenças entre genes ou diferenças entre espécies, e permitir investigar hipóteses de evolução dos genes e das próprias espécies. Um exemplo concreto é a comparação da distribuição empírica das distâncias inter-nucleótidos com a distribuição geométrica, que constitui o padrão de referência. A distância inter-nucleótidos foi introduzida por Nair and Mahalakshmi (2005) e corresponde a uma forma de mapear a sequência de ADN, que é uma sequência simbólica, em uma ou mais sequências numéricas que traduz(em) o número de nucleóticos entre dois nucleótidos do mesmo tipo. Este mapeamento entre a sequência de símbolos/nucleótidos no conjunto $\{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$ e uma ou mais sequências de valores inteiros, preserva toda a informação das sequências, permitindo reconstruir sempre (e de forma únivoca) a sequência simbólica a partir das sequências de valores inteiros e vice-versa. Adicionalmente, este mapeamento tem uma interpretação simples e clara.

Num trabalho anterior do nosso grupo de investigação, Afreixo et al. (2009) estudaram as sequências inter-nucleóticos associadas a cada nucleótido para melhor investigar o perfil de diferentes espécies. Mostrou-se que a distribuição empírica das distâncias apresenta diferenças estatisticamente significativas em relação à distribuição de referência, isto é, a distribuição obtida no caso em que os nucleótidos se distribuem de forma aleatória e independente ao longo da sequência de ADN. Este padrão de referência é a distribuição geométrica, e há estudos que defendem que esta distribuição reflecte o *background* aleatório do ADN (Qi et al., 2004). As diferenças entre as distribuições das distâncias empíricas e as de referência mostraram-se muito úteis na discriminação de espécies e na identificação de grupos de espécies, cuja estrutura entre (grupos de) espécies foi obtida por classificação hierárquica usando medidas de dissemelhança sobre as diferenças empírica/referência. Além de permitirem fazer um agrupamento de espécies compatível com o conhecimento que provém da Biologia e da Genómica em relação à diferença entre espécies, a literatura também sugere que estas diferenças empírica/referência refletem adequadamente a evolução seletiva do DNA da espécie (Qi et al., 2004). Surge assim a necessidade de desenvolver modelos estatísticos que permitam acomodar, simultaneamente, os dois componentes do

ADN acima referidos: o *background* aleatório e a evolução seletiva do ADN da espécie, esta última reflectida nas diferenças à distribuição de referência. Estes modelos permitirão reter informação importante de uma sequência no conjunto (pequeno) dos seus parâmetros, o que, além de permitir diminuir muito a quantidade de informação para analisar, permite também trazer maior interpretabilidade na análise de resultados. Mais concretamente, este trabalho visa analisar a relação entre uma série numérica de distâncias inter-simbólicas e a estrutura de autocorrelação das sequências binárias de sucesso/insucesso da ocorrência do símbolo. No caso do DNA, isso permitirá investigar se as diferenças entre as distribuições de distâncias inter-simbólicas e as distribuições de referência, que refletem também a contribuição da evolução selectiva, estão relacionadas com a estrutura de autocorrelação da sequência de DNA.

Este artigo traz um resumo dos passos na construção da distribuição geométrica autorregressiva. Para tal, na secção 2 começa-se por definir o processo binário autorregressivo (BinAR), apresentando também as suas principais características e resultados importantes associados a estes processos. Ainda na secção 2 apresenta-se também a ideia da construção da distribuição geométrica autorregressiva. A secção 2.1 apresenta e desenvolve a estratégia para a estimação dos parâmetros do modelo BinAR e determinação da ordem óptima do modelo, que são também usados para definir a distribuição geométrica autorregressiva. Finalmente, na secção 3, é apresentado um exemplo de aplicação em dados de ADN mitocondrial, mostrando-se o desempenho dos modelos em 34 espécies diferentes. Numa nota final, é importante referir que este texto é um resumo de trabalho já publicado na literatura científica, pelo que se aconselha a leitura de Gouveia et al. (2017) para mais desenvolvimento neste tópico.

2 Construção da distribuição geométrica autorregressiva

A construção da distribuição geométrica autorregressiva começa por assumir que a sequência binária de sucesso/insucesso que a gera, é modelada por um processo binário do tipo autorregressivo. Os modelos autorregressivos de ordem p para séries binárias, designados neste trabalho por BinAR(p) decorrente do inglês *binary autorregressive model of order p*, foram introduzidos por Kanter (1975). Formalmente, um processo discreto binário (X_t) é um processo BinAR(p) se (X_t) satisfaz

$$X_t = \alpha_{t,1} X_{t-1} \oplus \cdots \oplus \alpha_{t,p} X_{t-p} \oplus \varepsilon_t, \tag{1}$$

onde $\oplus$ representa a operação aritmética módulo 2, ε_t é uma variável aleatória i.i.d. tal que $\varepsilon_t \sim B(1,\mu)$ e ε_t é independente de $(X_s)_{s<t}$. As variáveis $\alpha_{t,1}$, …, $\alpha_{t,p}$ na definição anterior são indicadores binários associados à geração da observação em X_t tais que $(\alpha_{t,1},\ldots,\alpha_{t,p},\beta_{t,0}) \sim \text{MULT}(1;\phi_1,\ldots,\phi_p,\varphi_0)$ com $\beta_{t,0} = 1 - \sum_{i=1}^{p} \alpha_{t,i}$, $\varphi_0 \in (0,1)$, e MULT representa uma distribuição Multinomial. Segundo este esquema, X_t ou é igual a $X_{t-i} \oplus \varepsilon_t$ com probabilidade ϕ_i, para $i = 1,\ldots,p$, ou é igual a ε_t com probabilidade φ_0 sendo

$$\varphi_0 = 1 - \sum_{i=1}^{p} \phi_i. \tag{2}$$

A equação (1) pode ser simplesmente escrita como

$$X_t = \begin{cases} X_{t-1} \oplus \varepsilon_t & \text{com probabilidade } \phi_1 \\ X_{t-2} \oplus \varepsilon_t & \text{com probabilidade } \phi_2 \\ \vdots & \\ X_{t-p} \oplus \varepsilon_t & \text{com probabilidade } \phi_p \\ \varepsilon_t & \text{com probabilidade } \varphi_0 \end{cases}, \tag{3}$$

onde ϕ_i representa a probabilidade de activar o i-ésimo ramo da distribuição Multinomial e φ_0 representa a fracção de tempo em que X_t é igual a ε_t.

No trabalho introdutório sobre estes modelos BinAR, Kanter (1975) estabeleceu dois resultados muito importantes que são utilizados neste trabalho. O primeiro estabelece a relação entre os parâmetros ν e μ das distribuições $X_t \sim B(1,\nu)$ e $\varepsilon_t \sim B(1,\mu)$ respectivamente, por intermédio de

$$\nu = \frac{\mu}{1-(1-2\mu)(1-\varphi_0)}. \tag{4}$$

O segundo resultado mostra que a estrutura de autocorrelação de um processo BinAR é descrita através da relação

$$\rho(k) := Corr(X_t, X_{t-k}) = (1-2\mu)\sum_{i=1}^{p}\phi_i\,\rho(|k-i|) \text{ for } k \geq 1, \tag{5}$$

o que corresponde ao conjunto de equações de Yule-Walker estabelecido para, de forma recursiva, calcular as autocorrelações de um processo AR convencional com parâmetros $\phi_i^* = (1-2\mu)\phi_i$, $i = 1,\ldots,p$. Assim, é possível obter estimativas iniciais para os parâmetros do modelo BinAR por intermédio dos métodos convencionais para a obtenção de estimativas dos parâmetros de um modelo AR.

Um outro resultado importante relacionado com os modelos BinAR, foi mais recentemente publicado no trabalho de Weiß (2009), que deduziu a expressão para a probabilidade de transição do processo X_t como

$$\begin{aligned} &P(X_t = 1 \mid X_{t-1} = x_{t-1},\ldots,X_{t-p} = x_{t-p}) \\ &=: P(X_t = 1 \mid x_{t-1},\ldots,x_{t-p}) = \mu + (1-2\mu)\sum_{i=1}^{p}\phi_i\,x_{t-i}, \end{aligned} \tag{6}$$

escrevendo-a como uma função linear dos parâmetros BinAR e do histórico recente observado do processo. Note-se que $X_t \equiv \varepsilon_t$ e, portanto $X_t \sim B(1,\nu=\mu)$, para o caso em que $\phi_i = 0$, $\forall i$.

Seja agora Y a variável aleatória (v.a.) que representa o número de experiências até ocorrer o primeiro sucesso. O trabalho de Gouveia et al. (2017) mostrou que, se X_t for um processo binário com a probabilidade de transição dada na equação 6, então

$$P(Y=k) = P(X_{t+k-1}=1, X_{t+k-2}=0,\ldots,X_t=0 \mid 1, x_{t-2},\ldots,x_{t-p}) = f_k\prod_{i=1}^{k-1}(1-f_i), \tag{7}$$

onde as probabilidades de sucesso f_i para $i = 1,\ldots,k$ são dadas pela expressão

$$f_i = \begin{cases} \mu + \phi_i^* + \sum\limits_{j=i+1}^{p}\phi_j^* x_{t+i-j-1} & i \leq p, \\ \mu & i > p \end{cases} \tag{8}$$

e $\phi_i^* = (1-2\mu)\phi_i$ para $i = 1,\ldots,p$. Note que se X_t for uma sequência binária de valores independentes (isto é, $\phi_i = 0$, $\forall i$) então é bem conhecido que Y seguirá uma distribuição de probabilidade geométrica, onde $f_k = \mu$, $\forall k$ representa a probabilidade de sucesso (constante) em cada experiência.

2.1 Estimação dos parâmetros

As estimativas para os $p+1$ parâmetros μ e $\phi = [\phi_1 \cdots \phi_p]^{\mathrm{T}}$ são obtidas por maximização da função de log-verossimilhança condicional

$$\ell(\mu,\phi,X_t) := \sum_{t=p+1}^{\mathcal{L}}\log P(X_t = 1 \mid x_{t-1},\ldots,x_{t-p}), \tag{9}$$

para uma ordem p fixa e sendo $\mathcal{L}$ o número de observações do processo. No entanto, devido às propriedades específicas de X_t (por exemplo, relações entre parâmetros ou o facto dos coeficientes ϕ_i serem probabilidades), é inevitável terem de ser introduzidas algumas restrições na solução óptima do problema de maximização. De facto, Gouveia et al. (2017) mostraram que o problema de optimização pode ser escrito como

$$\begin{aligned} \max_{\mu,\phi_1,\ldots,\phi_p} \quad & \ell(\mu,\phi,X_t) \\ \text{sujeito a} \quad & \nu - \frac{\mu}{1-(1-2\mu)(1-\varphi_0)} = 0 \\ & 0 \leq \mu \leq \nu \\ & 0 \leq \phi_i \leq U_i,\ i = 1,2,\ldots,p. \end{aligned}$$

Este problema de optimização inicia-se com a estimativa para ν, a qual é obtida através da média amostral de X_t. Os valores iniciais para $(\hat{\mu},\hat{\phi})$ são determinados de forma eficiente (via o algoritmo de Levinson-Durbin) através das p equações de Yule-Walker, pela equivalência entre a estrutura de dependência do modelo BinAR e a do modelo AR convencional, e da equação (4). Adicionalmente, a significância estatística dos coeficientes do modelo é quantificada através dos respectivos erros padrão, obtidos pela raíz quadrada dos elementos da diagonal da inversa da matriz Hessiana da função objectivo ℓ. Finalmente, a estimativa de φ_0 é obtida através da equação (2).

No problema de optimização acima indicado, a restrição na solução corresponde à relação entre os parâmetros ν and μ das distribuições de X_t e ε_t respectivamente, estabelecida no trabalho de Kanter (1975). Desta restrição é possível também deduzir o primeiro limite à solução, isto é $0 \leq \mu \leq \nu$ onde $\nu = \mu$ no caso em que $\varphi_0 = 1$ (que acontece unicamente se $\phi_i = 0, \forall i$, ver equação (2)). Adicionalmente, ainda é possível estabelecer que $0 \leq \phi_i \leq 1$, uma vez que ϕ_i é uma probabilidade. No entanto, o espaço de valores possíveis para ϕ_i poderá ser ainda mais restrito tendo em conta um limite superior U_i determinado pelas estimativas iniciais: $U_i = 0$ para os coeficientes negativos ou não significativos e, para os restantes coeficientes, U_i é constante e é obtido pela equação (2) assumindo que $\varphi_0 = 0$, o que permite determinar um valor máximo para a soma dos parâmetros significativos.

Um outro ponto importante no processo de estimação é a identificação da ordem p do modelo BinAR. Neste trabalho consideraram-se modelos BinAR com ordens entre 0 (o caso da independência, que reproduz a distribuição geométrica convencional) e 15. O valor para p foi escolhido como a ordem que minimiza a discrepância entre as frequências observadas e as frequências esperadas pelo modelo BinAR(p), denotada neste trabalho por d_p, e avaliada por intermédio de uma estatística Qui-quadrado devidamente ajustada pelo número de parâmetros do modelo e tamanho da amostra (Pederson and Johnson, 1990).

