Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Anais

VIII SEMANÍSTICA

VIII Semana Acadêmica do Departamento de Estatística

da UFRGS

http://www.ufrgs.br/semanistica

Porto Alegre - 16 e 17 de outubro de 2017

Organização:

Promoção:

Conteúdo

1 Cartaz da VIII SEMANÍSTICA

SEMANÍSTICA

VIII SEMANA ACADÊMICA DA ESTATÍSTICA

16 e 17 de outubro de 2017

VIII SEMANA ACADÊMICA DA ESTATÍSTICA – CRONOGRAMA

Horário	Segunda-Feira - 16/10/2017	Terça-Feira - 17/10/2017
08:30 - 10:00	Minicurso 1 Uma breve introdução ao R Bruna Martini Dalmoro	Minicurso 2 Minicurso de Excel para Análises Estatísticas Aline Cafruni Gularte
10:00 - 10:30	Coffe Break	Coffe Break
10:30 - 11:15	Apresentações Orais	Conferência 1 Título: Introdução ao Processamento Estatístico de Imagens Fábio Mariano Bayer
11:15 - 12:00		Conferência 2 Título: Introdução à Estatística Espacial com Aplicações Márcia Helena Barbian

Local
Sala 115 do Prédio 43111 do Instituto de Matemática e Estatística – Campus do Vale – UFRGS

Informações e Inscrições
Site: http://www.ufrgs.br/semanistica
E-mail: semanistica@gmail.com

Datas Importantes
Inscrições até 13/10/2017
Submissão de Trabalhos: 28/08/2017 a 29/09/2017
Divulgação dos trabalhos aceitos: 06/10/2017

Organização: Departamento de Estatística IME-UFRGS

Apoio: Instituto de Matemática e Estatística UFRGS; UFRGS Universidade Federal do Rio Grande do Sul

2 Introdução

A VIII Semana Acadêmica da Estatística (VIII SEMANÍSTICA) foi realizada nos dias de 16 e 17 outubro de 2017, no Instituto de Matemática e Estatística - IME, Campus do Vale da UFRGS, Porto Alegre, RS. O evento engloba os mais variados temas dentro da área acadêmica e profissional.

O objetivo principal da SEMANÍSTICA é promover o desenvolvimento, aprimoramento e a divulgação da Estatística, entre diferentes perspectivas, acadêmica e/ou prática no campo de aplicação. A proposta da SEMANÍSTICA é promover a integração entre estudantes, professores e profissionais de diversas áreas que utilizam a Estatística como suporte de decisão em suas respectivas áreas de conhecimento. Propõe-se que o evento seja um cenário de aproximação e troca de experiências entre professores e alunos em diferentes áreas de conhecimento.

Como objetivos específicos da SEMANÍSTICA, podem-se citar: divulgar as contribuições recentes dos pesquisadores participantes promovendo-se o intercâmbio entre cientistas, alunos e profissionais aplicados; promover um maior contato entre pesquisadores do Departamento de Estatística da UFRGS e pesquisadores de outros departamentos, propiciando futuros trabalhos de pesquisa conjuntos; intensificar o contato e o intercâmbio científico entre profissionais da Região Sul e a iniciativa privada dentro das realidades do Estado do Rio Grande do Sul e do MERCOSUL; divulgar os diferentes métodos e aplicações de Estatística para discentes da graduação em Estatística, bem como discentes de pós-graduação e graduação das mais diversas áreas correlatas, tais como: Economia, Administração, Engenharia e Biomédicas.

Para maiores informações sobre a VIII SEMANÍSTICA (Semana Acadêmica da Estatística 2017) podem ser encontradas no site www.ufrgs.br/semanistica.

3 Agradecimentos

A VIII SEMANÍSTCA - Semana Acadêmica do Departamento de Estatística da UFRGS não teria sido possível sem o apoio das seguintes agências financiadoras e instituições:

- DEST-UFRGS - Departamento de Estatística da UFRGS
- IME-UFRGS - Instituto de Matemática e Estatística da UFRGS
- UFRGS - Universidade Federal do Rio Grande do Sul

A Comissão Organizadora da VIII SEMANÍSTICA agradece a colaboração de todos que se dedicaram anonimamente e sem interesses pessoais, em promover a integração entre alunos, professores e profissionais em estatística.

Comissão Organizadora

4 Comissão Organizadora Docente

- Cleber Bisognin (Departamento de Estatística-UFRGS)
- Danilo Marcondes Filho (Departamento de Estatística-UFRGS)
- Guilherme Pumi (Departamento de Estatística-UFRGS)
- Liane Werner (Departamento de Estatística-UFRGS)

5 Comissão Científica

- Cleber Bisognin (Departamento de Estatística-UFRGS)
- Danilo Marcondes Filho (Departamento de Estatística-UFRGS)
- Guilherme Pumi (Departamento de Estatística-UFRGS)
- Liane Werner (Departamento de Estatística-UFRGS)

6 Apresentação

A programação da VIII SEMANÍSTICA - Semana Acadêmica do Departamento de Estatística da Universidade Federal do Rio Grande do Sul englobou as seguintes atividades:

- 2 Conferências envolvendo pesquisas realizadas em diversas áreas da Estatística proferidas por pesquisadores convidados da Universidades do Rio Grande do Sul e da Universidade Federal de Santa Maria;
- 2 Minicursos sobre softwares para análise estatística de dados, ministrados por acadêmicas do curso de Bacharelado em Estatística da Universidade Federal do Rio grande do Sul.
- Comunicações orais apresentadas pelos participantes do evento;

7 Programação

VIII SEMANA ACADÊMICA DA ESTATÍSTICA – CRONOGRAMA

Horário	Segunda-Feira - 16/10/2017	Terça-Feira - 17/10/2017
08:30 - 10:00	Minicurso 1 Uma breve introdução ao R Bruna Martini Dalmoro	Minicurso 2 Minicurso de Excel para Análises Estatísticas Aline Cafruni Gularte
10:00 - 10:30	Coffe Break	Coffe Break
10:30 - 11:15	Apresentações Orais	Conferência 1 Título: Introdução ao Processamento Estatístico de Imagens Fábio Mariano Bayer
11:15 - 12:00		Conferência 2 Título: Introdução à Estatística Espacial com Aplicações Márcia Helena Barbian

Conferências:

(M1) Minicurso 1 - Bruna Martini Dalmoro - Acadêmica do Curso de Bacharelado em Estatística - UFRGS

Título: Uma breve introdução ao R

(M2) Minicurso 2 - Aline Cafruni Gularte - Acadêmica do Curso de Bacharelado em Estatística - UFRGS

Título: Minicurso de Excel para Análises Estatísticas

(C1) Conferência 1 - Prof. Dr. Fábio Mariano Bayer - Professor do Departamento de Estatística - UFSM

Título: Introdução ao Processamento Estatístico de Imagens

(C2) Conferência 2 – Profª. Drª. Márcia Helena Barbian - Professora do Departamento de Estatística - UFRGS

Título: Introdução a Estatística Espacial com Aplicações

8 Minicursos

Uma breve introdução ao R

Bruna Martini Dalmoro
Acadêmica do Curso de Bacharelado em Estatística - UFRGS

Resumo

Open source e gratuito. Multi-plataforma. Ele tem tudo, e se ainda não tem, alguém está fazendo. Linguagem criada especialmente para lidar com dados. Flexível, adaptável. Hoje é uma das linguagens de programação mais utilizadas, tanto cientificamente quanto analiticamente e é uma das linguagens que mais crescem no mundo. Ainda assim não se convenceu que valha a pena aprender R? Acha complicado? Nesta breve introdução ao R mostrarei minha visão como formanda sobre esta linguagem e te ajudarei a dar os primeiros passos.

Minicurso de Excel para Análises Estatísticas

Aline Cafruni Gularte
Acadêmica do Curso de Bacharelado em Estatística - UFRGS

Resumo

O Minicurso de Excel para Análise Estatística tem como objetivo apresentar aos alunos a importância de criar um banco de dados bem estruturado para facilitar a realização das análises estatísticas e conhecer as diversas funções que o Excel oferece para tais análises. Os conteúdos abordados são: - Conhecendo o banco de dados; - Validação de dados; - Ferramenta de filtro; - Função substituir; - Instalação do Suplemento Análise de Dados; - Medida Resumo; - Criação de Tabelas Dinâmicas; - Gráficos: Histograma, Setores, Pareto e de Linha.

9 Conferências

Conferência 1

Introdução ao Processamento Estatístico de Imagens

Prof. Dr. Fábio Mariano Bayer
Professor do Departamento de Estatística - UFSM

Resumo

Nesta palestra serão abordados aspectos básicos sobre a área de processamento de sinais e suas relações com a estatística. Será dada ênfase no processamento de imagens digitais. A relação entre análise de componentes principais e transformadas discretas será explorada, destacando resultados úteis para a compressão e a filtragem de imagens. Avanços recentes na linha de transformadas discretas de baixo custo computacional serão apresentados. Tópicos e trabalhos futuros na linha de processamento estatístico de imagens serão motivados.

Conferência 2

Introdução a Estatística Espacial com Aplicações

Profª. Drª. Márcia Helena Barbian
Professora do Departamento de Estatística - UFRGS

Resumo

Estatística espacial é um ramo da estatística que estuda métodos científicos para a coleta, descrição, visualização e análise de dados que possuem coordenadas geográficas. A grande diferença entre a estatística espacial e os outros métodos é considerar o georreferenciamento dos dados na modelagem. Nesse seminário apresentarei vários exemplos de diferentes aplicações da estatística espacial, além de abordar conceitos básicos dessa metodologia.

10 Comunicações Orais

Comunicação Oral 1:

Efeito da Má Especificação de Modelos nas Combinações de Previsão em Séries Temporais com Longa Dependência

Letícia Menegotto, Cleber Bisognin e Liane Werner

Resumo: Ao modelarmos processos estocásticos, é possível cometermos equívocos no tipo de processo ou mesmo no número de parâmetros do processo a ser ajustado em determinada série. O objetivo deste trabalho é verificar a influência da má especificação de modelos nas previsões e nas combinações de previsões através das medidas de acurácia quando a série apresenta a propriedade de longa dependência, uma vez que comumente séries temporais que apresentam esta propriedade são confundidas com séries temporais não estacionárias. Utilizando a técnica de Monte Carlo serão realizadas simulações para verificar esta influência, onde será calculada a média das medidas de acurácia calculadas para cada modelo a ser verificado. Analisando as simulações de Monte Carlo, observamos que na grande maioria das vezes as combinações de previsões têm melhor capacidade preditiva que o próprio modelo a partir do qual a série for gerada - neste caso, ARFIMA$(p; d; q)$. Finalmente será feita uma aplicação a dados reais, na qual será analisada a série temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores.

Comunicação Oral 2:

Avaliação do desempenho de Índices de Capacidade tradicionais diante de processos Não-Normais

Eduardo de Oliveira Correa e Danilo Marcondes Filho

Resumo: Índices de capacidade (IC) são amplamente usados para avaliar o desempenho dos processos industriais. Dado um processo operando sob condições estáveis e uma característica de qualidade representada por uma variável aleatória de interesse, os IC basicamente comparam a

variabilidade natural dessa variável em relação à amplitude das especificações do processo. Quanto maior a variabilidade, menor a capacidade do processo em produzir unidades dentro das especificações. Destacam-se no meio industrial os IC clássicos Cp, Cpk, disponíveis em rotinas de controle de qualidade. Entretanto, estes índices avaliam a capacidade do processo supondo distribuição Normal para a variável aleatória sob investigação. Devido à complexidade de processos produtivos atuais, as características de qualidade geram dados com distribuições com caldas longas e/ou assimétricas, tornando a avaliação da capacidade destes processos através dos IC clássicos bastante distorcida. Embora exista literatura disponível sobre IC para processos "não-normais", verifica-se pouca aplicação de tais abordagens nas indústrias. Este trabalho replica e estende o estudo de Somerville e Montgomery (1996), apresentando uma avaliação do erro de análise da capacidade de processos utilizando os índices Cp e Cpk, diante de variáveis apresentando distribuições de probabilidade com diferentes formas.

Comunicação Oral 3:

Perfil dos participantes em crimes de violência doméstica, no Rio Grande do Sul (Lei nº 11.340 - Lei Maria da Penha)

Helena Simeonidis Grillo e Patrícia Klarmann Ziegelmann

Resumo: Este estudo tem como objetivo apresentar o perfil dos participantes de crimes de feminicídio tentados e consumados no estado do Rio Grande do Sul de modo a auxiliar aos órgãos de segurança pública a responder à questão sobre a possibilidade de prevenção a este tipo de violência. A criação da Lei Maria da Penha (Lei n° 11.340/2006) criou mecanismos para coibir e prevenir a violência doméstica e familiar contra a mulher, permitindo que informações sejam coletadas, através dos registros de ocorrências, e estudos realizados. Para a análise foi utilizada a estatística descritiva. O resultado do estudo mostrou o perfil de um crime que acontece à noite, na residência da vítima, através do disparo de arma de fogo, no caso da morte, ou de uso de arma branca, no caso da tentativa, realizado por um homem branco, contra uma mulher branca, ambos com idade entre 18 e 24 anos, com pouca instrução, sem filhos, sem antecedentes registrados, devido ao fim o relacionamento, em sua maioria, quando o crime é consumado, ele é capturado pelos órgãos de segurança.

Comunicação Oral 4:

Mando de Campo e Gol Qualificado - Análise da Vantagem na Copa do Brasil

Alice Paul Waquil, Eduardo de Oliveira Horta e Jean Carlo Moraes

Resumo: No futebol, acredita-se que em confrontos de mata-mata – isto é, disputas eliminatórias com jogos de ida e volta – o time que faz o segundo jogo em seu estádio teria uma vantagem. Essa crença vem do fato, amplamente reconhecido na literatura científica, de que o fator local é uma vantagem numa partida de futebol. Logo, muitos pensam que fazer o segundo jogo com essa vantagem traria uma maior chance de classificação. Quando os confrontos estão empatados em número de pontos, precisa-se de um critério para definir o vencedor; os três mais usados são o saldo de gols, o gol qualificado (em que o vencedor do confronto será o time que marcar mais gols enquanto joga como visitante) e a disputa de pênaltis.

Esse estudo traz evidência de que decidir um confronto mata-mata em casa é um benefício quando olhado de forma geral, pois o mandante se classifica em aproximadamente 65% das disputas. Porém quando a decisão se dá por gol qualificado ou pênaltis o percentual de classificação é 20% menor, ou seja, esses critérios beneficiam o time visitante, se não dando a vantagem, ao menos equiparando as chances das duas equipes. Além disso, identificou-se que a probabilidade de classificação está relacionada com a diferença de qualidade entre os times.