3 Aplicações no contexto de análise de sequências de ADN

Este trabalho apresenta uma ilustração do uso das distribuições geométricas autorregressivas em dados genómicos mitocondriais de 34 espécies, incluindo primatas, carnívoros e outros, obtidos do projecto GenBank (http://ncbi.nlm.nih.gov/genbank). Estas sequências de ADN têm pelo menos 16000 nucleótidos (ou seja, $\mathcal{L} > 16000$) com proporções de ocorrência na sequência ν que variam entre 12% e 35% consoante o nucleótido e a espécie.

A sequência de ADN de cada espécie foi convertida em 4 séries binárias $X_t^{\mathcal{N}}$, $t = 1,\ldots,\mathcal{L}$ com valores no conjunto $\{0,1\}$ onde 1 representa a ocorrência do nucleótido $\mathcal{N} \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, i.e.

$$X_t^{\mathcal{N}} = \begin{cases} 1, & \mathcal{N} \text{ ocorre na posição } t \text{ da sequência,} \\ 0, & \text{outros casos.} \end{cases} \tag{10}$$

De seguida, constituiu-se a variável numérica $Y^{\mathcal{N}}$ que representa a distância inter-nucleótidos para o nucleótido $\mathcal{N}$. Por exemplo, para o fragmento AAACCCGTGTCAGTT de ADN, as séries $X_t^{\mathcal{N}}$ e $Y^{\mathcal{N}}$ para $\mathcal{N} \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$ são as seguintes:

$$\begin{aligned}
X_t^{\mathrm{A}} &: 111000000001000 \rightarrow Y^{\mathrm{A}} : 1,1,9,4\\
X_t^{\mathrm{C}} &: 000111000010000 \rightarrow Y^{\mathrm{C}} : 1,1,5,8\\
X_t^{\mathrm{G}} &: 000000101000100 \rightarrow Y^{\mathrm{G}} : 2,4,9\\
X_t^{\mathrm{T}} &: 000000010100011 \rightarrow Y^{\mathrm{T}} : 2,4,1,8
\end{aligned}$$

Neste mapeamento, os últimos elementos em $Y^{\mathcal{N}}$ são obtidos através de uma extensão circular da sequência simbólica assegurando-se assim, uma correspondência única entre $Y^{\mathcal{N}}$ and $X_t^{\mathcal{N}}$. Note-se que a soma dos valores de $Y^{\mathcal{N}}$ é igual ao tamanho da sequência binária $X_t^{\mathcal{N}}$ e que o número de nucleótidos do tipo $\mathcal{N}$ é igual ao tamanho de $Y^{\mathcal{N}}$.

A figura 1 apresenta alguns dos resultados obtidos para a sequência de ADN da espécie humana, sendo possível comparar as distribuições empíricas das distâncias inter-nucleótidos, com as distribuições esperadas segundo o modelo de independência e segundo o modelo BinAR(p). Em geral, os valores de ν obtidos para A e G são, respectivamente, os maiores e os menores para cada espécie e, portanto, a distribuição das distâncias inter-nucleótidos para A e G exibem, respectivamente, as caudas mais curtas e mais longas em relação às distribuições dos restantes nucleótidos. Em relação ao ajuste dos modelos, é possível observar que as frequências esperadas segundo a distribuição geométrica autorregressiva (linhas a preto) são bastante melhor ajustadas às frequências observadas (barras) em relação às da distribuição geométrica (linhas a cinzento). Este padrão é transversal a todos os nucleótidos e a todas as espécies analisados onde, predominantemente, a ordem do modelo é $p \leq 6$.

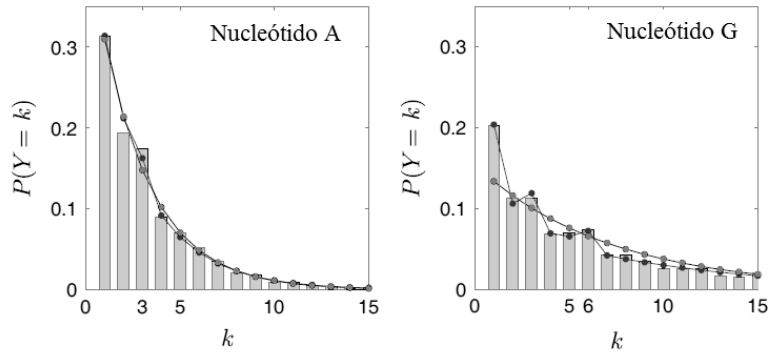


Figura 1: Função de probabilidade de Y^{A} e Y^{G} (espécie humana), onde as barras representam as frequências observadas para cada distância k. As linhas a preto e a cinza representam, respectivamente, a distribuição geométrica autorregressiva de ordem p e a distribuição de referência geométrica, com parâmetros estimados a partir dos dados. A ordem escolhida para os modelos BinAR foi $p = 3$ e $p = 6$, respectivamente, para os nucleótidos A e G.

Os modelos desenvolvidos neste trabalho foram avaliados tendo em conta dois aspectos. O primeiro quantifica o ajuste do modelo BinAR(p) à sequência binária X_t, para cada nucleótido e para cada espécie, por intermédio de

$$\mathcal{G} = \sum_{i=1}^{p} \phi_i. \tag{11}$$

Esta medida representa a probabilidade de activar um ramo da distribuição Multinomial relacionado com a história do processo X_t, uma vez que ϕ_i representa a probabilidade de ativar o ramo da distribuição Multinomial associado com X_{t-i} (ver equação (3)). O segundo aspecto quantifica a diminuição da discrepância ao considerar a distribuição geométrica autoregressiva de ordem p ao invés da distribuição geométrica, através de

$$\mathcal{R} = \frac{d_0 - d_p}{d_0}, \tag{12}$$

onde d_0 e d_p são as discrepâncias entre a distribuição empírica e, respectivamente, a distribuição geométrica e a distribuição geométrica autorregressiva. Assim, $\mathcal{R}$ mede a diminuição de discrepância como uma percentagem de d_0.

A figura 2 apresenta os resultados da avaliação dos modelos para os vários nucleótidos e espécies. Os resultados indicam que $\mathcal{G}$, a métrica de ajuste dos modelos, é superior para o nucleótido G (com valores na ordem dos $\approx 20\%$), seguido do nucleótido T ($\approx 15\%$), sendo que o valor para G é quase o dobro do valor avaliado para A e para C ($\approx 10\%$). Assim, os valores de $\mathcal{G}$ são inferiores a $\approx 35\%$ para todos os nucleótidos e todas as espécies. No entanto, este ajuste não muito elevado dos modelos tem um grande impacto na diminuição de discrepância entre a distribuição empírica e a distribuição geométrica. De facto, a figura 2 mostra que $\mathcal{R}$, a diminuição de discrepância, está bem acima dos 50% para os todos os nucleótidos, com excepção do nucleótido A, que apresenta uma diminuição mediana de discrepância na ordem dos $\approx 30\%$. Para este nucleótido, tal como ilustrado na figura 1 para a espécie humana, existe ainda um desvio evidente nas frequências em $1 \leq k \leq 4$, que o modelo BinAR de ordem p não consegue corrigir devidamente.

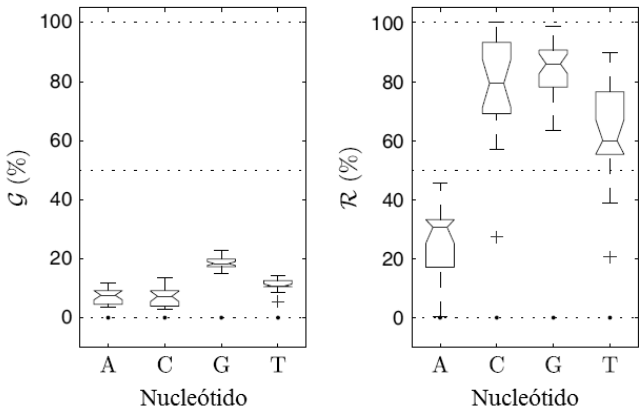


Figura 2: Boxplot da medida de ajuste do modelo BinAR(p) $\mathcal{G}$ e da medida de diminuição de discrepância $\mathcal{R}$, para as 34 espécies.

Assim, de uma forma geral, a diferença entre as frequências observadas das distâncias inter-nucleótidos e as frequências esperadas segundo o modelo de independência são explicadas pelos modelos BinAR. Este resultado tem grande impacto na análise de sequências de ADN pois, além de se modelar (quase na totalidade) a forma da distribuição observada, também a ordem dos modelos considerados é $p \leq 6$, o que obriga, no máximo, à estimação de 7 parâmetros por nucleótido. Assim, é claro que o uso de distribuições geométricas autorregressivas na análise de sequências de ADN poderá ser uma grande mais valia na determinação de perfis de ADN que permitam avaliar diferenças (ou semelhanças) entre espécies.

Agradecimentos

Este trabalho foi parcialmente financiado pela Fundação para a Ciência e a Tecnologia, I.P. (FCT) através de fundos nacionais do Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) e pelo Fundo Europeu de Desenvolvimento Regional (FEDER), no âmbito dos projectos IEETA (com referência UID/CEC/00127/2019) e CIDMA (com referência UID/MAT/04106/2019) da Universidade de Aveiro.

Referências

Gouveia, S., Scotto, M. G., Weiß, C. H., Ferreira, P. J. S. G. (2017) Binary autoregressive geometric modelling in a DNA context. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 66(2), 253–271.

Afreixo, V., Bastos, C. A. C., Pinho, A. J, Garcia, S. P. and Ferreira, P. J. S. G. (2009) Genome analysis with inter-nucleotide distances. *Bioinformatics*, **25**, 3064–3070.

Bastos, C. A., Afreixo, V., Pinho, A. J., Garcia, S. P., Rodrigues, J. M. and Ferreira, P. J. S. G. (2011) Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions. *Journal of Integrative Bioinformatics*, **8**, 172–184.

Cochran, W. G. (1954) Some methods for strengthening the common chi-squared tests. *Biometrics*, **10**, 417–451.

Kanter, M. (1975) Autoregression for discrete processes mod 2. *Journal of Applied Probability*, **12**, 371–375.

Nair, A. S. S. and Mahalakshmi, T. (2005) Visualization of genomic data using inter-nucleotide distance signals. *Proceedings of the IEEE International Conference on Genomic Signal Processing*, Bucharest, Romania.

Pederson, S. P. and Johnson, M. E. (1990) Estimating model discrepancy. *Technometrics*, **32**, 305–314.

Qi, J., Wang, B. and Hao, B. I. (2004) Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution*, **58**, 1–11.

Weiß, C. H. (2009) Properties of a class of binary ARMA models. *Statistics*, **43**, 131–138.

Cartas de controlo para o valor esperado de um processo INAR(1) com função ARL sem viés

Manuel Cabral Morais, *maj@math.ist.utl.pt*

CEMAT e Departamento de Matemática,
Instituto Superior Técnico, Universidade de Lisboa

1. Introdução

Em controlo de qualidade é habitual lidarmos com um processo $\{X_t : t \in \mathbb{N}_0\}$ associado à série temporal do número de defeitos em amostra de dimensão fixa. Mais, é usual assumir-se que as variáveis aleatórias (v.a.) X_t são independentes e identicamente distribuídas (i.i.d.) com distribuição marginal de Poisson. Contudo, estas v.a. são frequentemente autocorrelacionadas, facto que restringe substancialmente a utilização daquela que é, indubitavelmente, a carta de controlo de qualidade mais popular na detecção de alterações no número esperado de defeitos em amostras de dimensão fixa, a carta-c com limites 3-sigma descrita, por exemplo, em Montgomery (2009, pag. 309).

Há pouco mais de uma década, Weiß (2007) propôs uma carta para controlar o valor esperado de um processo inteiro autoregressivo de primeira ordem (INAR(1)) com marginais de Poisson. Representemo-lo por $\{X_t = \beta \circ X_{t-1} + \epsilon_t : t \in \mathbb{N}_0\}$, onde: $\beta \in (0,1)$; $\circ$ representa o operador de *thinning* binomial; $\epsilon_t \sim_{iid} \text{Poisson}(\lambda)$, $t \in \mathbb{N}$; ϵ_t e X_{t-1} são v.a. independentes; todas as operações de *thinning* binomial são independentes de $\{\epsilon_t : t \in \mathbb{N}\}$ e de $\{\ldots, X_{t-2}, X_{t-1}\}$ e, para além disso, são efectuadas de modo independente entre si.

Weiß (2007) reajustou os limites 3-sigma da carta-c de modo a ter-se em conta a autocorrelação do processo INAR(1) com marginais de Poisson e a detectar alterações no respectivo valor esperado do valor-alvo $\mu_0 = \lambda_0/(1-\beta_0)$ para $\mu = \lambda/(1-\beta)$.

A utilização da carta-c *modificada* daí resultante pressupõe o registo sequencial do número observado de defeitos em amostras de dimensão fixa, x_t, num gráfico com limite inferior de controlo (*lower control limit*, LCL) e limite superior de controlo (*upper control limit*, UCL) escritos à custa do tecto e da parte inteira seguintes:

$$LCL = \left\lceil \max\left\{0, \frac{\lambda_0}{1-\beta_0} - k\sqrt{\frac{\lambda_0}{1-\beta_0}}\right\}\right\rceil \quad \text{e} \quad UCL = \left\lfloor \frac{\lambda_0}{1-\beta_0} + k\sqrt{\frac{\lambda_0}{1-\beta_0}}\right\rfloor, \qquad (1)$$

onde k é uma constante real positiva, usualmente igual a 3 ou escolhida de tal forma que o valor esperado do número de amostras recolhidas até à emissão de um sinal (*average run length*, ARL) é relativamente elevado, quando o processo está sob controlo (leia-se quando $\mu = \mu_0 = \lambda_0/(1-\beta_0)$).

Esta carta de controlo de qualidade possui algumas desvantagens. Caso o valor esperado alvo $\mu_0 = \lambda_0/(1-\beta_0)$ não exceda k^2, temos $LCL = 0$ e, consequentemente, a carta-c modificada é incapaz de detectar de forma expedita qualquer diminuição no valor esperado do processo. Na verdade, mesmo que o LCL seja positivo esta carta leva mais tempo, em média, a detectar determinadas diminuições no valor esperado do processo que a emitir um falso alarme, pois a

função ARL não atinge o seu valor máximo quando $\mu = \mu_0 = \lambda_0/(1-\beta_0)$, ou seja, a função ARL possui viés, como bem ilustra a Figura 1 de Paulino *et al.* (2016b). Por este motivo lidamos com uma carta do tipo *ARL-biased*, designação esta que se deve a Pignatiello *et al.* (1995), ou seja, com *uma carta com função ARL com viés*.

A carta-c modificada padece de outro problema grave: devido ao carácter discreto de X_t, não é possível seleccionar os seus limites de controlo de forma a que o ARL sob controlo da carta seja exactamente igual a um valor pré-especificado $ARL^{\star}$.

O que se segue é um apanhado de Paulino *et al.* (2016b), que constitui um estudo aturado sobre as cartas-c *modificadas* com função ARL sem viés, i.e., do tipo *ARL-unbiased*, para controlar o valor esperado de processos INAR(1) com marginais de Poisson. Este parâmetro pode representar o número esperado de defeitos em amostras de dimensão fixo, ou o número esperado de camas ocupadas em intervalos de 5 minutos numa sala de exames do serviço de urgência de um hospital pediátrico, ou o valor esperado de qualquer outra característica de qualidade associada a um processo INAR(1) com marginais de Poisson.

2. Cartas-c modificadas para $\mu = \lambda/(1-\beta)$ com função ARL sem viés

Convém recordar que $\{X_t = \beta \circ X_{t-1} + \epsilon_t : t \in \mathbb{N}_0\}$ é uma cadeia de Markov em tempo discreto com espaço de estados $\mathbb{N}_0$ e probabilidades de transição $p_{i\,j} \equiv p_{i\,j}(\lambda, \beta)$ dadas por

$$p_{i\,j} = \sum_{m=0}^{\min\{i,j\}} \binom{i}{m} \beta^m \, (1-\beta)^{i-m} \times \frac{e^{-\lambda}\, \lambda^{j-m}}{(j-m)!}, \quad i, j \in \mathbb{N}_0, \tag{2}$$

de acordo com Weiß (2009, pag. 421).

À semelhança da carta-c com função ARL sem viés proposta por Paulino *et al.* (2016a) para controlar o valor esperado de um processo i.i.d. com marginais de Poisson, a eliminação do viés da função ARL da carta-c modificada pressupõe que procedamos do seguinte modo. Ao recolher a t-ésima amostra, deveremos emitir um sinal com:

- probabilidade um, caso $x_t < L$ ou $x_t > U$;
- probabilidade γ_L (resp. γ_U), caso $x_t = L$ (resp. $x_t = U$).

A obtenção dos limites de controlo, L e U, e das probabilidades de aleatorização, γ_L e γ_U, passa pelo recurso a um procedimento de pesquisa iterativo não trivial, descrito em detalhe em Paulino *et al.* (2016b) e implementado no *software* estatístico R (R Core Team, 2013).

A aleatorização da emissão do sinal, quando $x_t = L$ (resp. $x_t = U$), faz-se na prática incorporando a geração de um número pseudo-aleatório da distribuição de Bernoulli com parâmetro γ_L (resp. γ_U) no *software* usado para o tratamento dos dados provenientes da linha de produção (Paulino *et al.*, 2016b).

Importa notar ainda que a aleatorização da emissão do sinal significa também que a função ARL da carta-c modificada com função ARL sem viés está relacionado com o tempo esperado até absorção de uma cadeia de Markov cujas transições entre os respectivos estados transeuntes são regidas pela matriz sub-estocástica

$$\mathbf{Q}(\lambda, \beta, \gamma_L, \gamma_U) = \begin{bmatrix} (1-\gamma_L)\, p_{L\,L} & p_{L\,L+1} & \cdots & p_{L\,U-1} & (1-\gamma_U)\, p_{L\,U} \\ (1-\gamma_L)\, p_{L+1\,L} & p_{L+1\,L+1} & \cdots & p_{L+1\,U-1} & (1-\gamma_U)\, p_{L+1\,U} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ (1-\gamma_L)\, p_{U-1\,L} & p_{U-1\,L+1} & \cdots & p_{U-1\,U-1} & (1-\gamma_U)\, p_{U-1\,U} \\ (1-\gamma_L)\, p_{U\,L} & p_{U\,L+1} & \cdots & p_{U\,U-1} & (1-\gamma_U)\, p_{U\,U} \end{bmatrix}. \tag{3}$$

Se não, vejamos. Consideremos que o número de amostras recolhidas até à emissão de um sinal (*run length*, RL), condicional a $X_0 = u$ ($u \in \{L, L+1, \ldots, U-1, U\}$), é representado por

$$RL^u(\lambda, \beta, \gamma_L, \gamma_U) = \min\{t \in \mathbb{N} : X_t < L \text{ or } X_t > U \mid X_0 = u\}. \tag{4}$$

Então

$$ARL^u(\lambda, \beta, \gamma_L, \gamma_U) = \underline{\mathbf{e}}_u^\top \times [\mathbf{I} - \mathbf{Q}(\lambda, \beta, \gamma_L, \gamma_U)]^{-1} \times \underline{\mathbf{1}}, \tag{5}$$

onde:

- $\mathbf{e}_u^\top$ é o $(U - L + 1)$-ésimo vector da base ortonormada de $\mathbb{R}^{(U-L+1)}$;
- $\mathbf{I}$ representa a matriz identidade com característica $(U - L + 1)$;
- $\underline{\mathbf{1}}$ é um vector-coluna com $(U - L + 1)$ uns.

Uma vez que o valor X_0 é usualmente desconhecido, é plausível admitir que $X_1 \equiv X_1(\lambda, \beta) \sim$ Poisson$(\lambda/(1-\beta))$ e adoptar a medida de desempenho recomendada por Weiß e Testik (2009), o *overall ARL*, que neste caso particular se escreve:

$$\begin{aligned} ARL(\lambda, \beta, \gamma_L, \gamma_U) &= 1 + (1 - \gamma_L) \times ARL^L(\lambda, \beta, \gamma_L, \gamma_U) \times P[X_1(\lambda, \beta) = L] \\ &\quad + \sum_{u=L+1}^{U-1} ARL^u(\lambda, \beta, \gamma_L, \gamma_U) \times P[X_1(\lambda, \beta) = u] \\ &\quad + (1 - \gamma_U) \times ARL^U(\lambda, \beta, \gamma_L, \gamma_U) \times P[X_1(\lambda, \beta) = U]. \end{aligned} \tag{6}$$

$ARL(\lambda, \beta, 0, 0)$ corresponde ao *overall ARL* de uma carta-c modificada com limites de controlo $LCL = L$ e $UCL = U$ e função ARL com viés.

Escusado será referir que, devido ao carácter discreto de X_t, é absolutamente crucial aleatorizar a emissão do sinal por forma a obter um *overall ARL* sob controlo exactamente igual ao valor pré-estipulado $ARL^\star$. Para além disso, *função ARL* e *overall ARL* designarão, de ora em diante e sinonimicamente, o desempenho de qualquer carta-c modificada.

3. Ilustrações

Os resultados, no exemplo que se segue, referem-se à comparação do *overall ARL* das:

- cartas-c modificadas com limites 3-sigma;
- cartas-c modificadas com função ARL sem viés tal que $ARL(\lambda, \beta, \gamma_L, \gamma_U)$ atinge valor máximo quando $\lambda = \lambda_0$ (resp. $\beta = \beta_0$) e que designaremos de cartas-c modificadas com função ARL sem viés em λ (resp. β).

Exemplo 1 — Os limites de controlo da carta-c modificada com limites 3-sigma são iguais a $[LCL, UCL] = [0, 2], [6, 30]$, para $(\lambda_0, \beta_0) = (0.5, 0.2), (9, 0.5)$.

Na Tabela **??**, encontramos os limites de controlo, as probabilidades de aleatorização e alguns valores do *overall ARL* das cartas-c modificadas com função ARL sem viés em λ e sem viés em β, listados exactamente por esta ordem. Em qualquer dos casos, o *overall ARL* sob controlo é igual a $ARL^\star = 370.4$, daí que $ARL(\lambda_0, \beta_0, \gamma_L, \gamma_U)$ tenha sido omitido desta tabela.

A Tabela **??** e a Tabela 2 de Paulino *et al.* (2016b) levam a crer que os parâmetros das cartas-c modificadas com função ARL sem viés em λ pouco se distinguem dos das cartas-c modificadas

Tabela 1: Parâmetros e valores de $ARL(\lambda, \beta, \gamma_L, \gamma_U)$ das cartas-c modificadas com função ARL sem viés em λ (linha ímpar) e sem viés em β (linha par) — $(\lambda_0, \beta_0) = (0.5, 0.2), (9, 0.5)$ e $ARL^\star = 370.4$.

λ_0	β_0	$[L, U]$	(γ_L, γ_U)	$1.2\lambda_0$ β_0	$1.01\lambda_0$ β_0	$0.99\lambda_0$ β_0	$0.8\lambda_0$ β_0	λ_0 $1.2\beta_0$	λ_0 $1.01\beta_0$	λ_0 $0.99\beta_0$	λ_0 $0.8\beta_0$
0.5	0.2	$[0, 5]$	$(0.004432, 0.661663)$	352.9	370.4	370.4	356.4	369.5	370.4	370.4	369.4
		$[0, 5]$	$(0.004423, 0.674057)$	352.0	370.3	370.4	356.9	369.3	370.4	370.4	369.5
9	0.5	$[7, 32]$	$(0.304943, 0.363371)$	74.6	367.5	367.6	75.5	53.4	368.2	366.9	99.8
		$[7, 32]$	$(0.291560, 0.393249)$	73.6	366.8	368.3	76.4	52.8	367.5	367.6	101.0

com função ARL sem viés em β. Como se isso não bastasse, Paulino *et al.* (2016b) constataram que é muito raro a carta-c modificada com função ARL sem viés em λ (resp. β) verificar $ARL(\lambda_0, \beta, \gamma_L, \gamma_U) > ARL^\star$ (resp. $ARL(\lambda, \beta_0, \gamma_L, \gamma_U) > ARL^\star$), caso $\beta \neq \beta_0$ (resp. $\lambda \neq \lambda_0$). Consequentemente, atrevemo-nos a afirmar que as cartas-c modificadas com função ARL sem viés em λ e sem viés em β são intercambiáveis.

Da Figura **??** constam os gráficos do *overall ARL* das cartas-c modificadas com limites 3-sigma e com função ARL sem viés em λ (à esquerda) e sem viés em β (à direita), para $(\lambda_0, \beta_0) = (0.5, 0.2), (9, 0.5)$ e $ARL^\star = 370.4$.