Comunicação Oral 5:

Aplicação do método de séries temporais funcionais em linguagem R

Vitória Maria Martini Wendt e Eduardo de Oliveira Horta

Resumo: Dados funcionais estão cada vez mais em evidência principalmente no âmbito científico. Nesse sentido, propomos a implementação de uma ferramenta que padronize computacionalmente o uso de um método importante na área introduzido por Bathia et al (2010). Desta forma, foi desenvolvido o pacote em linguagem R ftsa2 que almeja tornar análises de séries temporais funcionais mais rápidas e universais, automatizando e melhorando processos.

Comunicação Oral 6:

Classificação de Doenças Cardíacas através de Eletrocardiogramas e Fonocardiogramas

Mikaela Baldasso, Marcio Valk e Airton Kist

Resumo: Uma grande parcela da população sofre de problemas relacionados a doenças do coração que estão entre as principais causas de morte em todo o mundo. Em particular, 1-2% da população mundial sofre de algum tipo de arritmia cardíaca que pode afetar pessoas das mais variadas faixas etárias. Recentemente o "National Institute of General Medical Science" (NIGMS) lançou um desafio com o objetivo de estimular a proposição de técnicas para classificação dos diferentes tipos de arritmias baseados em eletrocardiogramas (ECG's) e fonocardiogramas (PCG's) que podem ser vistos como séries temporais em que a técnica de classicação e agrupamento baseada em U-estatísticas pode ser aplicada. A utilização dessas técnicas depende fundamentalmente de medidas de distâncias ou similaridade que sejam capazes de capturar diferenças entre dois ECG's (ou PCG's), quando elas existem. Abordagens muito comuns na análise de sinais, como a filtragem, que elimina os ruídos que possivelmente poderiam afetar a classificação, devem ser consideradas. A partir disso, pode-se utilizar ferramentas comuns na análise de séries temporais, como a autocorrelação que é característica definidora podendo ser usada na classificação dos diferentes tipos de arritmia. Por m, neste estudo, os resultados são comparados aos disponibilizados pelo desafio sendo possível fazer uma comparação com outras técnicas propostas na literatura.

Comunicação Oral 7:

Avaliação da reconstrução de caractere em ancestral comum e estimação de correlações pelo modelo filogenético de variável latente

Lauren Alves Vieira e Gabriela Bettella Cybis

Resumo: O estudo de correlações entre variáveis fenotípicas ao longo da evolução é um dos problemas centrais da biologia evolutiva. O modelo filogenético de variável latente (Cybis et al. 2015) é uma opção para a estimação de tais correlações no contexto das filogenias bayesianas. O modelo permite a estimação simultânea de correlações entre variáveis contínuas, discretas ordenadas e discretas sem ordenamento, controlando para a história evolutiva compartilhada das amostras. Neste trabalho nós realizamos uma aplicação do modelo a um conjunto de dados de morcegos, que contem uma variável contínua e uma discreta não ordenada, na qual estimamos a correlação evolutiva entre as variáveis e reconstruímos o valor dessas variáveis no ancestral comum a todas as espécies em estudo. Como nos modelos ordenados a aplicação do modelo depende da escolha de um estado de referência, nós realizamos uma análise de sensibilidade, verificando que em geral estas estimativas são robustas à escolha do referencial.

Comunicação Oral 8:

Estimativa de casos de salmonelose humana atribuída às fontes de alimento de origem animal

Waldemir Santiago Neto, Luís Gustavo Corbellini, Vanessa Bielefeldt Leotti e Tine Hald

Resumo: Tem se estimado que alimentos contaminados estejam relacionados com diversas doenças infecciosas e sejam responsáveis por 2,2 milhões de mortes ao redor do mundo anualmente. Salmonella enterica é considerada uma das principais causas de gastroenterites e bacteremias e a maioria de seus subtipos é encontrada em animais de sangue quente. Dados do Brasil apontam que as salmonelas são as principais causas de toxinfecção alimentar. A fim de obter compreensão da dinâmica de infecções por salmonela em humanos, um modelo bayesiano comparando a ocorrência de sorovares de Salmonella em animais e humanos foi utilizado para atribuir casos de salmonelose a frangos de corte, perus, porcos, galinhas poedeiras e surtos no Rio Grande do Sul (RS). Dados de salmonela para animais e seres humanos, cobrindo o período de 2000 a 2015, foram obtidos principalmente de estudos e relatórios publicados pela Secretaria de Vigilância em Saúde do Ministério da Saúde. A disponibilidade de fontes de alimento para consumo foi derivada dos dados de produção do Instituto Brasileiro de Geografia e Estatística. A principal fonte de salmonelose humana no RS foi estimada como sendo galinhas poedeiras, com 92,1% [3963 casos, intervalo de credibilidade de 95% (ICr 95%) 3734-4159] de casos, seguido de 5,6% atribuídos a suínos de fora do RS (242 casos, ICr 95% 122-409). dos quais foi causada por S. Enteritidis. Este trabalho possibilita destacar diferenças na epidemiologia da Salmonella, foco de vigilância e hábitos alimentares no estado.

Efeito da Má Especificação de Modelos nas Combinações de Previsão em Séries Temporais com Longa Dependência

Cleber Bisognin[1]

Letícia Menegotto [2]

Liane Werner[3]

Resumo: Ao modelarmos processos estocásticos, é possível cometermos equívocos no tipo de processo ou mesmo no número de parâmetros do processo a ser ajustado em determinada série. O objetivo deste trabalho é verificar a influência da má especificação de modelos nas previsões e nas combinações de previsões através das medidas de acurácia quando a série apresenta a propriedade de longa dependência, uma vez que comumente séries temporais que apresentam esta propriedade são confundidas com séries temporais não estacionárias. Utilizando a técnica de Monte Carlo serão realizadas simulações para verificar esta influência, onde será calculada a média das medidas de acurácia calculadas para cada modelo a ser verificado. Analisando as simulações de Monte Carlo, observamos que na grande maioria das vezes as combinações de previsões têm melhor capacidade preditiva que o próprio modelo a partir do qual a série for gerada - neste caso, ARFIMA(p, d, q). Finalmente será feita uma aplicação a dados reais, na qual será analisada a série temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores.

Palavras-chave: *Combinação de Previsões, Modelagem Estatística, Previsões, Longa Dependência, Má Especificação de Modelos.*

1 Introdução

De acordo com Abraham e Ledolter (2009), o ser humano está sempre fazendo previsões, que consiste em uma atividade indispensável no planejamento, na definição da estratégia e na tomada de decisões orientadas para o futuro, tanto em nível individual como em nível organizacional.

Uma vez que previsões envolvem eventos futuros e estes, por sua vez, envolvem a incerteza, tem-se que as previsões, em geral, não são perfeitas. O objetivo, ao realizarmos uma previsão, é reduzir o erro da mesma (Abraham e Ledolter, 2009). Para produzir uma previsão que apresente um erro pequeno, é necessário utilizar uma técnica de previsão adequada, seja por meio de um

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *cbisognin@ufrgs.br*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *leticia.menegotto@gmail.com*
[3]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *liane.werner@ufrgs.br*

modelo ou uma combinação de previsões oriundas de várias técnicas de previsão, e para tanto, é preciso obter critérios de acurácia (Werner, 2005).

Conforme Morettin e Toloi (2006), uma das suposições mais frequentes que se faz a respeito de uma série temporal é que se desenvolva no tempo, aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável (estacionaridade). Todavia, a maior parte das séries que se encontra na prática apresenta alguma forma de não estacionariedade, pois mudam suas características estocásticas ao longo do tempo de observação, sendo conhecidas por séries não estacionárias. Segundo Box e Jenkins (1976) é possível obter séries estacionárias pela diferenciação (d), valor este assumido como número inteiro. Uma diferenciação fracionária é o caso geral do processo de diferenciação, modelos que usam este procedimento são conhecidos por modelos de longa dependência. Nas últimas décadas, tem ocorrido grande interesse no estudo de séries temporais com longa dependência, que iniciaram com os estudos de Hurst em 1951 quanto investigava a série temporal dos níveis mensais do rio Nilo (Bisognin, 2007). Segundo Lima et al. (2007) modelos de longa dependência são capazes de produzir previsões com menor erro quadrado médio, uma importante medida de acurácia.

Quando as medidas de acurácia são boas, acreditamos que um modelo adequado foi encontrado. Porém, é preciso ter cuidado na especificação do modelo. Para Queiroz (2016) a solução de problemas estatísticos está baseada na teoria da máxima verossimilhança, que tem como suposição básica de que o modelo escolhido para analisar os dados é, de fato, o modelo gerador destes. Quando isso não acontece, ou seja, quando ocorre uma má especificação do modelo, utilizar os procedimentos inferenciais usuais pode resultar em conclusões errôneas, gerando interpretações equivocadas.

Frente a isto, o objetivo deste trabalho é verificar a influência da má especificação de modelos na previsão e nas combinações de previsões através das medidas de acurácia, tendo como modelo gerador uma série que apresenta longa dependência. Tal objetivo deve-se ao fato que algumas séries temporais, podem ser tratadas como estacionária ou não estacionárias, ou seja, quando analisamos tais series com testes de raiz unitária, o *p-valor* de um teste é aproximadamente 0.1 e de outro menos que 0.05, como é o caso da série temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores. Maiores detalhes serão abordados na Seção 4.

2 Técnicas de Previsão

Nesta seção apresentamos os modelos utilizados para análise e previsão de séries temporais. Serão utilizados os modelos ARMA(p, q), ARIMA(p, d, q), ARFIMA(p, d, q) e suavização exponencial, além de três métodos para realizar combinações de previsões, a saber: variância mínima,

por regressão e média aritmética.

Inicialmente definimos os processos ARIMA(p,d,q) proposto por Box e Jenkins (1976).

Definição 1. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico satisfazendo a equação

$$\phi(\mathcal{B})(1-\mathcal{B})^d(X_t-\mu)=\theta(\mathcal{B})\varepsilon_t, \tag{1}$$

onde μ é a média do processo, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é o processo ruído branco, $\mathcal{B}$ é o operador *defasagem ou de retardo*, isto é, $\mathcal{B}^j(X_t)=X_{t-j}$, para $j\in\mathbb{N}$, $\phi(\cdot)$ e $\theta(\cdot)$ são os polinômios de ordem p e q, respectivamente, definidos por

$$\phi(z)=\sum_{\ell=0}^{p}(-\phi_\ell)\,z^\ell \quad \text{e} \quad \theta(z)=\sum_{m=0}^{q}(-\theta_m)\,z^m, \tag{2}$$

onde ϕ_ℓ, $1\leqslant\ell\leqslant p$ e θ_m, $1\leqslant m\leqslant q$, são constantes reais e $\phi_0=-1=\theta_0$. Então, $\{X_t\}_{t\in\mathbb{Z}}$ é um *processo auto-regressivo integrado de média móvel de ordem* (p,d,q), denotado por ARIMA(p,d,q), onde $d\in\mathbb{Z}_{\geqslant}$ é *o grau de diferenciação.*

Observação 1. Na Definição 1, quando $d=0$, temos os processos ARMA(p,q).

Durante as últimas décadas, houve muito interesse em estudar séries temporais com a propriedade de longa dependência. Utilizando a definição de longa dependência, Granger e Joyeux (1980), Hosking (1981), Hosking (1984) e Geweke e Porter-Hudak (1983) apresentam os *processos auto-regressivos fracionariamente integrados de média móvel* (ARFIMA(p,d,q)) como um exemplo de processos com a característica de longa dependência. A seguir definimos os processos ARFIMA(p,d,q).

Definição 2. Seja $\{X_t\}_{t\in\mathbb{Z}}$ um processo estocástico satisfazendo a equação

$$\phi(\mathcal{B})(1-\mathcal{B})^d(X_t-\mu)=\theta(\mathcal{B})\varepsilon_t, \tag{3}$$

onde μ é a média do processo, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ é o processo ruído branco, $\mathcal{B}$ é o operador de *defasagem* ou de *retardo*, isto é, $\mathcal{B}^j(X_t)=X_{t-j}$, para todo $j\in\mathbb{N}$, $\phi(\cdot)$ e $\theta(\cdot)$ são os polinômios de ordem p e q, respectivamente, definidos por

$$\phi(z) \quad = \quad \sum_{\ell=0}^{p}(-\phi_\ell)\,z^\ell, \quad \theta(z)=\sum_{m=0}^{q}(-\theta_m)\,z^m, \tag{4}$$

onde ϕ_ℓ, $1\leqslant\ell\leqslant p$ e θ_m, $1\leqslant m\leqslant q$, são constantes reais e $\phi_0=-1=\theta_0$. Então, $\{X_t\}_{t\in\mathbb{Z}}$ é um *processo auto-regressivo fracionariamente integrado de média móvel de ordem* (p,d,q) *com média* μ, denotado por ARFIMA(p,d,q), onde d é *o grau de diferenciação fracionário.*

Hosking (1981) demonstra que os processos ARFIMA(p,d,q) são estacionários se $d<\frac{1}{2}$ e as

raízes da equação $\phi(z) = 0$ estão fora do círculo unitário; e é inversível se $d > -\frac{1}{2}$ e as raízes da equação $\theta(z) = 0$ estão fora do círculo unitário.

Além destes, os modelos de suavização exponencial, devido a sua simplicidade, facilidade de ajustes e boa acurácia, são os mais utilizados frente a outras técnicas de previsão, segundo Pellegrini (2000). Como assumem que os valores extremos da série são flutuações aleatórias, o propósito destes modelos é identificar um padrão básico na série temporal a ser analisada (Morettin e Toloi, 2006). Estes modelos valorizam mais as últimas observações na série temporal através da ponderação exponencial das mesmas, de acordo com a proximidade ao período da previsão h. Os métodos mais tradicionais de suavização exponencial são: (i) a suavização exponencial simples, para séries que apresentam apenas variações em torno de um nível; (ii) o modelo linear de Holt, para as séries que apresentam a componente de tendência e (iii) os modelos de Holt-Winters, quando a série apresenta tanto o componente de tendência quanto o componente sazonal (Makridakis et al., 1998).

A seguir definimos os modelos lineares de Holt. Maiores detalhes sobre estes modelos e os de Holt-Winters podem ser encontrados em Makridakis et al. (1998) e em Morettin e Toloi (2006).

Modelos Lineares de Holt

Seja uma série temporal $\{X_t\}_{t=1}^{n}$. No caso dos modelos lineares de Holt consideramos que tal série é formada pela soma do nível, tendência e um erro aleatório, como segue:

$$X_t = L_t + T_t + \varepsilon_t, \quad \text{para} \quad t = 1, \cdots, n. \tag{5}$$

As estimativas do nível da série no tempo t, denotado por L_t e da tendência, denotada por T_t, são dadas, respectivamente por

$$L_t = \alpha X_t + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{6}$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}, \tag{7}$$

onde α é o coeficiente de ponderação exponencial do nível ($0 \leqslant \alpha \leqslant 1$) e β é o coeficiente de ponderação exponencial da tendência ($0 \leqslant \beta \leqslant 1$).

As previsões h passos a frente são dadas por

$$\widehat{X}_t(h) = L_t + hT_t. \tag{8}$$

A notação $\widehat{X}_t(h)$ indica a previsão de origem t e horizonte $h \geqslant 1$.

Além destes modelos, um método comumente utilizado para melhorar a acurácia das previsões é a combinação de previsões. Segundo Costantini e Pappalardo (2010), este método consiste em

utilizar um mecanismo para captar os diversos fatores que afetam cada técnica de previsão individual usada como base na obtenção da previsão combinada.

O método da variância mínima, proposto por Bates e Granger (1969) consiste em realizar a combinação linear de duas previsões com diferentes pesos. Neste método a combinação das previsões é obtida atribuindo-se um peso para cada uma das previsões individuais que serão combinadas. Sua estrutura é apresentada conforme equação (9).

$$F_c = wF_1 + (1 - w)F_2 \tag{9}$$

onde w é o peso atribuído a previsão de menor variância e F_1 e F_2 são as previsões individuais a serem combinadas.

Para a obtenção dos pesos descritos na equação (9) é interessante atribuir menor peso às previsões de maior variabilidade nos erros e considerar a correlação existente entre os erros das duas previsões individuais realizadas. O peso para a previsão com menor variabilidade nos erros é obtido conforme equação (10).

$$w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{10}$$

onde ρ é o valor da correlação linear entre os erros das previsões obtidas em F_1 e F_2, σ_1^2 é a variância dos erros de previsão de F_1 e σ_2^2 é a variância dos erros de previsão de F_2.

O método mais popular de combinação de previsões individuais é a média aritmética, pois além de ser um dos métodos mais conhecidos é fácil de calcular. Segundo De Menezes et al. (2000), uma possível resposta para o sucesso da média pode estar associada à instabilidade dos pesos ao longo do tempo na matriz de covariância dos erros das previsões individuais.

Um fato que chamou a atenção de Granger e Ramanathan (1984) é que a combinação de previsões poderia ser uma forma estruturada de regressão, utilizando o Método dos Mínimos Quadrados Ordinários (MQO), tendo a previsão combinada como variável resposta e as previsões individuais como variáveis explicativas.

De acordo com Makridakis et al. (1998), a palavra acurácia refere-se a habilidade do modelo ou da combinação em reproduzir os dados que já são conhecidos (qualidade do ajuste). Porém para optar qual técnica de previsão - individual ou combinação - é a mais adequada, faz-se necessário obter medidas de acurácia. Neste trabalho iremos utilizar as medidas de acurácia *Root Mean Squared Error* (RMSE), erro médio absoluto (MAE), erro percentual médio (MPE), erro percentual médio absoluto (MAPE) e erro médio de previsão (ME).

3 Simulações de Monte Carlo

Nesta seção serão apresentados os resultados tendo como base os procedimentos metodológicos de simulação de Monte Carlo. O procedimento consiste em gerar séries temporais (amostras) dos processos ARFIMA(p, d, q), com $0 < d < 0.5$ (ver Definição 2). As séries temporais foram geradas utilizando a rotina *fracdiff.sim*, do pacote *fracdiff* do *software* R 3.4.0. Após foram ajustados as séries temporais geradas processos ARFIMA(p, d, q), utilizando a rotina *arfima*, do pacote *forecast*, processos ARIMA(p, d, q) e ARMA(p, q), utilizando a rotina *auto.arima*, também do pacote *forecast*, e o modelo de suavização exponencial, mais conhecido como Modelo Linear de Holt, utilizando a rotina *HoltWinters*, do pacote *stats*.

No caso dos processos ARFIMA(p, d, q), a rotina seleciona automaticamente os valores de p e q usando o algoritmo Khandakar e Hyndman (2008) e o algoritmo de Haslett e Raftery (1989), que é baseado no método da máxima verossimilhança, para estimar o parâmetros incluindo o parâmetro de longa dependência d.

Para a estimação dos parâmetros dos processos dos modelos ARIMA e ARMA foi utilizado a rotina *auto.arima* que calcula a verossimilhança exata via representação de Estado de Espaço do modelo enquanto as inovações são encontradas via Filtro de Kalmann. A estimação dos coeficientes dos polinômios é baseada em Gardner et al. (1980).

Para os modelos de suavização exponencial foi utilizado a rotina *HoltWinters*. A função tenta encontrar valores ótimos para α, e/ou β minimizando o erro quadrado de previsão de um passo à frente quando nenhum dos parâmetros de suavização é informado pelo usuário.

Após ajuste de modelos e teste de resíduos (rotina *Box.test*) foram calculadas as previsões dos n valores da série temporal gerada e também serão aplicadas as técnicas de combinação previsão de variância mínima, média aritmética e por regressão, como base nos modelos individuais previamente obtidos combinados dois a dois. As técnicas de combinação de previsão foram implementadas no mesmo *software*.

Calculadas as previsões, o próximo passo é calcular as medidas de acurácia ME (média dos erros de previsão), RMSE (raiz do erro médio quadrático), MAE (erro médio absoluto de previsão), MPE (percentual médio de erro) e pelo MAPE (percentual médio absoluto de erro). As medidas foram calculadas utilizando-se a rotina *accuracy* do pacote *forecast*.

As Tabelas 1 a 5 contemplam os resultados de simulação de Monte Carlo para o procedimento descrito acima, e apresentam as médias das medidas de acurácia, para as $re = 1000$ replicações, das previsões utilizando os modelos e os três tipos de combinação de previsão. Foram geradas séries temporais, para cinco composições dos seguintes valores dos parâmetros $d = 0.3$, $p \in \{0, 1\}$,

$\phi_1 \in \{-0.8, 0.8\}$, $q \in \{0, 1\}$, $\theta_1 \in \{-0.2, 0.2\}$, com $n = 1000$.

Analisando a Tabela 1, quando geramos amostras dos processos ARFIMA(p, d, q), com $d = 0.3$, $p = 1$, $\phi_1 = -0.8$ e $q = 0$, concluímos que os modelos com menor ME são os modelos ARIMA(p, d, q), a combinação de previsões de variância mínima utilizando os modelos ARMA(p, q) e Holt, e a combinação de previsões por regressão e por média dos modelos ARIMA(p, d, q) e ARMA(p, q). Com menor RMSE, MAE e MAPE é a combinação de previsões de variância mínima utilizando os modelos ARIMA(p, d, q) e Holt. Já o menor MPE foi encontrado na combinação de previsões por variância mínima dos modelos ARIMA(p, d, q) e Holt.

Pela análise da Tabela 2, quando geramos amostras dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 1$, $\phi_1 = 0.8$ e $q = 0$, concluímos que o modelo com menor ME é o modelo ARMA(p, q) e a combinação de previsões por média dos modelos ARFIMA(p, d, q) e ARMA(p, q). O modelo ARMA(p, q) possui menor RMSE e MAPE, enquanto o modelo ARFIMA(p, d, q) possui menor MAE e a combinação de previsões por variância mínima utilizando os modelos ARIMA(p, d, q) e Holt possui menor MPE.

Pela Tabela 3, quando geramos amostras dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 0$, $q = 1$ e $\theta_1 = 0.2$, verificamos que o modelo ARIMA(p, d, q) e a combinação de previsões por regressão dos modelos ARFIMA(p, d, q) e ARMA(p, q) possuem menor ME, a combinação de previsões por média dos modelos ARFIMA(p, d, q) e Holt possui menores RMSE, MAE e MAPE, enquanto a combinação de previsões por regressão dos modelos ARIMA(p, d, q) e Holt possui menor MPE.

Através da Tabela 4, quando geramos amostras dos processos ARFIMA(p, d, q), com $d = 0.3$, $p = 0$, $q = 1$ e $\theta_1 = -0.2$, constatamos que o modelo Holt, a combinação de previsões por variância mínima dos modelos ARFIMA(p, d, q) e Holt e a combinação de previsões por regressão dos modelos ARFIMA(p, d, q) e ARMA(p, q) e dos modelos ARMA(p, q) e Holt possuem menor ME. A combinação de previsões por média dos modelos ARFIMA(p, d, q) e Holt possui menor RMSE, MAE, MPE e MAPE.

Analisando a Tabela 5, quando geramos amostras dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 1$, $\phi_1 = 0.8$, $q = 1$ e $\theta_1 = -0.2$, observamos que o modelo ARIMA(p, d, q), a combinação de previsões por variância mínima dos modelos ARIMA(p, d, q) e Holt e a combinação de previsões por média dos modelos ARIMA(p, d, q) e ARMA(p, q) e dos modelos ARMA(p, q) e Holt possuem menor ME. A combinação de previsões por média dos modelos ARFIMA(p, d, q) e ARMA(p, q) possui menor RMSE, a combinação de previsões por regressão dos modelos ARFIMA(p, d, q) e ARMA(p, q) possui menor MAE, enquanto os menores valores de MPE e MAPE ocorrem na

combinação de previsões por variância mínima dos modelos ARIMA(p, d, q) e ARMA(p, q).

Tabela 1: Medidas de Acurácia para séries temporais geradas a partir dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 1$, $\phi_1 = -0.8$, $q = 0$ e $n = 1000$.

Modelos Ajustados	ME	RMSE	MAE	MPE	MAPE
ARFIMA	-0.0002	0.9978	0.7963	-1.0660	8.2262
ARIMA	**0.0000**	1.0024	0.7996	-1.0519	8.2601
ARMA	-0.0001	0.9974	0.7959	-1.0669	8.2300
Holt	-0.0010	1.6295	1.2831	-2.2506	13.3495
Combinação de Previsões - Variância Mínima					
ARFIMA/ARIMA	0.0006	0.9977	0.7962	-1.0573	8.2265
ARFIMA/ARMA	0.0019	0.9989	0.7974	-1.0457	8.2360
ARFIMA/Holt	-0.0002	0.7434	0.5933	-0.8310	6.1339
ARIMA/ARMA	0.0003	1.0000	0.7977	-1.0521	8.2406
ARIMA/Holt	-0.0006	**0.7328**	**0.5847**	**-0.8137**	**6.0455**
ARMA/Holt	**0.0000**	0.7434	0.5934	-0.8270	6.1391
Combinação de Previsões - Regressão					
ARFIMA/ARIMA	0.0002	0.9958	0.7943	-1.0556	8.2047
ARFIMA/ARMA	-0.0001	0.9956	0.7942	-1.0637	8.2093
ARFIMA/Holt	-0.0002	0.7418	0.5919	-0.8328	6.1254
ARIMA/ARMA	**0.0000**	0.9954	0.7942	-1.0586	8.2046
ARIMA/Holt	0.0015	0.7342	0.5857	-0.7952	6.0571
ARMA/Holt	0.0001	0.7443	0.5941	-0.8277	6.1465
Combinação de Previsões - Média					
ARFIMA/ARIMA	-0.0001	0.9974	0.7962	-1.0598	8.2289
ARFIMA/ARMA	0.0001	0.9959	0.7948	-1.0633	8.2179
ARFIMA/Holt	0.0002	1.0301	0.8157	-1.4743	8.5070
ARIMA/ARMA	**0.0000**	0.9963	0.7953	-1.0566	8.2141
ARIMA/Holt	-0.0004	1.0273	0.8145	-1.4675	8.4935
ARMA/Holt	-0.0007	1.0314	0.8175	-1.4824	8.5233

Fonte: Autores.

Tabela 2: Medidas de Acurácia para séries temporais geradas a partir dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 1$, $\phi_1 = 0.8$, $q = 0$ e $n = 1000$.

Modelos Ajustados	ME	RMSE	MAE	MPE	MAPE
ARFIMA	-0.0001	0.9976	**0.7960**	-1.1196	12.0266
ARIMA	0.0004	1.0125	0.8079	-1.4654	13.3490
ARMA	**0.0000**	**0.9973**	0.7961	-1.3920	**11.8463**
Holt	-0.0010	1.0461	0.8348	-1.1510	12.1322
Combinação de Previsões - Variância Mínima					
ARFIMA/ARIMA	-0.0001	0.9984	0.7964	-3.5889	15.1945
ARFIMA/ARMA	0.0024	1.0021	0.7995	-1.7194	12.4937
ARFIMA/Holt	-0.0002	1.1847	0.9454	-1.4550	14.4596
ARIMA/ARMA	0.0002	1.0137	0.8085	-1.8338	13.2070
ARIMA/Holt	-0.0004	1.1928	0.9514	0.7557	15.8453
ARMA/Holt	-0.0001	1.1854	0.9461	-8.9787	22.2009
Combinação de Previsões - Regressão					
ARFIMA/ARIMA	0.0032	0.9992	0.7973	-4.9606	15.2409
ARFIMA/ARMA	0.0002	0.9984	0.7968	-3.2177	14.9527
ARFIMA/Holt	0.0007	1.1836	0.9445	3.6867	19.9684
ARIMA/ARMA	0.0021	0.9987	0.7971	1.2228	14.7097
ARIMA/Holt	0.0435	1.1887	0.9487	-1.6423	14.7383
ARMA/Holt	-0.0007	1.1862	0.9467	-2.3120	14.3208
Combinação de Previsões - Média					
ARFIMA/ARIMA	0.0003	1.0018	0.7994	-2.1029	12.3018
ARFIMA/ARMA	**0.0000**	0.9990	0.7972	-2.7820	13.8283
ARFIMA/Holt	-0.0001	1.1069	0.8836	-1.8229	13.9846
ARIMA/ARMA	0.0002	1.0015	0.7990	-1.9371	12.2046
ARIMA/Holt	-0.0001	1.1074	0.8836	**0.4485**	13.9419
ARMA/Holt	0.0003	1.1071	0.8833	-2.2674	13.9702

Fonte: Autores.

Tabela 3: Medidas de Acurácia para séries temporais geradas a partir dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 0$, $q = 1$, $\theta_1 = 0.2$ e $n = 1000$.

Modelos Ajustados	ME	RMSE	MAE	MPE	MAPE
ARFIMA	0.0006	0.9985	0.7967	-1.0285	8.1506
ARIMA	**0.0000**	1.0034	0.8002	-1.0077	8.1807
ARMA	0.0001	0.9982	0.7969	-1.0301	8.1384
Holt	0.0007	1.0993	0.8697	-0.9323	8.8585
Combinação de Previsões - Variância Mínima					
ARFIMA/ARIMA	-0.0001	0.9976	0.7964	-1.0268	8.1329
ARFIMA/ARMA	0.0008	0.9968	0.7957	-1.0193	8.1267
ARFIMA/Holt	-0.0001	0.8740	0.6975	-0.9154	7.1440
ARIMA/ARMA	-0.0006	1.0023	0.7996	-1.0200	8.1733
ARIMA/Holt	-0.0003	0.8876	0.7084	-0.8978	7.2432
ARMA/Holt	-0.0001	0.8898	0.7101	-0.9316	7.2750
Combinação de Previsões - Regressão					
ARFIMA/ARIMA	0.0005	0.9959	0.7953	-1.0187	8.1292
ARFIMA/ARMA	**0.0000**	0.9974	0.7960	-1.0326	8.1400
ARFIMA/Holt	0.0003	0.8743	0.6977	-0.9076	7.1291
ARIMA/ARMA	0.0003	0.9964	0.7954	-1.0225	8.1295
ARIMA/Holt	0.0037	0.8856	0.7067	-0.8535	7.2144
ARMA/Holt	0.0000	0.8890	0.7094	-0.9246	7.2476
Combinação de Previsões - Média					
ARFIMA/ARIMA	-0.0002	0.9968	0.7957	-1.0200	8.1394
ARFIMA/ARMA	0.0002	0.9969	0.7957	-1.0284	8.1302
ARFIMA/Holt	0.0001	**0.8463**	**0.6736**	-0.8135	**6.8733**
ARIMA/ARMA	0.0006	0.9968	0.7954	-1.0134	8.1374
ARIMA/Holt	0.0004	0.8546	0.6806	**-0.8049**	6.9493
ARMA/Holt	-0.0001	0.8524	0.6788	-0.8199	6.9314

Fonte: Autores.