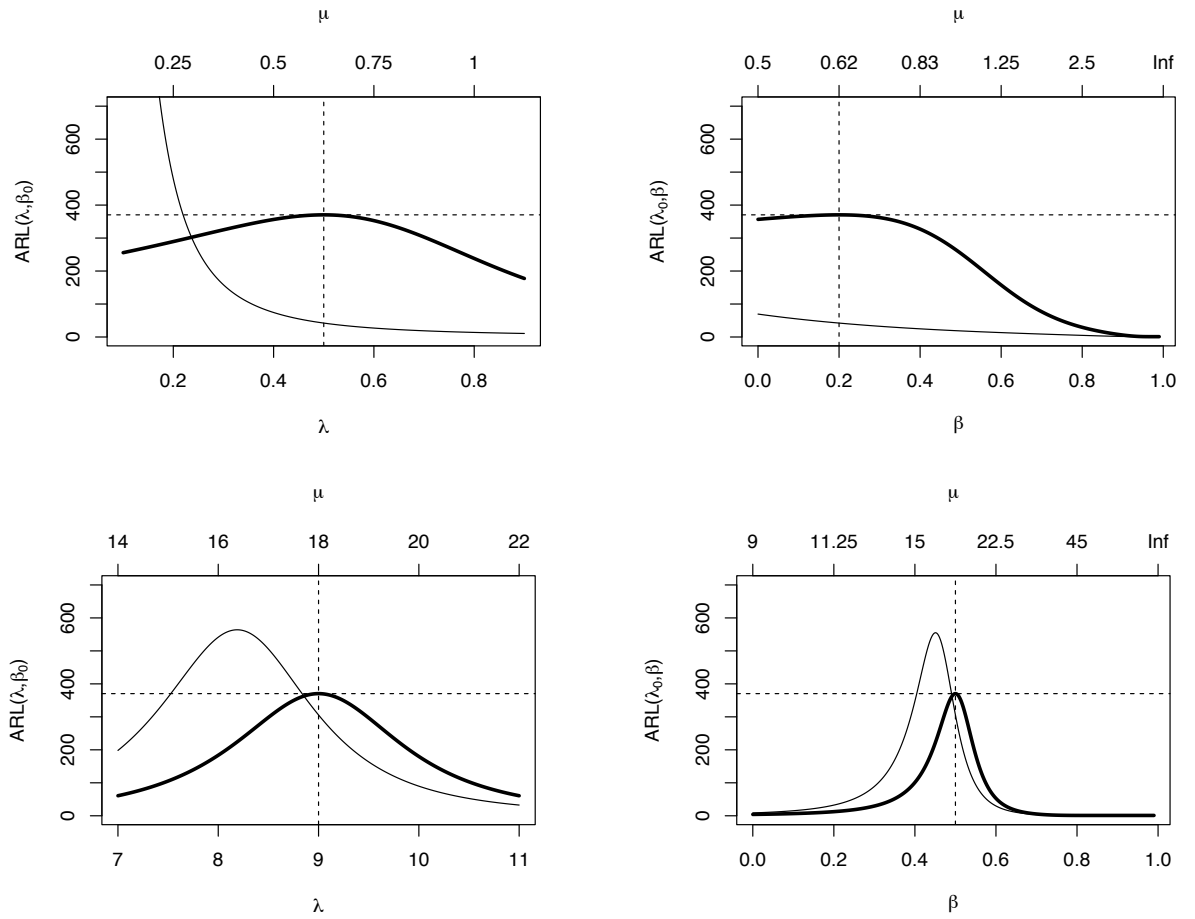


Figura 1: Gráficos de *overall ARL* enquanto função de λ (à esquerda) e de β (à direita) das cartas-c modificadas com limites 3-sigma e com ARL sem viés — $(\lambda_0, \beta_0) = (0.5, 0.2), (9, 0.5)$ e $ARL^\star = 370.4$.

Estes gráficos permitem concluir que as cartas-c modificadas com limites 3-sigma possui *overall ARL* sob controlo indesejavelmente abaixo do valor pré-especificado $ARL^\star$. Por este motivo e pelo facto de não possuírem função ARL com máximo em $\mu = \lambda_0/(1 - \beta_0)$, estas cartas são mais

rápidas a detectar aumentos em μ que as cartas-c modificadas com função ARL sem viés. No entanto, aquelas são incapazes de detectar diminuições no valor esperado do processo μ em tempo útil.

Caso μ represente o número esperado de defeitos numa amostra de dimensão fixa, podemos afirmar que as cartas-c modificadas com função ARL sem viés são, ao contrário das cartas-c modificadas com limites 3-sigma, sensíveis não só a melhorias como também a piorias da qualidade do processo de fabrico. ■

Exemplo 2 — Weiß e Testik (2011) consideraram o número de camas ocupadas em intervalos de 5 minutos (das 08:00 às 23:55) numa sala de exames do serviço de urgência de um hospital pediátrico e concluíram que um modelo INAR(1) com marginais de Poisson, com parâmetros $(\mu_0 = \lambda_0/(1-\beta_0), \beta_0) = (2.44, 0.81)$, se adequa ao conjunto de dados.

Uma vez que, do ponto de vista prático, importa detectar quer diminuições, quer aumentos no número esperado de camas ocupadas em intervalos de 5 minutos, é conveniente recorrer a uma carta-c modificada com função ARL sem viés em λ (resp. β) de modo a controlar alterações em tal parâmetro devido a *shifts* em λ (resp. β).

Admitamos que o valor-alvo de (λ, β) é $(\lambda_0, \beta_0) = (0.4636, 0.81)$ e que $ARL^\star = 500$. Neste caso, os parâmetros das cartas-c modificadas com função ARL sem viés em λ e sem viés em β são

- $[L, U] = [0, 9]$ e $(\gamma_L, \gamma_U) = (0.01707, 0.945655)$
- $[L, U] = [0, 8]$ e $(\gamma_L, \gamma_U) = (0.016434, 0.018586)$, respectivamente.

Os perfis de *overall ARL* associados a estas duas cartas encontram-se na Figura **??** e garantem que estas cartas permitem, efectivamente, detectar aumentos e diminuições no valor esperado do processo devido a alterações em λ (resp. β).

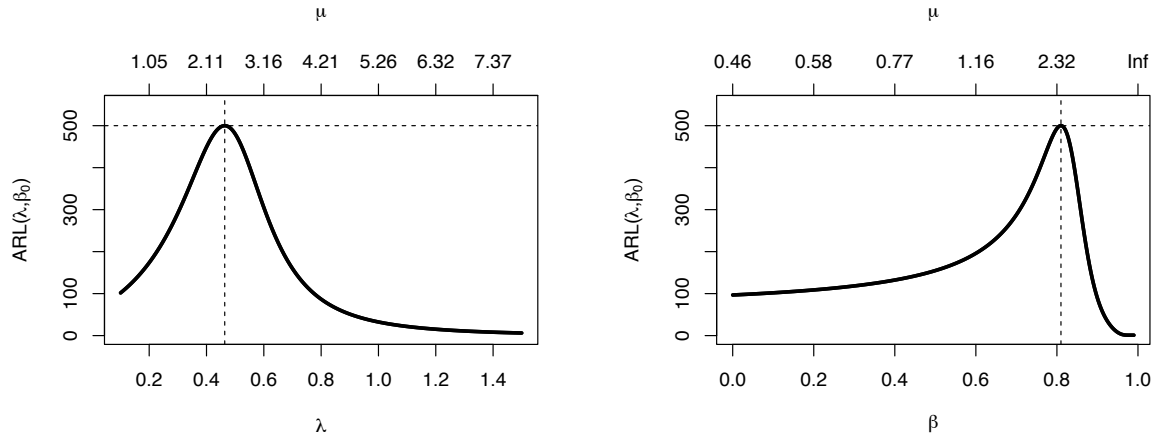


Figura 2: Gráficos de *overall ARL* enquanto função de λ (à esquerda) e de β (à direita) das cartas-c modificadas com ARL sem viés — $(\lambda_0, \beta_0) = (0.4636, 0.81)$ e $ARL^\star = 500$.

A título meramente ilustrativo, consideremos a carta-c modificada com função ARL sem viés em λ da Figura **??**, bem como as 40 observações simuladas que dela constam, das quais as 20 primeiras sob controlo e as 20 restantes associadas a uma alteração no valor esperado do processo de $\mu_0 = 2.44$ para $\mu_0 + 4 = 6.44$ devido a um aumento de $\lambda_0 = 0.4636$ para $\lambda_0 + 4 \times (1 - \beta_0) = 1.2236$. A carta compreende ainda os limites de controlo $[L, U] = [0, 9]$ e o valor esperado alvo $\mu_0 = 2.44$.

Acrescentemos também que um • corresponde a uma observação responsável por um sinal ou porque está para além dos limites de controlo, ou porque coincide com LCL (resp. UCL) e a correspondente geração do número pseudo-aleatório da distribuição de Bernoulli com parâmetro γ_L (resp. γ_U) é igual a 1.

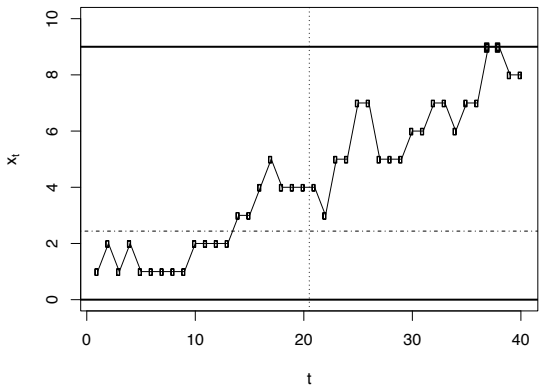


Figura 3: Carta-c modificada com função ARL sem viés em λ, para o número esperado de camas em intervalos de 5 min — $(\lambda_0, \beta_0) = (0.4636, 0.81)$, $[L, U] = [0, 9]$, $(\gamma_L, \gamma_U) = (0.017070, 0.945655)$.

A consulta da Figura **??**, leva-nos a concluir que as observações 37 e 38 são iguais ao UCL e responsáveis pela emissão de dois sinais válidos graças à aleatorização associada à carta-c modificada com função ARL sem viés em λ.

Por fim e uma vez que $ARL(\lambda_0 + 4 \times (1 - \beta_0), \beta_0, \gamma_L, \gamma_U) \simeq 14.0$, podemos afirmar que o primeiro destes dois sinais válidos ocorreu 3 observações depois do esperado. ■

4. À laia de conclusão

Chegadas/os aqui, o artigo termina necessariamente reafirmando que, ao contrário da carta-c modificada com limites k-sigma, a carta-c modificada com função ARL sem viés:

- permite que pré-especifiquemos o ARL sob controlo;
- emite sinais válidos, em média, mais rapidamente do que falsos alarmes;
- é capaz de detectar diminuições no valor esperado do processo em tempo razoável mesmo quando o limite inferior de controlo é nulo pois a carta depende de duas probabilidades de aleatorização.

Convém adiantarmos também que Clara (2018, Cap. 3) propõe um esquema bilateral constituído por um par de cartas unilaterais de somas acumuladas (*cumulative sum*, CUSUM) com função ARL sem viés para o valor esperado de processos INAR(1) com marginais de Poisson. Este esquema bilateral tem por objectivo detectar diminuições e aumentos de pequena ou média magnitude de modo mais célere que a carta-c modificada com função ARL sem viés.

(Por decisão pessoal, o autor deste texto não escreve segundo o Acordo Ortográfico de 1990.)

Agradecimentos

Muito agradecemos ao Editor do Boletim da SPE, Prof. Fernando Rosado, e ao Prof. Manuel Scotto a oportunidade de divulgação deste trabalho.

Este trabalho foi parcialmente financiado pela FCT (Fundação para a Ciência e Tecnologia) através do projecto UID/Multi/04621/2013.

Referências

Clara, M.F.V. (2018). *Cartas CUSUM ARL-unbiased para processos i.i.d. e INAR(1) com marginais de Poisson* (título provisório). Tese de mestrado. Instituto Superior Técnico, Universidade de Lisboa.

Montgomery, D.C. (2009). *Introduction to Statistical Quality Control* (6a. edição). New York: John Wiley & Sons.

Paulino, S., Morais, M.C. e Knoth, S. (2016a). An ARL-unbiased c-chart. *Quality and Reliability Engineering International* **32**, 2847–2858.

Paulino, S., Morais, M.C. e Knoth, S. (2016b). On ARL-unbiased c-charts for INAR(1) Poisson counts. *Statistical Papers*, 1–18. https://doi.org/10.1007/s00362-016-0861-9.

Pignatiello, J.J.Jr., Acosta-Mejia, C.A. e Rao, B.V. (1995). The performance of control charts for monitoring process dispersion. Em: *4th Industrial Engineering Research Conference Proceedings* 24-25 May, 1995, 320–328. Nashville, Tennessee, USA.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Weiß, C.H. (2007). Controlling correlated processes of Poisson counts. *Quality and Reliability Engineering International* **23**, 741–754.

Weiß, C.H. e Testik, M.C. (2009). CUSUM monitoring of first-order integer-valued autoregressive processes of Poisson counts. *Journal of Quality Technology* **41**, 389–400.

CP-INGARCH: uma classe geral de modelos para séries de contagem

Filipa Alexandra Cardoso da Silva, *mat0504@mat.uc.pt*

CMUC, Departamento de Matemática, Universidade de Coimbra

1. Introdução

O número diário de passageiros que desembarca no Aeroporto Sá Carneiro, o número de incêndios rurais verificados em Portugal mensalmente, o número de transacções do mercado de acções por minuto ou o número de peças defeituosas por hora produzidas numa fábrica são alguns dos muitos exemplos de séries de contagem que surgem de forma natural em diversas áreas e contextos. Sendo este tipo de sistemas estocásticos facilmente observado, não é de estranhar o interesse nas últimas décadas na procura de modelos que permitam descrevê-los tendo-se multiplicado na literatura as referências acerca deste tema (veja, por exemplo, [12]).
As séries de contagem começaram por ser analisadas como se o seu suporte fosse o conjunto dos números reais mas, adoptar este procedimento na maioria das situações conduzia a resultados sem grande significado o que tornou evidente a necessidade de introduzir modelos específicos que melhor as descrevessem e caracterizassem.
As primeiras modelações, de carácter linear e essencialmente inspiradas nos clássicos modelos ARMA, revelaram-se insuficientes para dar resposta a algumas características empíricas como a heteroscedasticidade condicional. De modo a ter em conta tal tipo de características, surgiram na literatura vários modelos para séries temporais de valores inteiros não negativos inspirados nos clássicos GARCH entre os quais se destacam os GARCH de valor inteiro com distribuição condicional de Poisson (designados de modelos INGARCH), propostos em 2006 por Ferland, Latour e Oraichi [2]. Rapidamente surgiram extensões a estes modelos sendo a lei condicional de Poisson do modelo INGARCH substituída por outras leis discretas de valores inteiros não negativos como a binomial negativa ([11], [13]) ou a Poisson generalizada ([14]).
Com o objectivo de unificar e alargar estes estudos, particularmente no que diz respeito à família de distribuições condicionais, Gonçalves, Mendes Lopes e Silva ([4]) introduziram uma classe de modelos de valores inteiros com evolução para a média condicional análoga à definida nos modelos INGARCH mas em que se considera associada uma família abrangente de leis condicionais, nomeadamente a das leis infinitamente divisíveis discretas com suporte em $\mathbb{N}_0$. Em consequência da equivalência entre leis infinitamente divisíveis e leis de Poisson compostas, no conjunto das leis discretas com suporte $\mathbb{N}_0$, este novo modelo denominou-se modelo GARCH de valor inteiro Poisson Composto.