Tabela 4: Medidas de Acurácia para séries temporais geradas a partir dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 0$, $q = 1$, $\theta_1 = -0.2$ e $n = 1000$.

Modelos Ajustados	ME	RMSE	MAE	MPE	MAPE
ARFIMA	0.0001	0.9979	0.7964	-1.0469	8.1889
ARIMA	-0.0005	1.0013	0.7989	-1.0171	8.1995
ARMA	0.0001	0.9977	0.7964	-1.0466	8.1839
Holt	**0.0000**	1.0975	0.8739	-0.8456	8.9308
Combinação de Previsões - Variância Mínima					
ARFIMA/ARIMA	-0.0005	0.9967	0.7953	-1.0475	8.1866
ARFIMA/ARMA	-0.0008	0.9977	0.7960	-1.0515	8.1781
ARFIMA/Holt	**0.0000**	0.9598	0.7659	-1.0030	7.8554
ARIMA/ARMA	-0.0002	0.9987	0.7967	-1.0242	8.1886
ARIMA/Holt	-0.0001	0.9488	0.7574	-0.9539	7.7708
ARMA/Holt	0.0002	0.9613	0.7676	-1.0059	7.8788
Combinação de Previsões - Regressão					
ARFIMA/ARIMA	0.0008	0.9980	0.7965	-1.0286	8.1682
ARFIMA/ARMA	**0.0000**	0.9966	0.7955	-1.0480	8.1818
ARFIMA/Holt	0.0004	0.9581	0.7645	-1.0036	7.8737
ARIMA/ARMA	0.0005	0.9960	0.7952	-1.0333	8.1689
ARIMA/Holt	0.0046	0.9479	0.7568	-0.9043	7.7581
ARMA/Holt	**0.0000**	0.9618	0.7676	-1.0141	7.8986
Combinação de Previsões - Média					
ARFIMA/ARIMA	-0.0002	0.9979	0.7965	-1.0308	8.1736
ARFIMA/ARMA	-0.0003	0.9975	0.7963	-1.0536	8.1892
ARFIMA/Holt	0.0007	**0.8348**	**0.6655**	**-0.7557**	**6.8197**
ARIMA/ARMA	0.0002	0.9969	0.7949	-1.0286	8.1638
ARIMA/Holt	-0.0018	0.8349	0.6657	-0.7659	6.8285
ARMA/Holt	-0.0004	0.8357	0.6667	-0.7739	6.8465

Fonte: Autores.

Tabela 5: Medidas de Acurácia para séries temporais geradas a partir dos processos ARFIMA(p, d, q), quando $d = 0.3$, $p = 1$, $\phi_1 = 0.8$ $q = 1$, $\theta_1 = -0.2$ e $n = 1000$.

Modelos Ajustados	ME	RMSE	MAE	MPE	MAPE
ARFIMA	0.0002	0.9979	0.7965	-3.9851	19.9434
ARIMA	**0.0000**	1.0106	0.8058	0.2030	17.2564
ARMA	0.0001	0.9969	0.7957	-4.0811	20.5178
Holt	0.0009	1.0915	0.8706	0.5765	20.2942
Combinação de Previsões - Variância Mínima					
ARFIMA/ARIMA	0.0002	0.9973	0.7958	5.8007	24.2898
ARFIMA/ARMA	0.0048	1.0030	0.8003	-30.4502	45.4219
ARFIMA/Holt	0.0002	1.4316	1.1431	-5.9641	29.9762
ARIMA/ARMA	0.0002	1.0110	0.8065	**0.0863**	**17.0525**
ARIMA/Holt	**0.0000**	1.4355	1.1455	-0.1687	26.1373
ARMA/Holt	-0.0002	1.4300	1.1412	-6.4398	32.0171
Combinação de Previsões - Regressão					
ARFIMA/ARIMA	0.0045	0.9970	0.7956	-4.2765	20.8610
ARFIMA/ARMA	0.0001	0.9967	0.7957	-0.8420	19.9359
ARFIMA/Holt	0.0006	1.4319	1.1433	-1.3433	24.5153
ARIMA/ARMA	0.0031	0.9965	**0.7949**	-2.6315	17.9124
ARIMA/Holt	0.0477	1.4343	1.1449	2.1888	27.7956
ARMA/Holt	-0.0008	1.4279	1.1398	7.0087	33.1381
Combinação de Previsões - Média					
ARFIMA/ARIMA	-0.0003	1.0009	0.7981	-0.5809	21.3508
ARFIMA/ARMA	0.0002	**0.9964**	0.7954	-2.0612	17.6468
ARFIMA/Holt	0.0004	1.2619	1.0068	-0.3275	23.5462
ARIMA/ARMA	**0.0000**	0.9997	0.7978	-3.8775	20.1004
ARIMA/Holt	0.0009	1.2627	1.0073	-25.1495	49.3979
ARMA/Holt	**0.0000**	1.2593	1.0050	-0.7913	22.2157

Fonte: Autores.

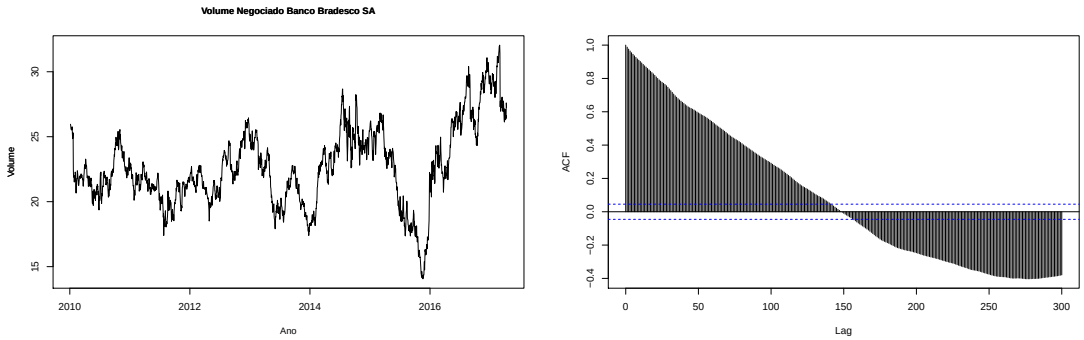
4 Aplicação a Dados Reais

A seguir analisamos a série temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores utilizando a metodologia desenvolvida neste trabalho. A etapa da obtenção de dados neste artigo, consistiu em resgatar dados históricos do site Yahoo Finanças (https://br.financas.yahoo.com/). Serão utilizadas as 1853 observações diárias disponíveis, de 04/01/2010 a 27/06/2017. Os dados foram acessados em 04/08/2017. O uso deste período se deve ao fato de o período de dados mais completo disponíveis na internet e que foram encontrados pelos autores.

Uma vez que se busca uma técnica adequada prever o valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores, obteve-se as previsões utilizando os modelos de Suavização Exponencial, ARFIMA(p, d, q), ARIMA(p, d, q) e ARMA(p, q) e suas respectivas combinações, utilizando dois modelos base.

A Figura 1 apresenta o gráfico da séries temporal e da função de autocorrelação amostral. Podemos perceber, pelo gráfico da série temporal e pela sua função de autocorrelação amostral que a série pode ser tratada como estacionária com a propriedade de longa dependência, mas também pode ser tratada como não estacionária. Foram aplicados os testes de raiz unitária de Dickey-Fuller, que apresentou $p-valor = 0.1551$, e de Phillips-Perron, que apresentou $p-valor =$

0.04145. Ambos testam as hipóteses: H_0: série temporal não estacionária versus H_1: série temporal estacionária. Os testes foram realizados utilizando, respectivamente, as rotinas *adf.test* e *pp.test*, do pacote *tseries* do R. Os resultados dos dois testes de raiz unitária foram inconclusivos quanto a estacionariedade da série temporal.



(a) Valor de Fechamento dos Ativos do Banco Bradesco SA.

(b) ACF

Figura 1: (a) Valor de Fechamento dos Ativos do Banco Bradesco SA, de 04/01/2010 a 27/06/2017. (b) ACF Amostral.

Fonte: Autores.

A seguir apresentamos os modelos ajustados a série temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores.

Modelo 1 - ARFIMA(p, d, q), com $\widehat{d} = 0.0458$, $p = 1$, onde $\widehat{\phi}_1 = 0.9890$ e $q = 0$. Para este modelo, obtivemos uma variância estimada dos resíduos igual a $\widehat{\sigma}^2 = 0.2104$, Critério de Informação de Akaike $AIC = 2375.877$ e $p - valor = 0.6571$ para o teste de Box - Pierce para os resíduos.

Modelo 2 - ARIMA(p, d, q), com $p = 0 = q$ e $d = 1$. Para este modelo, obtivemos uma variância estimada dos resíduos igual a $\widehat{\sigma}^2 = 0.2115$, Critério de Informação de Akaike $AIC = 2381.62$ e $p - valor = 0.4938$ para o teste de Box - Pierce para os resíduos.

Modelo 3 - ARMA(p, q), com $p = 1$, $q = 0$, onde $\widehat{\phi}_1 = 0.9790$. Para este modelo, obtivemos uma variância estimada dos resíduos igual a $\widehat{\sigma}^2 = 0.2125$, Critério de Informação de Akaike $AIC = 2348.86$ e $p - valor = 0.6587$ para o teste de Box - Pierce para os resíduos.

Modelo 4 - Modelo de Suavização Exponencial (Modelo Linear de Holt): as estimativas para os parâmetros do modelo são: $\widehat{\alpha} = 0.9856$, $\widehat{\beta} = 0.00564$. Para este modelo, obtivemos

uma variância estimada dos resíduos igual a $\widehat{\sigma}^2 = 0.212891$ e $p - valor = 0.9738$ para o teste de Box - Pierce para os resíduos.

As Tabelas 6 a 9 a seguir apresentam as medidas de acurácia dos para as previsões utilizando os Modelos 1 a 4 ajustados e a combinação de previsões combinadas dois a dois. Analisando tais tabelas, verificamos que o Modelo 2 apresenta menor ME, em valor absoluto. A combinação das previsões dos Modelos 2 e 3, por variância mínima, apresenta menor MAE e MAPE, com $\widehat{w} = 0.9110477$. A combinação das previsões dos Modelos 1 e 3, por regressão, apresenta menor RMSE, com $\widehat{\beta}_1 = -15.86$, $\widehat{\beta}_2 = 16.86$ e R^2 ajustado de 0.986. Por último a combinação de previsões dos modelos 3 e 4, por média, apresenta menor MPE, em valor absoluto.

Tabela 6: Medidas de Acurácia dos Modelos 1 a 4.

Modelo Ajustado	ME	RMSE	MAE	MPE	MAPE
Modelo 1	0.000953	0.458560	0.334767	-0.036906	1.485480
Modelo 2	**0.000014**	0.459645	0.334200	-0.020778	1.482779
Modelo 3	-0.001027	0.458546	0.334766	-0.045816	1.485664
Modelo 4	0.013157	0.461464	0.336099	0.047475	1.491103

Fonte: Autores.

Tabela 7: Medidas de Acurácia para as Combinações de Previsão: Variância Mínima.

Combinação de Previsão	ME	RMSE	MAE	MPE	MAPE
Modelos 1 e 2	0.000932	0.458547	0.334626	-0.035013	1.484809
Modelos 1 e 3	0.040178	0.460744	0.336245	0.139534	1.488609
Modelos 1 e 4	0.002594	0.458586	0.334513	-0.025046	1.484326
Modelos 2 e 3	0.000575	0.459439	**0.334170**	-0.020068	**1.482612**
Modelos 2 e 4	0.000432	0.459882	0.334482	-0.019219	1.484048
Modelos 3 e 4	0.000922	0.458579	0.334516	-0.032573	1.484486

Fonte: Autores.

Tabela 8: Medidas de Acurácia para as Combinações de Previsão: Média.

Combinação de Previsão	ME	RMSE	MAE	MPE	MAPE
Modelos 1 e 2	0.000843	0.458765	0.334233	-0.027228	1.482913
Modelos 1 e 3	-0.000037	0.458551	0.334765	-0.041361	1.485565
Modelos 1 e 4	0.006950	0.459091	0.334497	0.004856	1.483973
Modelos 2 e 3	-0.000148	0.458760	0.334228	-0.031683	1.482968
Modelos 2 e 4	0.006973	0.460303	0.334908	0.015062	1.485829
Modelos 3 e 4	0.005966	0.459076	0.334481	**0.000426**	1.483962

Fonte: Autores.

A Figura 2 apresenta as Predições e as Previsões da Série Temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores, utilizando o modelo e as combinações de previsão com menores medidas de acurácia. Observa-se que os modelos e combinação de previsões captam o comportamento dos dados e as previsões apresentadas possuem pouca variação entre si.

Tabela 9: Medidas de Acurácia para as Combinações de Previsão: Regressão.

Combinação de Previsão	ME	RMSE	MAE	MPE	MAPE
Modelos 1 e 2	0.000411	0.458546	0.334630	-0.037350	1.484872
Modelos 1 e 3	-0.000624	**0.458435**	0.334743	-0.046644	1.485753
Modelos 1 e 4	0.000971	0.458583	0.334521	-0.032412	1.484500
Modelos 2 e 3	0.000365	0.458533	0.334629	-0.037762	1.484887
Modelos 2 e 4	0.004310	0.459863	0.334531	-0.002468	1.484046
Modelos 3 e 4	0.000958	0.458579	0.334523	-0.032533	1.484513

Fonte: Autores.