2. O modelo GARCH de valor inteiro Poisson composto

Seja $X = (X_t, t \in \mathbb{Z})$ um processo estocástico com valores em $\mathbb{N}_0$ e designemos por $\underline{X}_{t-1}$ a σ-álgebra gerada por $\{X_{t-j},\ j \geq 1\}$.

2.1. Definição e casos particulares

Diz-se que X verifica um modelo GARCH de valor inteiro Poisson composto de ordens p e q (onde $p, q \in \mathbb{N}$), que passamos a designar por CP-INGARCH (p, q), se para todo $t \in \mathbb{Z}$, a função característica de X_t condicionada por $\underline{X}_{t-1}$ é da forma

$$\Phi_{X_t|\underline{X}_{t-1}}(u) = exp\left\{i\frac{\lambda_t}{\varphi_t'(0)}[\varphi_t(u)-1]\right\}, \quad u \in \mathbb{R}, \qquad (1)$$

com

$$E\left[X_t|\underline{X}_{t-1}\right] = \lambda_t = \alpha_0 + \sum_{j=1}^{p}\alpha_j X_{t-j} + \sum_{k=1}^{q}\beta_k\lambda_{t-k}, \qquad (2)$$

onde $\alpha_0 > 0, \alpha_j \geq 0$, para $j = 1, \dots, p$, $\beta_k \geq 0$, para $k = 1, \dots, q$, e $(\varphi_t, t \in \mathbb{Z})$ é uma família de funções características sobre $\mathbb{R}$, $\underline{X}_{t-1}$-mensuráveis, associada a uma família de leis discretas de suporte em $\mathbb{N}_0$ e média finita. i designa a unidade imaginária.
Dizemos que X segue um modelo CP-INARCH(p) se $q = 1$ e $\beta_1 = 0$.

Notemos que a derivada de φ_t em $u = 0$, $\varphi_t'(0)$, existe e não se anula já que φ_t é a função característica de uma lei discreta de suporte em $\mathbb{N}_0$ e média finita.
Supondo que as funções $(\varphi_t, t \in \mathbb{Z})$ são deriváveis pelo menos duas vezes, é possível especificar a evolução da variância condicional de X. De facto, deduz-se que

$$V\left[X_t|\underline{X}_{t-1}\right] = -\Phi''_{X_t|\underline{X}_{t-1}}(0) - \lambda_t^2 = -i\frac{\varphi_t''(0)}{\varphi_t'(0)}\lambda_t.$$

A forma funcional da função característica condicional (1) confere uma grande flexibilidade à classe dos modelos INGARCH Poisson compostos. Com efeito, o facto da família de funções características $(\varphi_t, t \in \mathbb{Z})$ ser $\underline{X}_{t-1}$-mensurável faz com que os seus elementos possam ser funções aleatórias ou deterministas, o que permite ao modelo CP-INGARCH incluir os principais processos de valores inteiros condicionalmente heteroscedásticos presentes na literatura como é o caso dos modelos INGARCH Poisson ([2]), INGARCH Poisson generalizado ([14]), INGARCH binomial negativo ([13]) e DINARCH binomial negativo ([11]) (ver [4] para uma explicação mais detalhada desta afirmação). Para além disso, é possível evidenciar ainda a partir desta formulação geral novos processos naturalmente interessantes na prática como os modelos INGARCH Poisson geométrico ([4]) e INGARCH Neyman tipo-A ([3]) já que as leis condicionais que lhes estão associadas são capazes de explicar fenómenos em várias áreas de aplicação. Apresentamos agora com pormenor estes dois novos processos.

a) Modelo NTA-INGARCH (p, q)

O processo X segue um modelo INGARCH Neyman tipo-A se

$$X_t|\underline{X}_{t-1} \sim \boldsymbol{NTA}\left(\frac{\lambda_t}{\phi}, \phi\right),$$

onde $\boldsymbol{NTA}$ representa abreviadamente a lei Neyman tipo-A ([7, Secção 9.6]), $\phi > 0$ e λ_t possui a evolução apresentada em (2). Neste caso, $(\varphi_t, t \in \mathbb{Z})$ em (1) é uma família de funções características deterministas e independentes de t, onde $\varphi_t = \varphi$ representa a função característica da lei de Poisson de parâmetro ϕ, e portanto tem-se

$$\Phi_{X_t|\underline{X}_{t-1}}(u) = exp\left\{\frac{\lambda_t}{\phi}\left[exp\left(\phi\left(e^{iu}-1\right)\right)-1\right]\right\}, \quad u \in \mathbb{R}.$$

Notemos que para este modelo $V\left[X_t|\underline{X}_{t-1}\right] = \lambda_t(1+\phi)$.

b) Modelo GEOMP2-INGARCH (p, q)

Dizemos que o processo X segue um modelo INGARCH Poisson geométrico quando

$$X_t|\underline{X}_{t-1} \sim \boldsymbol{GEOMP}(p^*\lambda_t, p^*),$$

onde $\boldsymbol{GEOMP}$ representa abreviadamente a lei Poisson geométrica ([7, Secção 9.7]), $p^* \in]0,1[$ e λ_t possui a evolução apresentada em (2). Tem-se então

$$\Phi_{X_t|\underline{X}_{t-1}}(u) = exp\left\{p^*\left(\frac{e^{iu}-1}{1-(1-p^*)e^{iu}}\right)\lambda_t\right\}, \quad u \in \mathbb{R},$$

já que $\varphi_t = \varphi$, em (1), é a função característica de uma lei geométrica de parâmetro p^*. Notemos ainda que $V\left[X_t|\underline{X}_{t-1}\right] = \lambda_t\left((2-p^*)/p^*\right)$.
Este modelo é designado de GEOMP2-INGARCH para distingui-lo do modelo apresentado em [4] (Example 1 (1)) onde a função característica φ_t é uma função aleatória já que o parâmetro envolvido na sua lei depende de λ_t e portanto depende das observações passadas do processo.

Para uma lista mais extensa de casos particulares do modelo CP-INGARCH ver [10].
Para além de ter a capacidade de descrever diferentes comportamentos distribucionais e, consequentemente, diferentes tipos de heteroscedasticidade condicional, o modelo INGARCH Poisson composto consegue incorporar outra característica empírica muito associada a séries de contagem, nomeadamente a sobredispersão. Com efeito, o facto da distribuição condicional ser uma lei de Poisson composta de valores inteiros não negativos (e portanto, à excepção da lei de Poisson, ser uma distribuição sobredispersa ([9])) faz com que

$$\frac{V[X_t]}{E[X_t]} \geq \frac{E\left[V\left[X_t|\underline{X}_{t-1}\right]\right]}{E[X_t]} > \frac{E\left[E\left[X_t|\underline{X}_{t-1}\right]\right]}{E[X_t]} = 1.$$

Notemos que mesmo quando a distribuição condicional é a lei de Poisson continuamos a ter uma lei marginal sobredispersa já que $V[X_t] = E[\lambda_t] + V[\lambda_t] > E[X_t]$, na presença de heteroscedasticidade condicional.

2.2. Estrutura probabilista

O estudo da estacionaridade e ergodicidade desta classe de modelos está fortemente ligado à construção de uma sucessão de processos estacionários em média.
Considere-se a função característica φ_t relativa a uma lei discreta, o correspondente modelo CP-INGARCH (p,q) tal que $\beta_1 + \cdots + \beta_q < 1$ e seja $\{\psi_j\}_{j\in\mathbb{N}_0}$ a sucessão de coeficientes associada à representação CP-INARCH(∞) do modelo, i.e., $\psi_0 = \alpha_0/\left(1-\beta_1-\cdots-\beta_q\right)$ e

$$\psi_j = \begin{cases} \alpha_1, & se\ j = 1, \\ \alpha_j + \sum_{k=1}^{j-1} \beta_k \psi_{j-k}, & se\ 2 \leq j \leq p, \\ \sum_{k=1}^{q} \beta_k \psi_{j-k}, & se\ j \geq p+1. \end{cases}$$

Seja $(U_t, t \in \mathbb{Z})$ uma sucessão de variáveis aleatórias inteiras não negativas independentes seguindo a distribuição de Poisson composta de função característica

$$\Phi_{U_t}(u) = exp\left\{\psi_0 \frac{i}{\varphi_t'(0)}[\varphi_t(u)-1]\right\}, \quad u \in \mathbb{R}.$$

Para cada $t \in \mathbb{Z}$ e $k \in \mathbb{N}$, seja $\mathcal{Z}_{t,k} = \{Z_{t,k,j}\}_{j\in\mathbb{N}}$ a sucessão de variáveis aleatórias inteiras não negativas independentes possuindo a lei de Poisson composta de função característica

$$\Phi_{Z_{t,k,j}}(u) = exp\left\{\psi_k \frac{i}{\varphi_t'(0)}[\varphi_t(u) - 1]\right\}, \quad u \in \mathbb{R}.$$

Notemos que $E[U_t] = \psi_0$, $E[Z_{t,k,j}] = \psi_k$ e que $Z_{t,k,j}$ são identicamente distribuídas para cada par $(t,k) \in \mathbb{Z} \times \mathbb{N}$. Assuma-se ainda que todas as variáveis U_s, $Z_{t,k,j}$ $(s,t \in \mathbb{Z}, k,j \in \mathbb{N})$, são independentes entre si. Baseada nestas variáveis aleatórias define-se a sucessão $\left\{\left(X_t^{(n)}, t \in \mathbb{Z}\right), n \in \mathbb{Z}\right\}$ da seguinte forma:

$$X_t^{(n)} = \begin{cases} 0, & n < 0, \\ U_t, & n = 0, \\ U_t + \sum_{k=1}^{n} \sum_{j=1}^{X_{t-k}^{(n-k)}} Z_{t-k,k,j}, & n > 0 \end{cases}$$

onde se convenciona que $\sum_{j=1}^{0} Z_{t-k,k,j} = 0$.
As boas propriedades desta sucessão permitem-nos, entre outros resultados, estabelecer a existência de uma solução fortemente estacionária e ergódica, nomeadamente o seu limite quase certo, na subclasse de processos CP-INGARCH para os quais a função característica φ_t é determinista e independente de t. O teorema seguinte verifica-se então, em particular, para os modelos INGARCH Poisson, GP-INGARCH, NB-DINARCH, NTA-INGARCH e GEOMP2-INGARCH referidos anteriormente. Deste resultado destaca-se a simplicidade da condição envolvida e o facto desta se exprimir apenas em termos dos coeficientes do modelo e independentemente da sua distribuição condicional.

Teorema. Seja X um processo seguindo o modelo CP-INGARCH (p,q) com φ_t determinista independente de t. Se $\sum_{j=1}^{p} \alpha_j + \sum_{k=1}^{q} \beta_k < 1$, existe um processo fortemente estacionário e ergódico, solução do modelo. Mais, os dois primeiros momentos deste processo são finitos.

A prova deste resultado pode ser consultada em [4].

3. Estimação do modelo

No que se segue, vamos apresentar alguns métodos de estimação para um qualquer processo CP-INARCH (1) estacionário em média onde as funções características φ_t são deterministas e independentes de t, $\varphi_t = \varphi$. Com esse objectivo, considere-se $\boldsymbol{x} = (x_1, \ldots, x_n)$ a série temporal observada que pretendemos modelar com um modelo CP-INARCH (1) de parâmetros $\theta = (\alpha_0, \alpha_1, v_0)$, onde $v_0 = -i\,\varphi''(0)/\varphi'(0)$ e que inclui o parâmetro adicional associado à distribuição condicional do modelo (por exemplo, $v_0 = 1 + \phi$ no modelo NTA-INARCH (1) e $v_0 = (2 - p^*)/p^*$ no modelo GEOMP2-INARCH (1)).

3.1. Método em duas etapas baseado nos mínimos quadrados

O primeiro método proposto é um método em duas etapas que consiste em, numa primeira etapa, determinar pelo método dos mínimos quadrados condicional (CLS) estimativas para o parâmetro $\alpha = (\alpha_0, \alpha_1)$ e numa segunda etapa, fixando estas estimativas, usar o método dos momentos para estimar o parâmetro v_0.
Minimizando a função

$$Q_n(\alpha) = \sum_{t=2}^{n} \left(x_t - E\left[X_t | \underline{X}_{t-1} = x_{t-1}\right]\right)^2 = \sum_{t=2}^{n} (x_t - \alpha_0 - \alpha_1 x_{t-1})^2,$$

obtemos o estimador CLS de α, digamos $\hat{\alpha}_n = \left(\hat{\alpha}_{0,n}, \hat{\alpha}_{1,n}\right)$, onde

$$\hat{\alpha}_{1,n} = \frac{\sum_{t=2}^{n} X_t X_{t-1} - \frac{1}{n-1} \cdot \sum_{t=2}^{n} X_t \cdot \sum_{s=2}^{n} X_{s-1}}{\sum_{t=2}^{n} X_{t-1}^2 - \frac{1}{n-1} \cdot (\sum_{t=2}^{n} X_{t-1})^2},$$

$$\hat{\alpha}_{0,n} = \frac{\sum_{t=2}^{n} X_t - \hat{\alpha}_{1,n} \cdot \sum_{t=2}^{n} X_{s-1}}{n-1}. \qquad (3)$$

A consistência e a distribuição assimptótica deste estimador são estabelecidas no próximo teorema e decorrem da aplicação dos resultados de Klimko e Nelson [8, Secção 3].