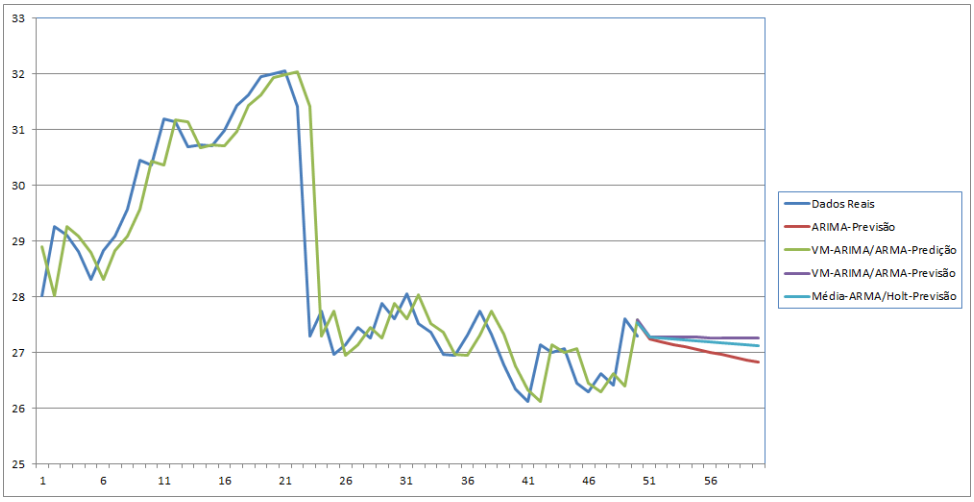


Figura 2: Previsão da Série Temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores.

Fonte: Autores.

5 Considerações Finais

Ao concluirmos nossas análises, podemos verificar que na maioria dos casos as menores medidas de acurácia, nas simulações de Monte Carlo, foram obtidas através de combinação de previsões. Cabe ressaltar que, em alguns casos as menores medidas médias foram obtidas quando a previsão foi feita somente com um modelo, e nestes casos, em apenas uma situação o modelo ARFIMA(p, d, q) obteve menores medidas de acurácia médias. Ou seja, mesmo tendo uma série gerada a partir deste processo, as combinações exerceram melhor papel preditivo do que o modelo propriamente dito. O mesmo pode ser observado na aplicação em dados reais, uma vez que quatro das cinco menores medidas de acurácia são obtidas a partir de combinação de previsões.

Desta forma, é possível concluir com base nas simulações de Monte Carlo e aplicação realizada neste artigo, que mesmo tendo um problema na especificação do modelo, como é o caso da aplicação utilizando a série temporal do valor do ativo do Banco Bradesco SA na hora do fechamento da bolsa de valores, podemos obter boas previsões.

Neste caso, a combinação de previsões pode ser uma ótima alternativa para aperfeiçoar a

previsão, uma vez que, como na maioria dos casos apresentados na seção de simulações de Monte Carlo deste trabalho, é possível aprimorar a capacidade preditiva do modelo visto que as mesmas apresentam menores medidas de acurácia quando comparadas com previsões geradas utilizando-se apenas um modelo.

Referências

Abraham, B. e Ledolter, J. (2009). *Statistical methods for forecasting*, volume 234. John Wiley & Sons.

Bates, J. M. e Granger, C. W. (1969). The combination of forecasts. *Or*, pages 451–468.

Bisognin, C. (2007). *Estimação e previsão em processos $SARFIMA(p, d, q) \times (P, D, Q)_s$ na presença de outlier. Tese de Doutorado. Universidade Federal do Rio grande do Sul. Programa de Pós-Graduação em Matemática. Porto Alegre.* PhD thesis.

Box, G. E. e Jenkins, G. M. (1976). Time series analysis, control, and forecasting. *San Francisco, CA: Holden Day*, 3226(3228):10.

Costantini, M. e Pappalardo, C. (2010). A hierarchical procedure for the combination of forecasts. *International journal of forecasting*, 26(4):725–743.

De Menezes, L. M., Bunn, D. W., e Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, 120(1):190–204.

Gardner, G., Harvey, A. C., e Phillips, G. D. (1980). Algorithm as 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of kalman filtering. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):311–322.

Geweke, J. e Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of time series analysis*, 4(4):221–238.

Granger, C. W. e Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of time series analysis*, 1(1):15–29.

Granger, C. W. e Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of forecasting*, 3(2):197–204.

Haslett, J. e Raftery, A. E. (1989). Space-time modelling with long-memory dependence: Assessing ireland's wind power resource. *Applied Statistics*, pages 1–50.

Hosking, J. R. (1981). Fractional differencing. *Biometrika*, 68(1):165–176.

Hosking, J. R. (1984). Modeling persistence in hydrological time series using fractional differencing. *Water resources research*, 20(12):1898–1908.

Khandakar, Y. e Hyndman, R. J. (2008). Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, 27(03).

Lima, R. C., Góis, M. R., e Ulises, C. (2007). Previsão de preços futuros de commodities agrícolas com diferenciações inteira e fracionária, e erros heteroscedásticos. *Brazilian Journal of Rural Economy and Sociology (RESR)*, 45(3).

Makridakis, S., Wheelwright, S. C., e Hyndman, R. J. (1998). *Forecasting methods and applications*. John wiley & sons.

Morettin, P. A. e Toloi, C. (2006). *Análise de séries temporais*. Blucher.

Pellegrini, F. R. (2000). Metodologia para implementação de sistemas de previsão de demanda. *Dissertaçãoo de Mestrado. Porto Alegre: UFRGS*.

Queiroz, F. F. (2016). Estudo sobre má especificação na família de posição e escala. *Monografia (Especialização) - Curso de Estatística. Natal: UFRN*.

Werner, L. (2005). *Um modelo composto para realizar previsão de demanda através da integração da combinação de previsões e do ajuste baseado na opinião. Tese de Doutorado. Universidade Federal do Rio grande do Sul. Programa de Pós-Graduação em Engenharia de Produção. Porto Alegre.* PhD thesis.

Avaliação do desempenho de Índices de Capacidade tradicionais diante de processos Não-Normais

Eduardo de Oliveira Correa [1]

Danilo Marcondes Filho[2]

Resumo: Índices de capacidade (IC) são amplamente usados para avaliar o desempenho dos processos industriais. Dado um processo operando sob condições estáveis e uma característica de qualidade representada por uma variável aleatória de interesse, os IC basicamente comparam a variabilidade natural dessa variável em relação à amplitude das especificações do processo. Quanto maior a variabilidade, menor a capacidade do processo em produzir unidades dentro das especificações. Destacam-se no meio industrial os IC clássicos Cp, Cpk, disponíveis em rotinas de controle de qualidade. Entretanto, estes índices avaliam a capacidade do processo supondo distribuição Normal para a variável aleatória sob investigação. Devido à complexidade de processos produtivos atuais, as características de qualidade geram dados com distribuições com caldas longas e/ou assimétricas, tornando a avaliação da capacidade destes processos através dos IC clássicos bastante distorcida. Embora exista literatura disponível sobre IC para processos "não-normais", verifica-se pouca aplicação de tais abordagens nas indústrias. Este trabalho replica e estende o estudo de Somerville e Montgomery (1996), apresentando uma avaliação do erro de análise da capacidade de processos utilizando os índices Cp e Cpk, diante de variáveis apresentando distribuições de probabilidade com diferentes formas.

Palavras-chave: *Índices de Capacidade, Distribuição Não-Normal*.

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *eduardo.correa@ufrgs.br*

[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marcondes.danilo@gmail.com*

1 Índices de Capacidade Clássicos

Os índices de capacidade são medidas adimensionais que quantificam a capacidade de um processo estável. São medidos através da relação entre a variabilidade natural do processo e a variabilidade que é permitida a esse processo, dada pelos limites de especificação da variável (característica de qualidade).

1.1 Índices de dados Normais (IC-N)

1.1.1 Índice C_p

$$C_p = \frac{LSE - LIE}{6\,\sigma},$$

onde:

LSE: Limite Superior Especificado

LIE: Limite Inferior Especificado

σ: Desvio Padrão do processo

O Índice C_p compara a variabilidade da variável em relação a amplitude da especificação sem avaliar a centralidade da distribuição Normal.

1.1.2 Índice C_{pk}

$$C_{pk} = \text{mín}\left(\frac{LSE - \mu}{3\,\sigma}; \frac{\mu - LIE}{3\,\sigma}\right),$$

onde:

μ: Média do processo

O índice C_{pk} avalia tanto a variabilidade quanto a centralidade da distribuição Normal.

1.2 Índices de Clements para dados Não-Normais (IC-NN)

1.2.1 $C_p^{'}$

$$C_p^{'} = \frac{LSE - LIE}{F_{\left(1-\frac{\alpha}{2}\right)} - F_{\left(\frac{\alpha}{2}\right)}},$$

onde :

$F_{\left(1-\frac{\alpha}{2}\right)}$ e $F_{\left(\frac{\alpha}{2}\right)}$: Percentis da distribuição de probabilidade considerada

1.2.2 C'_{pk}

$$C'_{pk} = \left(\frac{LSE - \mu}{F_{\left(1-\frac{\alpha}{2}\right)} - \mu} \; ; \; \frac{\mu - LIE}{\mu - F_{\left(\frac{\alpha}{2}\right)}} \right)$$

Neste estudo admitimos que a distribuição de probabilidade sob investigação possui média centrada, isto é, no ponto médio do limite de especificação. Dessa forma, $C_p = C_{pk}$ (dados Normais) e $C'_p = C'_{pk}$ (dados Não-Normais).

2 Estudo comparativo do Índice C_{pk} (IC-N) em distribuições de probabilidade Não-Normais (Simétricas)

Neste estudo apresentaremos os erros de avaliação de capacidade do processo via o uso de C_{pk} (IC-N) diante de diferentes em duas distribuições de probabilidade simétricas: A distribuição t-Student e a distribuição Beta com parâmetros iguais. A distribuição de t-Student apresenta caudas mais longas e a distribuição Beta apresenta caudas mais curtas em relação a distribuição Normal.

2.1 IC-N Variável t-Student

Considere como exemplo a variável aleatória X $\sim$ t (gl) , onde gl = 5. Temos então $\mu = 0$ e $\sigma = \sqrt{\frac{gl}{gl-2}} = \sqrt{\frac{5}{5-2}} = 1,2910.$

Sejam $F_N \rightarrow$ *Função acumulada distribuição Normal*, $F_t \rightarrow$ *Função acumulada distribuição t-Student* e $NNC \rightarrow$ *Número de não conformes em partes por milhão*.
Logo:

$C_{pk_{2\sigma}}$ = mín$\left(\frac{F_{N_{0,9772}} - \mu}{F_{t_{0,9772}} - \mu}; \frac{\mu - F_{N_{0,0228}}}{\mu - F_{t_{0,0228}}} \right)$0,67 = mín$\left(\frac{2,5820 - 0}{2,6486 - 0}; \frac{0 - (-2,5820)}{0 - (-2,6486)} \right)$0,67 =

$C_{pk_{2\sigma}}$ **= mín(0,6531 ; 0,6531)** **e** **NNC = 24656,54**

$C_{pk_{3\sigma}}$ = mín$\left(\frac{F_{N_{0,9987}} - \mu}{F_{t_{0,9987}} - \mu}; \frac{\mu - F_{N_{0,0013}}}{\mu - F_{t_{0,0013}}} \right)$1,00 = mín$\left(\frac{3,8730 - 0}{5,5071 - 0}; \frac{0 - (-3,87300)}{0 - (-5,5071)} \right)$1,00 =

$C_{pk_{3\sigma}}$ **= mín(0,7032 ; 0,7032)** **e** **NNC = 5862,41**

$C_{pk_{4\sigma}}$ = mín$\left(\frac{F_{N_{0,99997}} - \mu}{F_{t_{0,99997}} - \mu}; \frac{\mu - F_{N_{0,00003}}}{\mu - F_{t_{0,00003}}} \right)$1,33 = mín$\left(\frac{5,1640 - 0}{12,2814 - 0}; \frac{0 - (-5,1640)}{0 - (-12,2814)} \right)$1,33 =

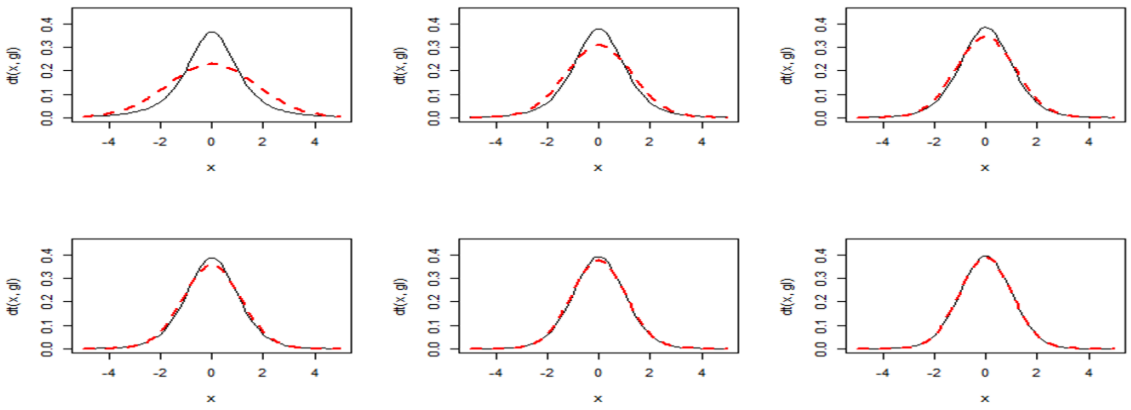
$C_{pk_{4\sigma}}$ **= = mín(0,5592 ; 0,5592) e NNC = 1786,41**

A tabela 1 apresenta os resultados do C_{pk} (IC-N) para distribuição t-Student com diversos valores de gl. Observa-se de uma maneira geral que a capacidade do processo é subestimada, visto que a distribuição t tem caudas mais longas do que a distribuição Normal. Este erro de estimativa é atenuado cada vez mais na medida em que aumentamos o parâmetro gl, o que é esperado, pois as caudas tornam-se cada vez mais parecidas em relação a distribuição Normal (ver figura 1).

Tabela 1: Cpk^r (Cpk real), NNC^r (número real de não-conformes, em partes por milhão), Cpk e NNC (supondo distribuição Normal para variável aleatória com distribuição t-student).

t-Student (gl)													
		gl = 3				**gl = 5**				**gl = 8**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	0,7018	0,7018	20259,66	20259,66	0,6531	0,6531	24656,54	24656,54	0,6538	0,6538	24867,78	24867,78
1,00	**1350**	0,5636	0,5636	6923,42	6923,42	0,7032	0,7032	5862,41	5862,41	0,8100	0,8100	4258,13	4258,13
1,33	**32,00**	0,2825	0,2825	6923,42	6923,42	0,5592	0,5592	1786,41	1786,41	0,8088	0,8088	856,44	856,44
		gl = 10				**gl = 30**				**gl = 50**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	0,6538	0,6538	24666,1	24666,1	0,6621	0,6621	23913,74	23913,74	0,6667	0,6667	23261,13	23261,13
1,00	**1350**	0,8476	0,8476	3657,31	3657,31	0,9240	0,9240	2449,99	2449,99	0,9698	0,9698	1767,7	1767,7
1,33	**32,00**	0,9057	0,9057	596,73	596,73	1,1136	1,1136	212,08	212,08	1,2433	1,2433	80,27	80,27

Figura 1: Distribuição t-Student (linha sólida) e distribuição suposta Normal (linha pontilhada).