Teorema. Seja $\hat{\alpha}_n$ o estimador CLS de α dado em (3). Tem-se

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{d} N(\mathbf{0}_{2\times 1}, \boldsymbol{V}^{-1}\boldsymbol{W}\boldsymbol{V}^{-1}),$$

quando $n \to \infty$, onde as entradas da matriz $\boldsymbol{V}^{-1}\boldsymbol{W}\boldsymbol{V}^{-1} = (b_{ij}), i, j = \{1,2\}$, são dadas por

$$b_{11} = \frac{\alpha_0}{1-\alpha_1}\left(\alpha_0(1+\alpha_1) + \frac{v_0 + (d_0 - v_0^2)\alpha_1(1+\alpha_1-\alpha_1^2) + (3v_0^2 - d_0)\alpha_1^4}{v_0(1+\alpha_1+\alpha_1^2)}\right),$$

$$b_{12} = b_{21} = v_0\alpha_1 - \alpha_0(1+\alpha_1) - \frac{\alpha_1(1+\alpha_1)(d_0 + (3v_0^2 - d_0)\alpha_1^2)}{v_0(1+\alpha_1+\alpha_1^2)},$$

$$b_{22} = (1-\alpha_1^2)\left(1 + \frac{\alpha_1(d_0 + (3v_0^2 - d_0)\alpha_1^2)}{v_0\alpha_0(1+\alpha_1+\alpha_1^2)}\right),$$

onde $\xrightarrow{d}$ significa convergência em lei, $v_0 = -i\,\varphi''(0)/\varphi'(0)$ e $d_0 = -\varphi'''(0)/\varphi'(0)$.

Para estimar o parâmetro v_0, começamos por recordar a expressão geral do momento de segunda ordem de um processo CP-INARCH (1) estabelecida em [5, Exemplo 3.1]:

$$E[X_t^2] = \frac{\alpha_0(v_0 + \alpha_0(1+\alpha_1))}{(1-\alpha_1)(1-\alpha_1^2)}.$$

Um estimador para v_0, digamos $\hat{v}_0$, obtém-se então resolvendo a equação

$$\frac{\hat{\alpha}_{0,n}\left(\hat{v}_0 + \hat{\alpha}_{0,n}(1+\hat{\alpha}_{1,n})\right)}{(1-\hat{\alpha}_{1,n})(1-\hat{\alpha}_{1,n}^2)} = \frac{1}{n}\sum_{t=1}^{n} X_t^2.$$

Em particular, deduzimos o seguinte estimador de ϕ para o modelo NTA-INARCH (1)

$$\hat{\phi}_n = -1 - \hat{\alpha}_{0,n}(1+\hat{\alpha}_{1,n}) + \frac{(1-\hat{\alpha}_{1,n})(1-\hat{\alpha}_{1,n}^2)}{n\hat{\alpha}_{0,n}}\sum_{t=1}^{n} X_t^2,$$

e para o modelo GEOMP2-INARCH (1), o estimador de p^*

$$\widehat{p^*}_n = 2\left[-1 - \hat{\alpha}_{0,n}(1+\hat{\alpha}_{1,n}) + \frac{(1-\hat{\alpha}_{1,n})(1-\hat{\alpha}_{1,n}^2)}{n\hat{\alpha}_{0,n}}\sum_{t=1}^{n} X_t^2\right]^{-1}.$$

A consistência forte de $\hat{v}_0$ resulta da aplicação do teorema ergódico, tendo em conta a estacionaridade forte e ergodicidade de X, e das propriedades da convergência quase certa.

3.2. Método em duas etapas baseado na quase máxima verosimilhança de Poisson

O segundo método proposto é, tal como o anterior, um método em duas etapas que consiste em determinar pelo método da quase máxima verosimilhança condicional de Poisson estimativas para α_0 e α_1 e depois usar o método dos momentos para estimar o parâmetro v_0.
Os estimadores de α são então obtidos maximizando a função de pseudo-verosimilhança

$$\check{L}_n(\theta|\boldsymbol{x}) = \sum_{t=2}^{n} (x_t \log(\lambda_t) - \lambda_t - \log(x_t!)).$$

Através das condições de regularidade estabelecidas por Ahmad e Francq ([1]) conclui-se que este estimador é consistente e assimptoticamente Gaussiano sempre que $\alpha_1 < 1$.

3.3. Método da máxima verosimilhança condicional

Um dos métodos mais usuais e que conduz a estimadores com boas propriedades assimptóticas é o da máxima verosimilhança. Ao contrário do que acontece com os métodos de estimação anteriores, neste é essencial o conhecimento da lei envolvida. Sendo a lei condicional a única de que se dispõe, recentemente, Gonçalves, Mendes Lopes e Silva ([6]) desenvolveram o procedimento conducente à obtenção dos estimadores de máxima verosimilhança de θ para os modelos NTA-INARCH (1) e GEOMP2-INARCH (1) usando o método da máxima verosimilhança condicional (CML).
Na literatura é possível também encontrar estudos nesse sentido para os outros modelos listados neste texto (veja-se, por exemplo, [2] para o modelo INGARCH, [14] para o INGARCH Poisson generalizado e [11] para o DINARCH binomial negativo).

Uma análise comparativa entre os três métodos indicados acima foi também apresentada em [6] e permitiu concluir que ambos conduzem a bons resultados sendo que o método CML oferece erros quadráticos médios ligeiramente melhores.

4. Aplicação a uma série de contagem

Com o objectivo de ilustrar a metodologia apresentada nas secções anteriores, vamos considerar a série temporal representada na Figura 1 **(a)** relativa ao número mensal de dias de chuva (dia em que a quantidade de precipitação é superior ou igual a 0,2 mm) observados entre Janeiro de 1982 e Março de 2018 registados pela Estação Meteorológica de Changi (https://data.gov.sg/dataset).
A média e variância empíricas desta série são 14.0460 e 24.0071, respectivamente, indicando que a sua distribuição marginal é sobredispersa. Observemos também que esta série temporal exibe características de heteroscedasticidade condicional e que a função de autocorrelação apresentada na Figura 1 **(c)** sugere uma dependência de ordem 1. Com base nestas características empíricas, tem sentido considerar na sua modelação um processo CP-INARCH (1).
Na tentativa de obter um modelo adequado para esta série de contagem, apresenta-se de seguida um estudo comparativo entre processos do tipo INARCH (1) considerando como distribuições condicionais as leis de Poisson, Poisson generalizada, Neyman tipo-A, Poisson geométrica e binomial negativa e utilizando o método da máxima verosimilhança condicional.
A maximização da função de log-verosimilhança foi implementada usando a função *fmincon* (disponível no software MATLAB) considerando como restrições $\alpha_0 > 0$ e a condição de estacionaridade forte $0 < \alpha_1 < 1$. Os valores iniciais requeridos foram baseados no método de duas etapas apresentado na secção 3.1. Para aferir a adequação dos modelos estimados utilizaram-se o valor da função de log-verosimilhança, o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC).

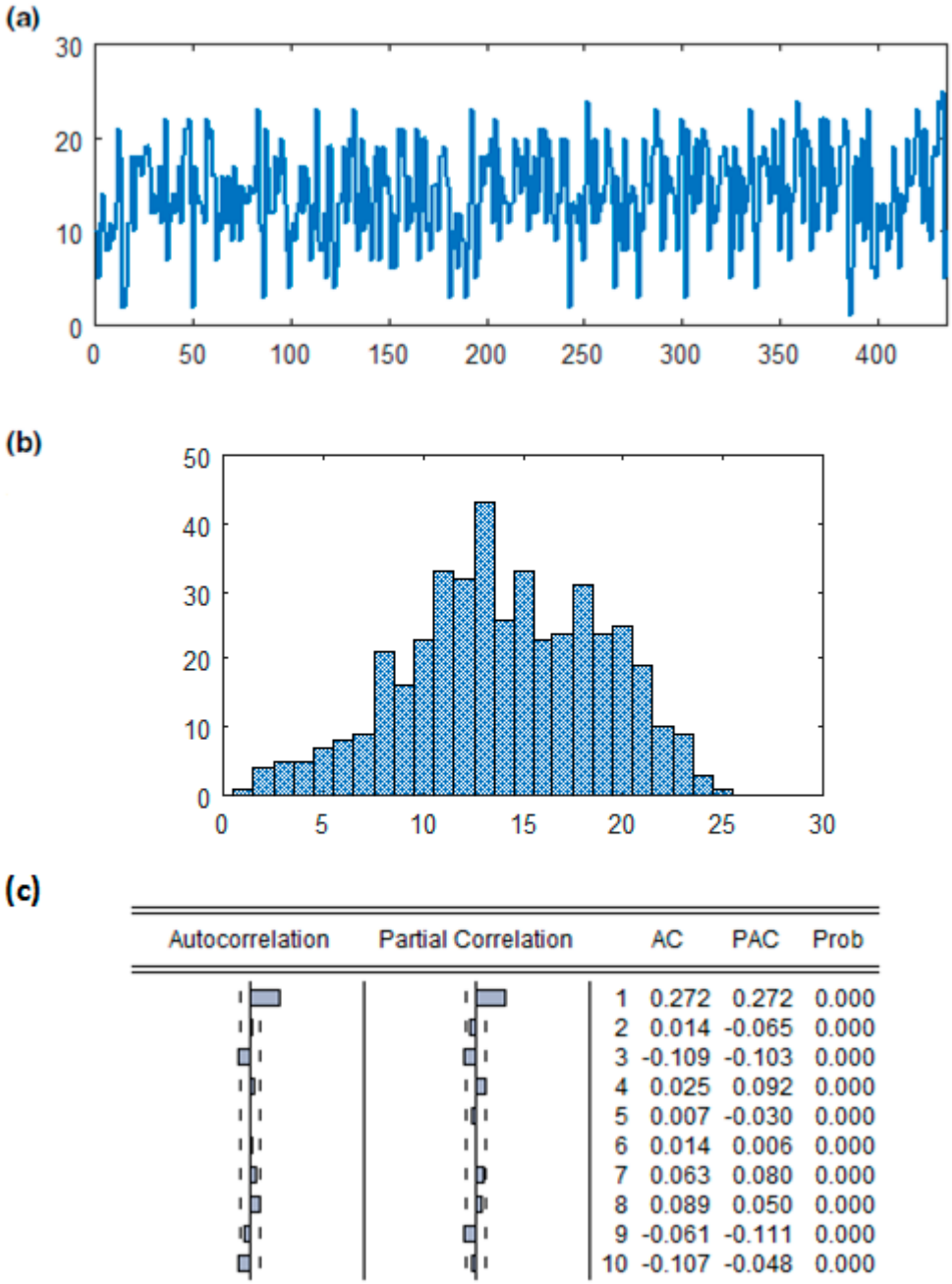


Figura 1: Número mensal de dias de chuva observados entre Janeiro de 1982 e Março de 2018 **(a)**, correspondente histograma **(b)** e autocorrelações e autocorrelações parciais empíricas **(c)**.

Modelo	$\hat{\alpha}_{0,435}$	$\hat{\alpha}_{1,435}$	Parâmetro adicional	-Log L	AIC	BIC
Poisson INARCH (1)	10.1254 **(0.0401)**	0.2796 **(0.0012)**		1336.2	2676.4	2684.5
GP-INARCH (1)	10.0050 **(0.0075)**	0.2882 **(0.0007)**	$\hat{\kappa}_{435}$=0.1214 **(0.0013)**	1305.8	2617.6	2629.9
NTA-INARCH (1)	10.0564 **(0.0401)**	0.2845 **(0.0129)**	$\hat{\phi}_{435}$=0.7435 **(0.0061)**	1301.8	2609.7	2621.9
GEOMP2-INARCH (1)	10.0339 **(0.0231)**	0.2861 **(0.0017)**	$\widehat{p*}_{435}$=0.7350 **(0.0023)**	1303.4	2612.9	2625.1
NB-DINARCH (1)	10.0150 **(0.5419)**	0.2875 **(0.0353)**	$\hat{\beta}_{435}$=1.6960 **(0.0433)**	1304.9	2615.9	2628.1

Tabela 1: Estimativa da máxima verosimilhança condicional para os parâmetros de vários modelos CP-INARCH (1) e respectivos erros entre parêntesis.

Da Tabela 1 é possível concluir que o modelo NTA-INARCH (1) é o que melhor se ajusta aos dados apresentados na Figura 1 **(a)** já que é o que exibe os menores valores de -Log L, AIC e BIC. Os modelos GEOMP2-INARCH (1) e NB-DINARCH (1) estão bastante próximos e o Poisson INARCH (1) é o que apresenta o pior ajustamento. A média, variância e coeficiente de autocorrelação de primeira ordem (FOAC) para os modelos CP- INARCH (1) ajustados estão sintetizados na Tabela 2. Os resultados estão de acordo com as conclusões anteriores.