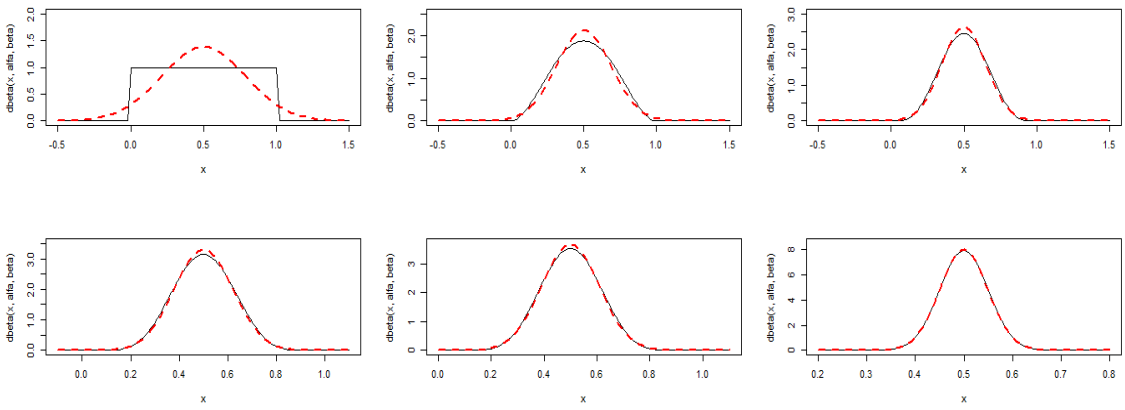
2.2 IC-N Variável Beta

A distribuição Beta com parâmetros α e β iguais é simétrica, assim como a distriuiçcao t-student. Entretanto, observamos na tabela 2 que o Cpk (IC-N) superestima a capacidade do processo, sendo que o erro de estimativa diminui na medida que aumenta o valor do parâmetro, visto que as caudas mais curtas da Distribuição Beta começam a se alongar e se assemelhar as caudas da distribuição Normal (ver figura 2).

Tabela 2: Cpk^r (Cpk real), NNC^r (número real de não-conformes, em partes por milhão), Cpk e NNC (supondo distribuição Normal para variável aleatória com distribuição Beta).

Beta (α ; β) , $\alpha = \beta$												
		(1 ; 1)				**(3 ; 3)**				**(5 ; 5)**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	0,8105	0,8105	0	0	0,7067	0,7067	15009,87	15009,87	0,6901	0,6901	18964,49	18964,49
1,00	**1350**	1,7367	1,7367	0	0	1,2674	1,2674	0	0	1,1579	1,1579	26,55	26,55
1,33	**32,00**	3,0716	3,0716	0	0	2,0720	2,070	0	0	1,7803	1,7803	0	0
		(8 ; 8)				**(10 ; 10)**				**(50 ; 50)**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	0,6818	0,6818	20655,17	20655,17	0,6792	0,6792	21142,79	21142,79	0,6717	0,6717	22470,29	22470,29
1,00	**1350**	1,0971	1,0971	310,33	310,33	1,0772	1,0772	473,34	473,34	1,0150	1,0150	1153,92	1153,92
1,33	**32,00**	1,6095	1,6095	0	0	1,5526	1,5526	0,06	0,06	1,3735	1,3735	17,16	17,16

Figura 2: Distribuição Beta (linha sólida) e distribuição suposta Normal (linha pontilhada).



3 Estudo comparativo do Índice C_{pk} (IC-N) em distribuições de probabilidade Não-Normais (Assimétricas)

3.1 IC-N Variável Beta assimétrica à direita

Considere como exemplo a variável aleatória X $\sim$ B $(\alpha;\beta)$, onde $\alpha = 1; \beta = 5$. Temos então $\mu = \dfrac{\alpha}{\alpha+\beta} = 0,1667$ e $\sigma = \sqrt{\dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}} = 0,1419$

Sejam $F_N \rightarrow$ *Função acumulada distribuição Normal*, $F_B \rightarrow$ *Função acumulada distribuição Beta*

e $NNC \rightarrow$ *Número de não conformes em partes por milhão*.

Logo:

$$C_{pk_{2\sigma}} = \text{mín}\left(\frac{F_{N_{0,9772}} - \mu}{F_{B_{0,9772}} - \mu}; \frac{\mu - F_{N_{0,0228}}}{\mu - F_{B_{0,0228}}}\right) 0{,}67 = \text{mín}\left(\frac{0,4484 - 0,1667}{2,6486 - 0,1667}; \frac{0,1667 + 0,1150}{0,1667 - 0,0046}\right) 0{,}67 =$$

$C_{pk_{2\sigma}}$ **= mín(1,1643 ; 0,5184)** **e** **NNC = (0 ; 51071,82)**

$$C_{pk_{3\sigma}} = \text{mín}\left(\frac{F_{N_{0,9987}} - \mu}{F_{B_{0,9987}} - \mu}; \frac{\mu - F_{N_{0,0013}}}{\mu - F_{B_{0,0013}}}\right) 1{,}00 = \text{mín}\left(\frac{0,5892 - 0,1667}{0,7333 - 0,1667}; \frac{0,1667 + 0,2559}{0,1667 - 0,0003)}\right) 1{,}00 =$$

$C_{pk_{3\sigma}}$ **= mín(2,5396 ; 0,7458)** **e** **NNC = (0 ; 11692,86)**

$$C_{pk_{4\sigma}} = \text{mín}\left(\frac{F_{N_{0,99997}} - \mu}{F_{B_{0,99997}} - \mu}; \frac{\mu - F_{N_{0,00003}}}{\mu - F_{B_{0,00003}}}\right) 1{,}33 = \text{mín}\left(\frac{0,7301 - 0,1667}{8741 - 0,1667}; \frac{0,1667 + 0,3968}{0,1667 - 0,0000}\right) 1{,}33 =$$

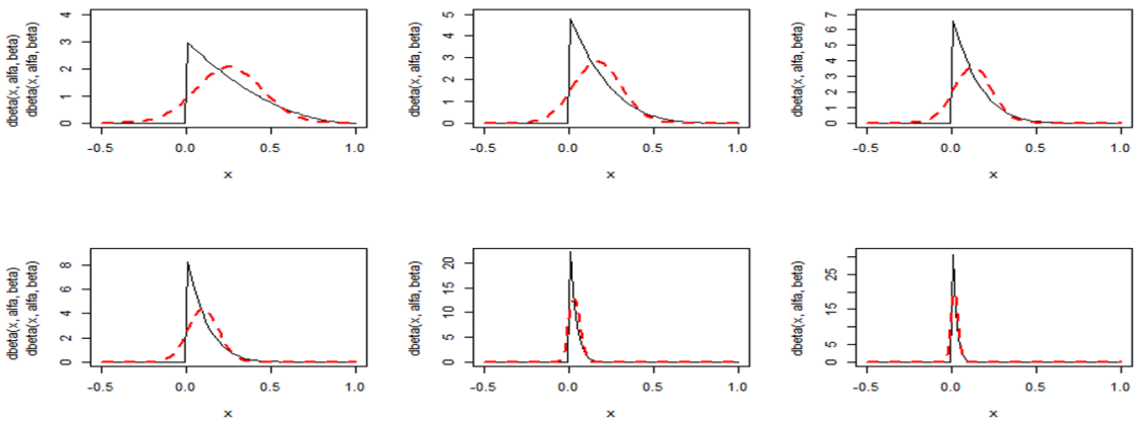
$C_{pk_{4\sigma}}$ **= mín(4,4964 ;1,0593)** **e** **NNC = (0 ; 1432,16)**

A tabela 3 mostra os valores de C_{pk} (IC-NN) para distribuição Beta com $\alpha = 1$ e diferentes valors do parâmetro β . Esta configuração de parâmetros gera distribuições Beta com caudas esquerdas mais curtas e direitas mais longas em relação a distribuição Normal (ver figura 3), desta forma, o C_{pk} (IC-NN) superestima a capacidade do processo à esquerda e subestima à direita.

Tabela 3: Cpk^r (Cpk real), NNC^r (número real de não-conformes, em partes por milhão), Cpk e NNC (supondo distribuição Normal para variável aleatória com distribuição Beta).

Beta (α ; β **)** , $\alpha = 1$													
		(1 ; 3)				**(1 ; 5)**				**(1 ; 7)**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	1,0707	0,5561	0	47714,31	1,1646	0,5184	0	51071,82	1,2136	0,5050	0	51459,29
1,00	**1350**	2,328	0,9085	0	4831,31	2,5396	0,7458	0	11692,86	2,6498	0,6806	0	14150,2
1,33	**32,00**	4,121	1,4341	0	0	4,4964	1,0593	0	1432,16	4,6920	0,9059	0	2902,14
		(1 ; 9)				**(1 ; 30)**				**(1 ; 50)**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	1,2438	0,4984	0	51412,25	1,3290	0,4854	0	50513,57	1,3455	0,4837	0	50247,69
1,00	**1350**	2,7176	0,6459	0	15332,63	2,9087	0,5664	0	17617,73	2,9458	0,5536	0	17917,75
1,33	**32,00**	4,8122	0,8244	0	3787,86	5,1512	0,6396	0	5913,49	5,2168	0,6105	0	6252,02

Figura 3: Distribuição Beta (linha sólida) e distribuição suposta Normal (linha pontilhada).



3.2 IC-N Variável Beta assimétrica à esquerda

Considere como exemplo a variável aleatória X $\sim$ B $(\alpha;\beta)$, onde $\alpha = 5; \beta = 1$. Temos então
$\mu = \dfrac{\alpha}{\alpha+\beta} = 0,8333$ e $\sigma = \sqrt{\dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}} = 0,1419$

Sejam $F_N \rightarrow$ *Função acumulada distribuição Normal*, $F_B \rightarrow$ *Função acumulada distribuição Beta* e $NNC \rightarrow$ *Número de não conformes em partes por milhão*.
Logo:

$$C_{pk_{2\sigma}} = \text{mín}\left(\frac{F_{N_{0,9772}} - \mu}{F_{B_{0,9772}} - \mu}; \frac{\mu - F_{N_{0,0228}}}{\mu - F_{B_{0,0228}}}\right) 0{,}67 = \text{mín}\left(\frac{1,1151 - 0,8333}{0,9954 - 0,8333}; \frac{0,8333 - 0,5515}{0,8333 - 0,4692}\right) 0{,}67 =$$

$C_{pk_{2\sigma}}$ **= mín(0,5184 ; 1,1646)** **e** **NNC = (51071,82 ; 0)**

$$C_{pk_{3\sigma}} = \text{mín}\left(\frac{F_{N_{0,9987}} - \mu}{F_{B_{0,9987}} - \mu}; \frac{\mu - F_{N_{0,0013}}}{\mu - F_{B_{0,0013}}}\right) 1{,}00 = \text{mín}\left(\frac{1,2560 - 0,8333}{0,9997 - 0,8333}; \frac{0,8333 - 0,4106}{0,8333 - 0,2667)}\right) 1{,}00 =$$

$C_{pk_{3\sigma}}$ **= mín(0,7458 ; 2,5396)** **e** **NNC = (11692,86 ; 0)**

$$C_{pk_{4\sigma}} = \text{mín}\left(\frac{F_{N_{0,99997}} - \mu}{F_{B_{0,99997}} - \mu}; \frac{\mu - F_{N_{0,00003}}}{\mu - F_{B_{0,00003}}}\right) 1{,}33 = \text{mín}\left(\frac{1,3969 - 0,8333}{0,9999 - 0,8333}; \frac{0,8333 - 0,2697}{0,8333 - 0,1259}\right) 1{,}33 =$$

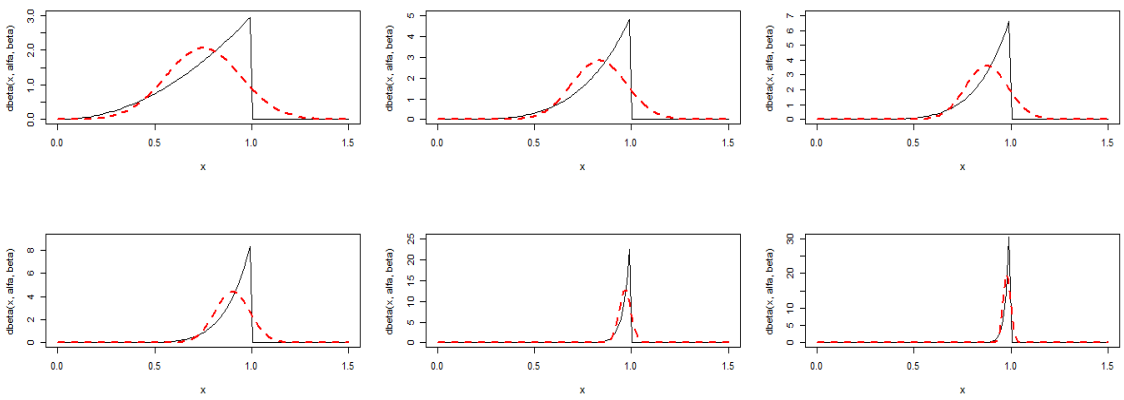
$C_{pk_{4\sigma}}$ **= mín(1,0593 ; 4,4964)** **e** **NNC = (1432,16 ; 0)**

A tabela 4 mostra os valores de C_{pk} (IC-N) para distribuição Beta com $\beta = 1$ e diferentes valores do parâmetro α . Esta configuração de parâmetros gera distribuições Beta com caudas esquerdas mais longas e direitas mais curtas em relação a distribuição Normal (ver figura 4), desta forma, o C_{pk} (IC-N) subestima a capacidade do processo à esquerda e superestima à direita.

Tabela 4: Cpk^r (Cpk real), NNC^r (número real de não-conformes, em partes por milhão), Cpk e NNC (supondo distribuição Normal para variável aleatória com distribuição Beta).

Beta (α ; β) , $\beta = 1$														
		(3 ; 1)				**(5 ; 1)**				**(7 ; 1)**				
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC		
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	
0,67	**22750**	0,5561	1,0707	47714,31	0	0,5184	1,1646	51071,82	0	0,5050	1,2136	51459,29	0	
1,00	**1350**	0,9085	2,328	4831,31	0	0,7458	2,5396	11692,86	0	0,6806	2,6498	14150,2	0	
1,33	**32,00**	1,4341	4,121	0	0	1,0593	4,4964	1432,16	0	0,9059	4,6920	2902,14	0	
		(9 ; 1)				**(30 ; 1)**				**(50 ; 1)**				
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC		
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	
0,67	**22750**	0,4984	1,2438	51412,25	0	0,49	1,329	50513,57	0	0,4837	1,3455	50247,7	0	
1,00	**1350**	0,6459	2,7176	15332,63	0	0,57	2,9087	17617,73	0	0,5536	2,9458	17917,75	0	
1,33	**32,00**	0,8244	4,8122	3787,86	0	0,6396	5,1512	5913,49	0	0,6105	5,2168	6252,02	0	

Figura 4: Distribuição Beta (linha sólida) e distribuição suposta Normal (linha pontilhada).



3.3 IC-N Variável Qui-Quadrado e Gama

As distribuições Qui-Quadrado e Gama são assimétricas à direita, conforme ilustrado nas figuras 5 e 6, respectivamente. As tabelas 5 e 6 apresentam, respectivamente, os resultados do C_{pk} (IC-N) para tais distribuições. Investigamos a distribuição Qui-Quadrado com diversos valores do parâmetro gl e a distribuição Gama fixando o valor do parâmetro $\beta = 1$ e variando os valores do parâmetro α. De maneira análoga ao resultado mostrado na seção 3.2, para ambas as distribuições observamos que o C_{pk} (IC-N) superestima a capacidade à esquerda e subestima a capacidade à direita.

Tabela 5: Cpk^r (Cpk real), NNC^r (número real de não-conformes, em partes por milhão), Cpk e NNC (supondo distribuição Normal para variável aleatória com distribuição Qui-Quadrado).

Qui-Quadrado (gl)													
		gl = 3				**gl = 5**				**gl = 8**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	1,1731	0,5007	0	48146,29	1,0081	0,5252	0	45311,49	0,9346	0,5407	0	42380,11
1,00	**1350**	2,4739	0,5818	0	15824,85	1,9921	0,6400	0	12795,52	1,7694	0,6770	0	10336,05
1,33	**32,00**	4,3472	0,6354	0	5094,54	3,3997	0,7217	0	3419,89	2,9333	0,7781	0	2291,79
		gl = 10				**gl = 20**				**gl = 50**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	0,8921	0,5518	220,61	40976,25	0,7735	0,5966	4634,71	36854,11	0,7469	0,6111	11164,78	32374,11
1,00	**1350**	1,6404	0,7034	0	9309,63	1,2877	0,8120	0	6717,83	1,2113	0,8480	46,95	4482,66
1,33	**32,00**	2,6568	0,8193	0	1880,92	1,9022	0,9959	0	1005,19	1,7453	1,0571	0	449,25

Figura 5: Distribuição Qui-Quadrado (linha sólida) e distribuição suposta Normal (linha pontilhada).

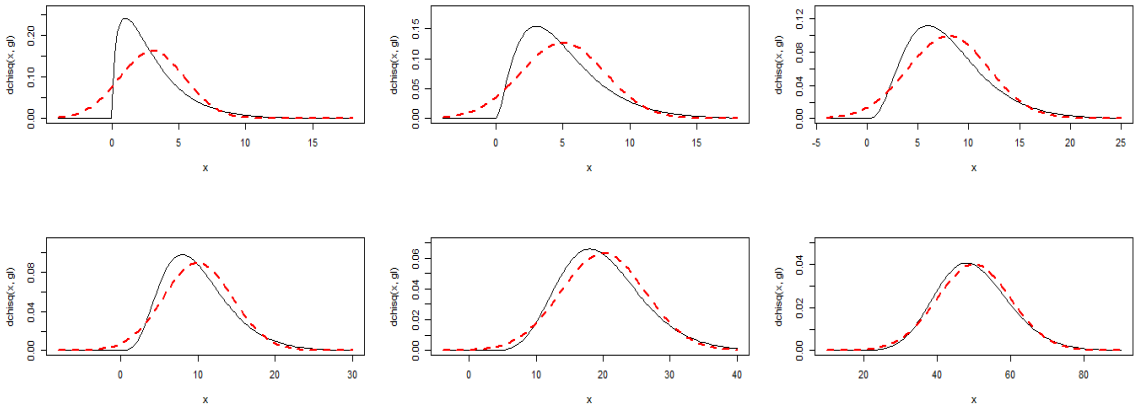
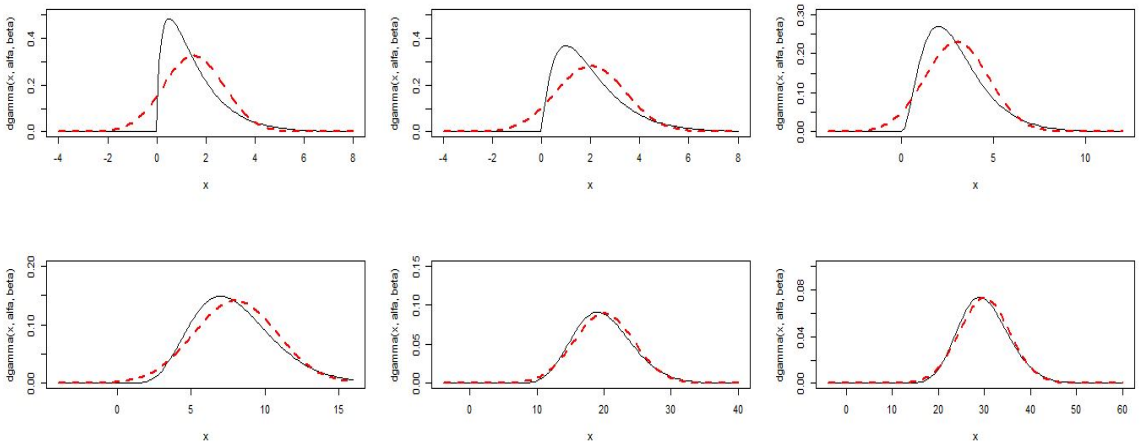


Tabela 6: Cpk^r (Cpk real), NNC^r (número real de não-conformes, em partes por milhão), Cpk e NNC (supondo distribuição Normal para variável aleatória com distribuição Gama).

Gama (α ; β) , $\beta = 1$													
		(1,5 ; 1)				**(2 ; 1)**				**(3 ; 1)**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	1,1731	0,5007	0	48146,29	1,0707	0,5145	0	46622,13	0,9655	0,5337	0	44190,36
1,00	**1350**	2,4739	0,5818	0	15824,85	2,1789	0,6148	0	14084,86	1,8635	0,6602	0	11796,25
1,33	**32,00**	4,3472	0,6354	0	5094,54	3,7768	0,6840	0	4092,89	3,1324	0,7524	0	2937,14
		(8 ; 1)				**(20 ; 1)**				**(30 ; 1)**			
Cpk^r	NNC^r	Cpk		NNC		Cpk		NNC		Cpk		NNC	
		Esquerda	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**	**Esquerda**	**Direita**
0,67	**22750**	0,8222	0,5751	2894,21	38125,28	0,7574	0,6051	9766,98	33349,5	0,7393	0,6156	12204,2	31632,3
1,00	**1350**	1,4303	0,7593	0	7456,84	1,2414	0,8331	19,21	4923,3	1,1899	0,8594	79,84	4163,75
1,33	**32,00**	2,2042	0,9087	0	1230,64	1,8061	1,0314	0	543,19	1,7009	1,0768	0	386,3

Figura 6: Distribuição Gama (linha sólida) e distribuição suposta Normal (linha pontilhada).



4 Considerações Finais

Este trabalho replicou e estendeu o estudo de Sommerville e Montgomery (1996), quantificando os erros de avaliação da capacidades de processos utilizando os Índices C_p e C_{pk} quando a variável envolvida não possui distribuição Normal. As distorções de tais índices foram evidenciadas considerando processos com varáveis apresentando distribuições t-Student, Qui-Quadrado, Gamma e Beta, configurando uma ampla variedade de formas distintas da distribuição Normal.

Referências

[1] MONTGOMERY, D. C. Introdução ao Controle Estatístico da Qualidade. 4 ed. Rio de Janeiro: *LTC ? Livros Técnicos e Científicos Editora S.A. 2004.*

[2] WERNER, Liane; GONÇALEZ, Patrícia U. Comparação dos índices de capacidade do processo para distribuições não-normais, Gest. Prod., São Carlos, v. 16, n. 1, p. 121-132, jan.-mar. 2009;

[3] SOMERVILLE, S. E., MONTGOMERY, D. C. (1996). Process capability indices and non-normal distributions. Quality Engineering, 9(2), 305-316.

Perfil dos participantes em crimes de violência doméstica, no Rio Grande do Sul (lei nº 11.340 - Lei Maria da Penha)

GRILLO, Helena Simeonidis [1]

ZIEGELMANN, Patrícia Klarmann [2]

Resumo: Este estudo tem como objetivo apresentar o perfil dos participantes de crimes de feminicídio tentados e consumados no estado do Rio Grande do Sul de modo a auxiliar aos órgãos de segurança pública a responder à questão sobre a possibilidade de prevenção a este tipo de violência. A criação da Lei Maria da Penha (Lei n° 11.340/2006) criou mecanismos para coibir e prevenir a violência doméstica e familiar contra a mulher, permitindo que informações sejam coletadas, através dos registros de ocorrências, e estudos realizados. Para a análise foi utilizada a estatística descritiva. O resultado do estudo mostrou o perfil de um crime que acontece à noite, na residência da vítima, através do disparo de arma de fogo, no caso da morte, ou de uso de arma branca, no caso da tentativa, realizado por um homem branco, contra uma mulher branca, ambos com idade entre 18 e 24 anos, com pouca instrução, sem filhos, sem antecedentes registrados, devido ao fim o relacionamento, em sua maioria, quando o crime é consumado, ele é capturado pelos órgãos de segurança.

Palavras-chave: *Violência Doméstica; Perfil criminal; Lei Maria da Penha.*

Introdução

Para os efeitos da Lei Maria da Penha (BRASIL, 2006), configura violência doméstica e familiar contra a mulher qualquer ação ou omissão baseada no gênero que lhe cause morte, lesão, sofrimento físico, sexual ou psicológico e dano moral ou patrimonial. Está presente no mundo todo, motivando crimes hediondos e graves violações de direitos humanos.

Maria da Penha Maia Fernandes, em 1983, foi alvo de duas tentativas de homicídio por parte de seu marido e acabou ficando paraplégica. Foram mais de 20 anos de luta, para que seu agressor fosse condenado. O caso de Maria da Penha Maia Fernandes se tornou um marco e, motivou a criação da lei que trata da violência familiar e doméstica contra as mulheres, em 2006, popularmente chamada de Lei Maria da Penha. Esta lei tem por objetivo erradicar ou minimizar a violência familiar e

---

[1]Graduanda do Curso de Estatística da Universidade Federal do Rio Grande do Sul, hsgrillo@hotmail.com
[2]Professor orientador: Professora Doutora Patrícia Klarmann Ziegelmann, Universidade Federal do Rio Grande do Sul, patrícia.ziegelmann@ufrgs.br
Porto Alegre – RS, outubro de 2017.

doméstica contra as mulheres, e define em seu artigo 5º, que a violência doméstica e familiar ocorre no âmbito da unidade doméstica, no âmbito da família ou em qualquer relação íntima de afeto, e em seu artigo 7º cita cinco formas de violência doméstica e familiar, a violência física, a violência psicológica, a violência sexual, a violência patrimonial e a violência moral. As relações pessoais enunciadas neste artigo independem de orientação sexual. (BRASIL, 2006).

Como esse tipo de violência é de difícil acesso e controle e, com o déficit de efetivos policiais, viaturas e equipamentos, conhecer o perfil do agressor e de sua vítima, assim como as situações em que os crimes acontecem pode ajudar no enfrentamento e prevenção a estes crimes.

Este artigo então tem por objetivo caracterizar, através de seus perfis, os detalhes dos crimes de feminicídio, que é a morte de mulheres, com recorte de gênero, resultante de violência doméstica, as vítimas e seus agressores, possibilitando ações de segurança pública, a fim de minimizar as consequências desta forma de violência, atuando na prevenção destes crimes.

Métodos

Estudo transversal realizado com dados extraídos do "Sistema Integrado de Dados - Consultas Integradas" da Secretaria da Segurança Pública do Estado do Rio Grande do Sul. Através deste sistema são obtidas informações das ocorrências policiais registradas, que foram organizadas em dois grandes bancos, a saber, feminicídios consumados (morte de uma mulher por razões de sua condição feminina, ou seja, quando o crime envolver violência doméstica) e tentados (tentativas de morte destas mulheres). (Código Penal. art. 121, § 2º, VI), conforme as informações pertinentes disponíveis. Neste estudo, serão utilizadas apenas as ocorrências que envolvem uma única vítima e um único agressor. O banco dos feminicídios tentados abrange 1353 observações no período entre 2012 e Julho de 2017. O banco dos feminicídios consumados abrange 802 observações registradas no período entre Agosto de 2006 e Julho de 2017. As informações contidas nos bancos de dados estão descritas na Tabela 1.

Tabela 1 - Informações contidas no banco de dados

Informações	FATO	VÍTIMA	AUTOR
Ano	X	X	X
Id	X	X	X
Data fato	X	X	X
Dia semana	X		
Mês	X		
Horário	X		
Turno	X		
Local	X		
Instrumento	X		
Motivo	X		
Sob efeito alucinógeno	X		
Idade		X	X
Sexo			X

Cor	X	X
Escolaridade	X	X
Relação da vítima com o agressor	X	
Filhos com o agressor	X	
Possui antecedentes registrados	X	X
Agressões prévias registradas	X	X
Ameaça/Quantidade	X	X
Lesão corporal/Quantidade	X	X
Crime 1/Quantidade	X	X
Crime 2/Quantidade	X	X
Última agressão registrada	X	
Data da última agressão	X	
Tempo entre a última agressão e o homicídio (dias)	X	
Possui antecedentes registrados com outro autor	X	
Possui antecedentes registrados com outra vítima		X
Agressões prévias registradas	X	X
Status prisional na época		X
Suicídio		X

Fonte: SSP/RS

Análise Estatística: os dados são apresentados através de frequências absolutas e relativas. Os métodos de Estatística Descritiva ajudam a organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos de dados. As análises foram realizadas utilizando o programa computacional SPSS 20.0, licenciado na Secretaria da Segurança Pública/RS. Sobre a categorização dos dados sobre a relação da vítima com o autor, na Lei Maria da Penha (Brasil, Lei N 11.340,2006) que define as relações entre os participantes, temos que o relacionamento atual refere-se às esposos, namorados, noivos, companheiros, ficantes, amantes; o relacionamento anterior cita ex-esposos, ex-companheiros, ex-namorados, ex-sogros, ex-cunhados , ex-genros; o relacionamento familiar inclui mãe, pai, filho(a), avó, madrasta, padrasto, irmão(ã), sogro(a), sobrinho(a), primo(a), cunhado(a), enteado(a), marido da sobrinha, e ainda a categoria outros relacionamentos trás companheira do ex-companheiro, amante do companheiro, namorada do caso.

Resultados

Feminicídio Consumado: O número total de ocorrências, por ano, é apresentado na Figura 1.

Figura 1 – Total de ocorrências de feminicídio consumado por ano **(**Fonte: SSP/RS)

Caracterização do fato: os crimes ocorrem mais no período noturno (30,80%), seguidos de manhã (26,48%), tarde (25,51%) e madrugada (17,29%), as mulheres sofrem a violência na residência (72,32%) onde mais acontecem os crimes, e outros 12,59% ocorrem em via pública, sendo atingidas com armas de fogo (38,40%) e arma branca, por exemplo facas, espetos, facões, canivetes, etc. (38,09%). Toda essa violência tem como um dos motivos principais a separação entre os casais (14,71%) e as brigas, desentendimentos e vinganças (12,84%), porém na maioria das ocorrências (65,59%) a motivação não foi identificada. (Tabela 2)

Tabela 2 - Características do fato - Feminicídio Consumado - RS

Características do fato	Feminicídio Consumado (n=802)
Turno	
Noite	247(30,80)
Manhã	219(26,48)
Tarde	211(25,51)
Madrugada	143(17,29)
Local	
Residência	580(72,32)
Via pública	101(12,59)
Ni	52(6,48)
Outros	69(8,60)
Instrumento	
Arma branca	312(38,90)
Arma de fogo	308(38,40)
Força Física/Uso das Mãos	70(8,73)
Outros	112(13,97)
Motivação	
Ni	526(65,59)
Separação	118(14,71)
Briga/Desentendimento/Vingança	103(12,84)
Outros	55(6,86)

Dados são apresentados por totais (percentuais)

Caracterização da vítima: a vítima de crime de feminicídio consumado tem idade entre 18 e 29 anos (32,92%), com ensino fundamental como nível de instrução (53,12%), auto declaradas brancas (85,91%), não possuem filhos com seu agressor (34,29%), porém em 37,28% dos casos não foi possível identificar se a vítima possui filhos com o autor, que é uma pessoa de seu relacionamento atual (52,87%). Ainda, não possuíam antecedentes registrados com este autor (57,6%) ou com outro autor (67,71%). (Tabela 3)

Tabela 3 - Características da vítima - Feminicídio Consumado - RS

Características da vítima	Feminicídio Consumado (n=802)
Escolaridade	
Ensino Fundamental	426(53,12)
Ni	182(22.69)
Ensino Médio	123(15,34)
Outros	71(8,85)

Raça/Cor	
Branca	689(85,91)
Negra	81(10,09)
Mulata	22(2,74)
Outros	10(1,25)
Idade (anos)	
0 - 12	14(1,74)
13 - 17	40(4,99)
18 - 24	155(19,33)
25 - 29	109(13,59)
30 - 34	106(13,22)
35 - 39	97(12,09)
40 - 44	74(9,23)
45 - 49	64(7,98)
50 - 54	35(4,36)
55 - 59	36(4,49)
60 - 64	23(2,87)
65 - 69	18(2,24)
70 - 79	20(2,49)
> 80	11(1,37)
Relação com o agressor	
Relacionamento Atual	424(52,87)
Relacionamento Anterior	260(32,42)
Relacionamento Familiar	100(12,47)
Outro Relacionamento	18(2,24)
Filhos com o agressor	
Ni	299(37,28)
Não	275(34,29)
Sim	228(28,43)
Antecedentes registrados com o agressor	
Não	462(57,61)
Sim	339(42,27)
Ni	1(0,12)
Antecedentes registrados com outro agressor	
Não	543(67,71)
Sim	250(31,17)
Ni	9(1,12)

Dados são apresentados por totais (percentuais)

Caracterização do agressor: o autor de violência doméstica pode ser do sexo masculino ou feminino, conforme a Lei Maria da Penha, neste estudo 97,38% são homens e 2,62% são mulheres, a idade dos autores homens varia entre 18 e 34 anos (42,13%), e das mulheres autoras entre 25 e 34 anos (38,09%), assim como as vítimas, 57,11% dos homens e 42,86% das mulheres possuem nível de instrução baixo e são auto declarados brancos 84,25% dos autores e 85,71% das autoras. Foram recolhidos pelas instituições da Segurança Pública 46,99% dos homens autores e 33,33% das mulheres autoras, e 21,90% dos agressores e 14,29% das agressoras morreram após o delito, sendo que destes homens 21,51% e 14,29% das mulheres cometeram suicídio. Também, 67,22% dos autores e 85,71% das autoras não possuíam ocorrências registradas com outras vítimas. (Tabela 4).

Tabela 4 - Características do agressor - Feminicídio Consumado - RS

Características do agressor	Feminicídio Consumado (n=802)	
	Masculino (n=781(97,38))	**Feminino (n=21(2,62))**
Escolaridade		
Ensino Fundamental	446(57,11)	9(42,86)
Ni	156(19.45)	7(0.87)
Ensino Médio	119(14,94)	2(0,25)
Outros	60(7,48)	3(0,37)
Raça/Cor		
Branca	658(84,25)	18(85,71)
Negra	89(11,40)	2(9,52)
Mulata	22(2,82)	1(4,76)
Outros	12(6,17)	0(0,0)
Idade(anos)		
13 - 17	10(1,28)	2(9,52)
18 - 24	112(14,34)	7(33,33)
25 - 29	108(13,83)	2(9,52)
30 - 34	109(13,96)	6(28,57)
35 - 39	113(14,47)	0(0,0)
40 - 44	80(10,24)	2(9,52)
45 - 49	74(9,48)	0(0,0)
50 - 54	61(7,81)	0(0,0)
55 - 59	43(5,51)	1(4,76)
60 - 64	24(3,07)	0(0,0)
65 - 69	23(2,94)	0(0,0)
70 - 79	15(1,92)	1(4,76)
> 80	5(0,64)	0(0,0)
Ni	4(0,51)	0(0,0)
Status Policial na época do crime		
Recolhido	367(46,99)	7(33,33)
Liberdade	198(25,35)	8(38,10)
Morto	171(21,90)	3(14,29)
Outros	45(5,76)	3(14,29)
Cometeu suicídio após o crime		
Não	604(77,34)	18(85,71)
Sim	168(21,51)	3(14,29)
Ni	9(1,15)	0(0,0)
Antecedentes registrados com outra vítima		
Não	525(67,22)	18(85,71)
Sim	247(31,63)	3(14,29)
Ni	9(1,15)	0(0,0)

Dados são apresentados por totais (percentuais)

Feminicídio Tentado

O gráfico da distribuição dos crimes por ano, Figura 2, mostra uma tendência de queda, principalmente pelas ações de prevenção contra os crimes de violência doméstica.

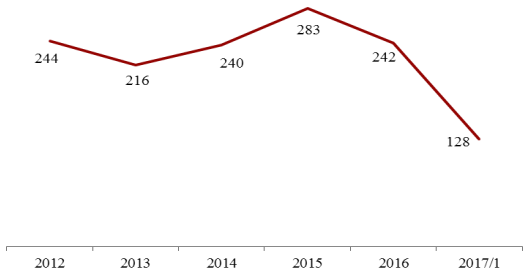


Figura 2 – Total de ocorrências de feminicídio tentado por ano **(**Fonte: SSP/RS)

Sobre o fato: os crimes ocorrem mais no período noturno (38,29%), seguidos de madrugada (22,39%), tarde (21,51%) e manhã (17,81%), as mulheres sofrem a violência na residência (69,77%) onde mais acontecem os crimes, e outros 19,22% ocorrem em via pública, sendo atingidas por arma branca (45,68%) e armas de fogo (23,87%). A principal causa dessa violência é a separação entre os casais (15,45%).(Tabela 5)

Tabela 5 - Características do fato - Feminicídio Tentado - RS

Características do fato	**Feminicídio Tentado (n=1353)**
Turno	
Noite	518(38,29)
Madrugada	303(22,39)
Tarde	291(21,51)
Manhã	241(17,81)
Local	
Residência	944(69,77)
Via pública	260(19,22)
Ni	74(5,47)
Outros	75(5,54)
Instrumento	
Arma branca	618(45,68)
Arma de fogo	323(23,87)
Força Física/Uso das Mãos	172(12,71)
Outros	240(17,74)
Motivação	
NI	812(60,01)
Separação	209(15,45)
Briga/Desentendimento/Vingança	159(11,75)
Outros	173(12,79)

Dados são apresentados por totais (percentuais)

Sobre a vítima: A vítima tem idade entre 18 e 34 anos (21,06%), são brancas (80,19%), em sua maioria, 59,57% tem apenas ensino fundamental e, não possuem filhos com seu agressor (44,86%) que é uma pessoa de seu relacionamento anterior (44,49%), com o qual não possui antecedentes registrados (52,18%), mais ainda não possui antecedentes registrados com outro agressor (60,83%).(Tabela 6).

Tabela 6 - Características da vítima - Feminicídio Tentado - RS

Características da vítima	Feminicídio Tentado (n=1353)
Escolaridade	
Ensino Fundamental	806(59,57)
Ensino Médio	288(21,29)
Ni	143(10,57)
Outros	116(8,57)
Raça/Cor	
Branca	1085(80,19)
Negra	190(14,04)
Mulata	22(2,74)
Outros	19(1,40)
Idade (anos)	
0 - 12	18(1,33)
13 - 17	65(4,80)
18 - 24	285(21,06)
25 - 29	193(14,26)
30 - 34	226(16,70)
35 - 39	192(14,19)
40 - 44	125(9,24)
45 - 49	99(7,32)
50 - 54	60(4,43)
55 - 59	15(1,11)
60 - 64	29(2,14)
65 - 69	15(1,11)
70 - 79	11(0,81)
> 80	2(0,15)
Relação com o agressor	
Relacionamento Atual	553(40,87)
Relacionamento Anterior	602(44,49)
Relacionamento Familiar	190(14,04)
Outro Relacionamento	8(0,59)
Filhos com o agressor	
Não	607(44,86)
Sim	417(30,82)
Ni	329(24,32)
Antecedentes registrados com o agressor	
Sim	706(52,18)
Não	646(47,75)
Ni	1(0,07)
Antecedentes registrados com outro agressor	
Não	823(60,83)
Sim	530(39,17)

Dados são apresentados por totais (percentuais)

Sobre o agressor: quanto ao autor neste período, 95,27% são homens e 4,73% são mulheres, suas idades variam entre 18 e 34 anos (49,10% dos homens e das mulheres 38,09%), também 65,32% dos homens e 51,56% das mulheres possuem apenas ensino fundamental e são auto declarados brancos 77,50% dos autores e 76,56% das autoras. Neste tipo de crime 56,71% dos homens autores e 78,13%

das mulheres autoras permaneceram em liberdade, e apenas 2,64% dos agressores homens cometeram suicídio, e 38,09% dos autores e 15,63% das autoras possuíam ocorrências registradas com outras vítimas. (Tabela 7).

Tabela 7 - Características do agressor - Feminicídio Tentado - RS

Características do agressor	Feminicídio Tentado (n=1353)	
	Masculino (n=1289(95,27))	**Feminino (n=64(4,73))**
Escolaridade		
Ensino Fundamental	842(65,32)	33(51,56)
Ensino Médio	208(16,14)	17(26,56)
Ni	144(11,17)	10(15,63)
Outros	95(7,06)	4(12,50)
Raça/Cor		
Branca	999(77,50)	49(76,56)
Negra	214(16,60)	7(10,94)
Outros	76(5,90)	8(6,17)
Idade (anos)		
13 - 17	35(2,72)	5(7,81)
18 - 24	213(16,52)	12(18,75)
25 - 29	207(16,06)	11(17,19)
30 - 34	213(16,52)	13(20,31)
35 - 39	182(14,12)	6(9,38)
40 - 44	145(11,25)	3(4,69)
45 - 49	109(8,46)	6(9,38)
50 - 54	77(5,97)	1(1,56)
55 - 59	45(3,49)	2(3,13)
60 - 64	20(1,55)	0(0,00)
65 - 69	16(1,24)	1(1,56)
70 - 79	8(0,62)	1(1,56)
> 80	2(0,16)	0(0,00)
Ni	17(1,32)	3(4,69)
Status Policial na época do crime		
Liberdade	731(56,71)	50(78,13)
Recolhido	391(30,33)	8(12,50)
Outros	160(12,41)	6(9,38)
Cometeu suicídio após o crime		
Não	1182(91,70)	58(90,63)
Sim	34(2,64)	0(0,00)
Ni	73(5,66)	6(9,38)
Antecedentes registrados com outra vítima		
Não	779(60,43)	51(79,69)
Sim	491(38,09)	10(15,63)
Ni	19(1,47)	3(4,69)

Dados são apresentados por totais (percentuais)

Distribuição Populacional de Raça ou Cor - RS

A Tabela 8, abaixo, mostra a distribuição da população do Rio Grande do Sul, segundo o Censo 2010 (IBGE), servindo para a análise da vítima e do autor.

Tabela 8 – Distribuição da população do RS segundo Raça/Cor

IBGE - População Residente - Percentual do Total Geral							
Unidade da Federação RS - Censo 2010							
Sexo	Cor ou Raça (%)						
	Total	Branca	Negra	Amarela	Parda	Indígena	Sem Declaração
Total	100,00	83,22	5,57	0,33	10,57	0,31	0,00
Homem	48,67	40,23	2,75	0,16	5,37	0,15	0,00
Mulher	51,33	42,99	2,81	0,17	5,20	0,15	0,00

Discussão

Este estudo mostrou o perfil de uma morte que acontece à noite, na residência da vítima, através do disparo de arma de fogo realizado por um homem branco, vivendo um relacionamento atual com uma mulher branca, ambos com idade entre 18 e 24 anos, com pouca instrução, sem filhos, sem antecedentes registrados, devido ao fim o relacionamento, em sua maioria ele é capturado pelos órgãos de segurança.

O segundo perfil é de uma tentativa de morte que ocorre entre a noite e a madrugada, dentro de casa, com uma arma branca por um motivo ainda desconhecido onde um homem ataca uma mulher ambos brancos entre 18 e 24 anos, com pouca instrução, sem filhos e sem antecedentes registrados, saídos de um relacionamento, e que após o crime permaneceu em liberdade.

Este trabalho permite ainda outros estudos que ajudem a complementar estes perfis e como era o objetivo inicial, como por exemplo, criar através de uma técnica estatística, um indicador preditivo de mortalidade para mulheres em situação de vulnerabilidade, por violência doméstica.

Referências

[1] BRASIL, Lei N 11.340, de 7 de agosto de 2006. Diário Oficial da República Federativa do Brasil.

[2] CATRACA LIVRE: <https://catracalivre.com.br/geral/cidadania/indicacao/maria-da-penha-uma-mulher-que-sobreviveu-na-luta/>.Acesso em: 10 de setembro de 2017.

[3] IBGE: https://www.ibge.gov.br/estatisticas-novoportal/sociais/populacao/2098-np-censo-demografico/9662-censo-demografico-2010.html.Acesso em: 22 de setembro de 2017.

[4] SECRETARIA DA SEGURANÇA PÚBLICA/RS. *Banco de dados sobre Feminicídio.* Observatório Estadual de Segurança Pública. Período Agosto 2006 a Julho 2017. Coleta Agosto 2