		Modelo				
	Amostra	Poisson	GP	NTA	GEOMP2	NB disperso
Média	14.0460	14.0552	14.0559	14.0551	14.0550	14.0561
Variância	24.0071	15.2472	19.8558	26.6632	26.3466	25.9871
FOAC	0.272	0.2796	0.2882	0.2845	0.2861	0.2875

Tabela 2: Média, variância e FOAC amostrais e estimados através do método CML para vários modelos CP-INARCH (1).

Agradecimentos Este trabalho foi financiado parcialmente pelo Centro de Matemática da Universidade de Coimbra -- UID/MAT/00324/2019, financiado pelo Governo Português através da FCT / MEC e co-financiado pelo Fundo Europeu de Desenvolvimento Regional através do acordo de parceria PT2020.

Referências

[1] Ahmad, A., Francq, C. (2016). Poisson QMLE of count time series models. *Journal of Time Series Analysis* **37**, 291–314.

[2] Ferland, R., Latour, A., Oraichi, D. (2006). Integer-valued GARCH process. *Journal of Time Series Analysis* **27**, 923–942.

[3] Gonçalves, E., Mendes Lopes, N., Silva, F. (2015). A new approach to integer-valued time series modeling: The Neyman Type-A INGARCH model. *Lithuanian Mathematical Journal*, **55(2)**, 231–242.

[4] Gonçalves, E., Mendes Lopes, N., Silva, F. (2015). Infinitely divisible distributions in integer valued GARCH models. *Journal of Time Series Analysis* **36**, 503–527.

[5] Gonçalves, E., Mendes Lopes, N., Silva, F. (2016). Zero-inflated compound Poisson distributions in integer-valued GARCH models. *Statistics* **50**, 558–578.

[6] Gonçalves, E., Mendes Lopes, N., Silva, F. (2017). Two-step estimation procedures for compound Poisson INARCH processes. *Pré-Publicações do Departamento de Matemática* **17–34**.

[7] Johnson, N. L., Kotz, S., Kemp, A. W. (2005). Univariate discrete distributions. Wiley, New York, 3rd edition.

[8] Klimko, L.A., Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes. *The Annals of Statistics* **6(3)**, 629–642.

[9] Minkova L. D., Balakrishnan, N. (2013). Compound weighted Poisson distributions. *Metrika*,**76**, 543–558.

[10] Silva, F. (2016). Compound Poisson Integer-Valued GARCH Processes. Tese de Doutoramento do Programa Inter-Universitário de Doutoramento em Matemática, Departamento de Matemática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra, https://estudogeral.sib.uc.pt/handle/10316/83579.

[11] Xu, H.-Y., Xie, M., Goh, T. N., and Fu, X. (2012). A model for integer-valued time series with conditional overdispersion. *Computational Statistics and Data Analysis* **56**, 4229–4242.

[12] Weiss, C. H. (2018). An Introduction to Discrete-Valued Time Series. Wiley.

[13] Zhu, Fk. (2011). A negative binomial integer-valued GARCH model. *Journal of Time Series Analysis* **32**, 54–67.

[14] Zhu, Fk. (2012). Modelling overdispersed or underdispersed count data with generalized Poisson integer-valued GARCH models. *Journal of Mathematical Analysis and Applications* **389 1**, 58–71.

• Teses de Doutoramento

Título: Análise de distribuições de distâncias entre palavras genómicas
Autora: Ana Helena Marques de Pinho Tavares, *ahtavares@ua.pt*
Orientadoras: Vera Afreixo e Paula Brito

A minha tese foi dedicada ao desenvolvimento de novos procedimentos estatísticos com vista à extração de informação de sequências genómicas, com especial enfoque na exploração de dissimilaridades, na deteção de observações atípicas e no agrupamento de dados. As sequências genómicas foram previamente transformados em distribuições de frequências, pelo que os procedimentos desenvolvidos têm potencial para serem aplicados noutras áreas científicas, que não a genómica.

Uma cadeia de ADN pode ser analisada como uma sequência num alfabeto de quatro letras, onde cada letra representa uma unidade genómica – o nucleótido. Uma palavra genómica é uma sequência definida nesse alfabeto. Uma característica com particular interesse no estudo de sequências genómicas é a distribuição das palavras ao longo da sequência, podendo esta ser caracterizada pelas distâncias entre palavras. A contagem das distâncias entre palavras fornece distribuições discretas passíveis de análise estatística – as distribuições de distâncias (DD) entre palavras.

As DD empíricas exibem diferenças estatisticamente significativas em relação à distribuição de referência, ou seja, a distribuição que é obtida se os nucleótidos forem gerados aleatoriamente e de forma independente. Para além de se observarem desvios na tendência global (*baseline*) das distribuições, pode observar-se a ocorrência de "picos" de frequências nas distribuições empíricas. Tais picos revelam que a palavra exibe uma preferência por se repetir a uma determinada distância.

Com vista ao estudo comparativo de sequências genómicas e à definição de assinaturas de espécies, foram descritos modelos teóricos que descrevam DD entre palavras em cenários aleatórios. O agrupamento hierárquico dos perfis obtidos (por confronto entre a DD real e a teórica) colocou em evidência a sua capacidade de discriminação entre espécies, mimetizando relações evolutivas entre elas.

Um dos principais tópicos de investigação consistiu na deteção de distribuições com comportamentos anormais, designadas por distribuições atípicas, tendo sido exploradas diversas abordagens. Por exemplo, a avaliação da concordância entre uma DD e a correspondente DD esperada num cenário teórico, recorrendo a medidas de efeito; ou a quantificação da dissemelhança entre as DD de duas palavras que são complemento invertido.

Devido à forte presença de picos nas DD, foi proposta uma medida de dissimilaridade entre distribuições. Esta combina diferenças entre as magnitudes e diferenças entre as posições dos 'n' maiores picos de frequência das duas distribuições. Conjetura-se que palavras que apresentam distribuições de distâncias similares poderão estar associadas a uma função biológica ou estrutural semelhante, e que palavras com distribuições de distâncias atípicas poderão estar relacionadas com alguma função biológica (motivos).

Assumindo que as DD são resultado da soma de duas componentes, uma que define a tendência de base (*baseline*) e outra que define os seus picos, foi proposto um procedimento de agrupamento inovador que conjuga elementos das duas componentes. O procedimento é validado através de um estudo de simulação e é aplicado a dados do genoma humano.

Em suma, esta tese contribui com a proposta de uma nova medida de dissimilaridade entre distribuições, que se baseia nas diferenças entre os seu picos de frequências, assim como com a proposta de um procedimento de agrupamento de distribuições, que tem em conta quer a *baseline* quer a estrutura de picos das distribuições. Do ponto de vista da aplicação prática, são ainda delineados novos procedimentos estatísticos para a identificação de características gerais e específicas das sequências genómicas.

Ana Tavares

Título: Contributos Computacionais e Metodológicos na Estimação de Valores Extremos
Autora: Helena Penalva, *helena.penalva@esce.ips.pt*
Orientadoras: M. Manuela Neves e Sandra Nunes

Na minha tese foi abordada a estimação, essencialmente em contexto semi-paramétrico, do parâmetro primordial em Teoria de Valores Extremos, o índice de valores extremos (EVI), ξ, para modelos de caudas pesadas, i.e, modelos para os quais $\xi > 0$, apresentando alguns contributos computacionais e metodológicos. Foi realizado um estudo sobre uma classe de estimadores baseada na média de Lehmer de ordem p de números positivos, que generaliza a média aritmética ($p = 1$) e a média harmónica ($p = 0$). Essa classe de estimadores de Lehmer de ordem p, $\mathbf{L}p$, constitui uma generalização do clássico estimador de Hill. Dada uma amostra aleatória $(X_1, \dots, X_n)$ e a correspondente amostra das estatísticas ordinais ascendentes $(X_{1:n}, \dots, X_{n:n})$, o estimador de Hill pode ser considerado como a média de Lehmer de ordem 1 dos k log-excessos $V_{ik} := \ln X_{n-i+1:n} - \ln X_{n-k:n}, 1 \le i \le k < n$. Também foi estudada outra classe de estimadores, a classe de estimadores de Lehmer de ordem p de viés reduzido de segunda ordem, L_p^{RB}. O estudo consistiu em obter o comportamento assintótico, em distribuição, dessas duas classes de estimadores, como também os seus comportamentos em amostras de dimensão finita, tendo-se efetuado algumas comparações entre elas e com outros estimadores do EVI, que têm constituído investigação recente. Os vários estimadores abordados neste trabalho foram aplicados a dois conjuntos de dados reais. De referir ainda que na obtenção dos resultados foram desenvolvidas funções no *software* R.

Admitindo que a função de distribuição, F, subjacente à amostra aleatória, pertence ao domínio de atração da distribuição de valores extremos, com $\xi > 0$, através de uma condição limite de primeira ordem relativa à cauda da distribuição F, e admitindo que $k = k_n$ é uma sucessão intermédia de valores inteiros, prova-se que a classe de estimadores de Lehmer de ordem p é consistente para a estimação de $\xi > 0$, com p>0. Adicionando ainda uma condição limite de segunda ordem, que regula a convergência da condição de primeira ordem, dependendo de um parâmetro $\rho \le 0$, e de uma função $A(t)$ **convergente para 0 quando** $t \to \infty$, e supondo $k = k_n$ uma sucessão intermédia de valores inteiros tais que $\lambda_A := \lim_{n\to\infty} \sqrt{k} A(n/k)$ é um valor finito, também se garante a convergência em distribuição desses estimadores para a distribuição normal se $p > 1/2$.

Além disso, se se trabalhar com a classe de modelos Hall-Welsh, para a qual a condição limite de segunda ordem se verifica com $A(t) = \xi \beta t^{\rho}, \rho < 0$, prova-se que se β e ρ forem estimados consistentemente por $\hat{\beta}$ e $\hat{\rho}$ com $\hat{\rho} - \rho = o_p(1/\ln n)$, consegue-se eliminar a componente dominante do viés de L_p^{RB}, e manter a sua variância assintótica igual à do correspondente estimador do EVI de viés não reduzido.

No estudo comparativo do comportamento assintótico realizado, baseado numa medida de eficiência, genericamente denominada por AREFF, que compara a raíz dos erros quadrados médios assintóticos dos estimadores, no nível ótimo, k_0, i.e., o nível que minimiza os correspondentes erros quadráticos médios assintóticos, concluiu-se que nenhum dos estimadores do EVI considerados domina as alternativas, mas a classe de estimadores $\mathbf{L}p$ revelou um bom desempenho.

No estudo comparativo do comportamento, dos estimadores considerados, em amostras de dimensão finita, baseado na simulação Monte Carlo multi-amostra, referente a alguns modelos pertencentes à classe de Hall-Welsh, concluiu-se que, os valores de p que parecem melhorar o desempenho dos estimadores $\mathbf{L}p$ e $\mathbf{L}p^{\mathbf{RB}}$ dependem fortemente do valor de ρ, constatando-se que, esses valores são, para $\rho > 1$, consideravelmente maiores do que aqueles que se verificam para $\rho \le -1$.

Neste trabalho foram também apresentados dois casos de estudo de aplicação dos métodos descritos. O primeiro conjunto de dados consistiu nas descargas médias diárias medidas na estação hidrométrica de Fragas da Torre no rio Paiva, e o segundo conjunto de dados consistiu nos valores diários da área ardida, em Portugal, relativos a cada um dos incêndios existentes nesse dia.

Helena Penalva

Título: Contribuições para o Desenho de Modelos de Previsão da Procura: Aplicação no Planeamento Energético para a Cidade de Cabinda

Autor: António Casimiro Puindi, *acpuindi@yahoo.com.br*
Orientadora: Maria Eduarda Silva

Na minha tese foi desenvolvido um quadro de modelos estruturais dinâmicos com a integração dos efeitos das covariáveis. O principal objetivo foi o de fornecer contribuições inerentes à construção de modelos de previsão para séries temporais com padrões sazonais complexos. A esse respeito, foi proposto um modelo estrutural básico com covariáveis, SCov, para previsão de séries temporais de sazonalidade não complexa. No contexto da previsão de séries temporais com padrões sazonais complexos, foi proposto um modelo estrutural trigonométrico com efeitos das covariáveis, TSCov, onde a componente sazonal é modelada mediante séries de Fourier e o seu ruído é projetado no intuito de: (i) ser a fonte de aleatoriedade para a componente sazonal em si e; (ii) propagar o efeito dessa aleatoriedade nos coeficientes dos termos trigonométricos estocasticamente variantes ao longo do tempo. A formulação dos dois modelos (pode admitir a transformação Box-Cox) integra as três principais componentes não observáveis: o nível, a tendência e a sazonalidade.

Do ponto de vista da extração do sinal de uma série temporal, um filtro de Kalman com as matrizes de covariância do sistema calculadas recursivamente é construído. O cálculo recursivo dessas matrizes é baseado nas inovações, a priori e a posteriori, do modelo. As inovações são incorporadas nas estruturas das matrizes de covariâncias de modo a influenciarem o ajusto das mesmas num processo recursivo até melhorar a precisão da estimativa do estado.

No domínio da estimação, um procedimento computacional de estimação é construído. É um procedimento único, recursivo e sistemático, baseado na estimativa de máxima verossimilhança e congrega no mesmo processo o filtro de Kalman e o método de regressão múltipla para a seleção do número de harmônicas para os termos trigonométricos na componente sazonal. O filtro de Kalman construído permite calcular não só as previsões pontuais e os intervalos de previsão, como também permite calcular os erros padrão de cada estimativa de parâmetro. A previsão das covariáveis é baseada na abordagem de média móvel exponencialmente ponderada. Ainda no âmbito da previsão, um procedimento bootstrap, Boot.TSCov, não paramétrico para previsão de séries temporais com padrões sazonais complexos é proposto.

Os resultados do estudo empírico demonstram o potencial do quadro de modelos propostos incluindo o processo de estimação como uma metodologia promissora para a previsão de séries temporais com padrões sazonais complexos, especialmente quando o objetivo é incluir os efeitos das covariáveis na previsão. O procedimento bootstrap formulado fornece bons resultados para o conjunto de dados usados. Os intervalos de previsão obtidos pelo modelo TSCov são menos precisos quando comparados com os obtidos pelo modelo Boot.TSCov. Com esses resultados obtidos, uma pergunta pode ser feita: qual dos modelos, TSCov e TBATS, usar? A resposta é imediata, como TBATS é
um modelo automático no sentido que a integração das covariáveis é improvável, a abordagem TSCov é preferível se houver covariáveis que são preditores úteis, pois podem ser adicionados como regressores e melhorar as previsões.

António Casimiro Puindi

• Livros

Título: *Time Series Clustering and Classification*
Autores: Elizabeth Ann Maharaj, Pierpaolo D'Urso, Jorge Caiado
Ano: 2019. Editora: Chapman & Hall.
ISBN: 9781498773218.

Título: *Computational Bayesian Statistics: An Introduction*
Autores: Amaral Turkman, M. A., Paulino, C. D. e Müller, P.
Ano: 2019. Editora: Cambridge University Press.
doi: 10.1017/9781108646185.

• Capítulos de Livros

Caeiro, F., Cabral I. and Gomes, M.I. (2018). Improving asymptotically unbiased extreme value index estimation. In T.A. Oliveira, C. Kitsos, A. Oliveira and L.M. Grilo (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Springer, 155-163.
https://www.springer.com/gp/book/9783319766041; https://doi.org/10.1007/978-3-319-76605-8_11

Costa, C., Pereira, I. and Scotto, M.G. (2018) Surveillance in Discrete Time Series. In *Recent Studies on Risk Analysis and Statistical Modeling*, (Oliveira, T.A., Kitsos, C.P., Oliveira, A., Grilo, L., Eds), pp 197-212, Springer Series: Contributions to Statistics, Springer.

Figueiredo, F., Figueiredo A. and Gomes, M.I. (2018). Acceptance-sampling plans for reducing the risk associated with chemical compounds. In T.A. Oliveira, C. Kitsos, A. Oliveira and L.M. Grilo (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Springer, 99-111.
https://doi.org/10.1007/978-3-319-76605-8_7

Pereira, I. and Silva, N. (2018) Statistical Modelling of Counts with a Simple Integer-valued Bilinear Process. In *Recent Studies on Risk Analysis and Statistical Modeling*, (Oliveira, T.A., Kitsos, C.P., Oliveira, A., Grilo, L., Eds), pp 345-357, Springer Series: Contributions to Statistics, Springer, ISBN 978-3-319-76605-8

Retrospetiva do Boletim SPE

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

- Outono de 2018 - Destaque:Equações diferenciais e algumas aplicações
- Primavera de 2018 - Destaque: Estatística Multivariada – perspetiva no século XXI
- Outono de 2017 - Destaque: O Tema Central da Estatística - um novo olhar
- Primavera de 2017 - Destaque: Incerteza em Engenharia
- Outono de 2016 - Destaque: O Tema Central da Estatística
- Primavera de 2016 - Destaque: Séries Temporais e suas aplicações
- Outono de 2015 - Destaque: Estatística em Genética
- Primavera de 2015 – Destaque: Estatística no Desporto
- Outono de 2014 – Destaque: Estatística no Ensino Básico e Secundário
- Primavera de 2014 – Destaque: (Um) Ano Internacional da Estatística
- Outono de 2013 – Destaque: A "Escola Bayesiana" em Portugal
- Primavera de 2013 – Destaque: Estatística não-paramétrica
- Outono de 2012 – Destaque: Métodos Estatísticos em Medicina
- Primavera de 2012 – Destaque: Estatística no Ensino Superior Politécnico
- Outono de 2011 – Destaque: Análise de Sobrevivência
- Primavera de 2011 – Destaque: Sondagens e Censos
- Outono de 2010 – Destaque: Estatística Espacial
- Primavera de 2010 – Destaque: Data Mining - Prospecção (Estatística) de Dados
- Outono de 2009 – Destaque: Modelos Econométricos
- Primavera de 2009 – Destaque: Investigação (em) Estatística
- Outono de 2008 – Destaque: Processos Estocásticos
- Primavera de 2008 – Destaque: ALEA - Um sítio do nosso mundo
- Outono de 2007 – Destaque: Bioestatística
- Primavera de 2007 – Destaque: A "Escola de Extremos" em Portugal
- Outono de 2006 – Destaque: Ensino e Aprendizagem da Estatística

também disponíveis em *http://www.spestatistica.pt/index.php/publicacoes-57/boletins*

Edições SPE - Mini Cursos

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 Giovani 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

PRÉMIOS "ESTATÍSTICO JÚNIOR 2019"
REGULAMENTO

Está aberto **até 25 de Maio de 2019** o concurso para atribuição dos prémios **"Estatístico Júnior 2019"**, de acordo com o seguinte regulamento:

1. A atribuição dos prémios **"Estatístico Júnior 2019"** é promovida pela Sociedade Portuguesa de Estatística (SPE), com o apoio do Centro de Matemática da Universidade de Coimbra (CMUC), e tem como objetivo estimular e desenvolver o interesse dos alunos dos Ensinos Básico e Secundário pelas áreas de Probabilidades e de Estatística.

2. Podem candidatar-se aos prémios **"Estatístico Júnior 2019"** os alunos inscritos, no ano lectivo 2018/19, no 3.° 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 Probabilidades ou Estatística.

5. O **trabalho** deverá ser constituído por um texto escrito em Português com um máximo de 10 páginas A4 dactilografadas e um *poster* formato A2 que resuma os principais aspetos do trabalho.

6. Poderão ser atribuídos prémios **"Estatístico Júnior 2019**" a sete trabalhos: aos três primeiros classificados de entre os trabalhos candidatos do 3º 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: 3 livros para cada aluno com trabalho classificado em primeiro lugar e 2 livros para cada aluno com trabalho classificado em segundo ou terceiro lugares.

7. Ao professor orientador do trabalho classificado em 1º lugar, em cada grau de ensino, é atribuída uma anuidade grátis como sócio da SPE e um livro de divulgação científica.

8. Aos grupos com trabalhos classificados em 1º lugar será também oferecida uma ampliação do correspondente *poster* que será exposta na Sessão de Entrega do Prémio.

9. A candidatura é composta pelo **Boletim de Candidatura**, devidamente preenchido, e pelo **trabalho (**poster e texto**)**. 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 para efeitos da avaliação** para **Sociedade Portuguesa de Estatística – Bloco C6, Piso 4 – Campo Grande – 1749-016 Lisboa.**

10. 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.

11. 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.

12. A atribuição dos prémios **"Estatístico Júnior 2019"** será anunciada logo que conhecida a decisão do júri e a sua entrega formal será realizada numa sessão expressamente dedicada a essa entrega.

13. Os prémios **"Estatístico Júnior 2019"** poderão não ser atribuídos.

14. O Boletim de candidatura e este regulamento podem ser obtidos em www.spestatistica.pt

Apoio

Centre for Mathematics
University of Coimbra

Prémio SPE 2019

Pretendendo estimular a atividade de estudo e investigação científica em Probabilidades e Estatística, a Sociedade Portuguesa de Estatística institui em 2019 o Prémio SPE, regido pelo seguinte regulamento.

Está aberto até 31 de Agosto de 2019 o concurso para atribuição do Prémio SPE 2019, doravante referido como prémio. O prémio é constituído por uma quantia de 1000 euros.

Ao prémio podem concorrer trabalhos originais sobre temas de Probabilidades e Estatística, desde que não tenham sido objeto de qualquer prémio atribuído por outra instituição. O trabalho deverá ser apresentado em português ou em inglês e não poderá exceder 25 páginas A4.

Podem candidatar-se ao prémio sócios da SPE que não completem 35 anos de idade até 31 de Dezembro de 2019 e que não tenham recebido o prémio nas quatro edições anteriores.

A decisão de admissibilidade e a apreciação dos trabalhos submetidos a concurso é da competência de um Júri, cuja constituição é da responsabilidade da Direção da SPE.

Os critérios de seleção pautar-se-ão pela exigência e precisão nos vários aspetos que o Júri considerar pertinentes, nomeadamente: i) qualidade e clareza do texto; ii) inovação e rigor científico; iii) contribuição para o desenvolvimento da área de Probabilidades e Estatística nos planos teórico, metodológico e/ou aplicado.

O Júri é soberano nas suas decisões, não havendo lugar a recurso.

O Júri reserva-se o direito de não atribuir o Prémio SPE 2019.

As candidaturas ao prémio, dirigidas à Presidente da SPE, são constituídas pelos trabalhos concorrentes e pelo *curriculum vitae* dos autores. Podem ser enviadas por correio eletrónico para spe@spestatistica.pt ou, em carta registada, para a morada a seguir indicada. O carimbo do correio valida a data de entrega.

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

A entrega formal do Prémio SPE 2019, com apresentação do trabalho galardoado, terá lugar numa sessão do XXIV Congresso da SPE que decorrerá em Amarante entre 6 e 9 de Novembro de 2019.

SOCIEDADE PORTUGUESA DE ESTATÍSTICA

PRÉMIO SPE 2018

Modelos espaço-temporais para dados georreferenciados do desemprego

Soraia Alexandra Gonçalves Pereira, *sapereira@fc.ul.pt*

Este trabalho foi feito no âmbito do doutoramento na Faculdade de Ciências da Universidade de Lisboa, com a orientação do Prof. Kamil Feridun Turkman e Dr. Luís Fernandes Correia.

Em Portugal, o Instituto Nacional de Estatística (INE) publica trimestralmente as estimativas oficiais do mercado de trabalho a nível nacional e para as regiões NUTS I e NUTS II.

Tem sido cada vez mais importante conhecer o estado do mercado de trabalho a níveis geográficos mais detalhados. Contudo, utilizando o método de estimação atual, não é possível produzir estimativas com uma precisão aceitável. Este problema é conhecido na literatura como "estimação em pequenos domínios". Têm sido propostos alguns métodos alternativos, entre os quais, o modelo Fay-Herriot, um modelo nível área que assume normalidade. Contudo, as suposições feitas neste modelo são muito restritivas e não são adequadas num contexto de dados de desemprego.

Neste sentido, Pereira *et al* (2018) propuseram modelos lineares generalizados usando uma abordagem bayesiana para a estimação do desemprego. Os autores fizeram uma análise espaço-temporal das estimativas obtidas por região NUTS III.

A partir do 4º trimestre de 2014, todos os edifícios residenciais da amostra do Inquérito ao Emprego (IE) foram georreferenciados. Para ter em conta esta informação, Pereira *et al* (2019) propuseram utilizar estes novos dados juntamente com informação dos indivíduos para modelar a intensidade dos pontos e as respetivas marcas usando um modelo de Cox log Gaussiano marcado. Nesta abordagem de processos pontuais espaciais, os pontos correspondem aos edifícios residenciais, e as marcas correspondem ao número de desempregados que vivem nesses edifícios.
Contudo, recentemente, os autores tiveram ainda acesso à georreferenciação de todos os edifícios residenciais no território nacional, incluindo os edifícios fora da amostra do IE. Consequentemente, já não é necessária a modelação da intensidade dos pontos. O método que propomos baseia-se num modelo para dados referenciados por pontos, também conhecido como modelo geoestatístico. Este modelo assume que os pontos são fixos e o interesse é a modelação das marcas. A inferência é feita com base nas aproximações de Laplace encaixadas e integradas (INLA).

Referências

Pereira, S., Turkman, K. F., Correia, L. (2018) Spatio-temporal analysis of regional unemployment rates: A comparison of model based approaches. *REVSTAT*. 16(4), 515-536.

Pereira, S., Turkman, K. F., Correia, L., Rue, H. (2019) Unemployment estimation - spatial point referenced methods and models. *Spatial Statistics* (in press). https://doi.org/10.1016/j.spasta.2019.01.004

Soraia Alexandra Gonçalves Pereira, **galardoada com o Prémio SPE 2018,** é licenciada em Matemática pela Faculdade de Ciências da Universidade do Porto, doutorada em Estatística e Investigação Operacional pela Faculdade de Ciências da Universidade de Lisboa e Mestre em Estatística pela mesma instituição.
Atualmente é Investigadora no Centro de Estatística e Aplicações da Universidade de Lisboa (CEAUL) e Professora Auxiliar Convidada na Faculdade de Ciências da Universidade de Lisboa.

Índice

Este boletim tem o apoio de:

FCT Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR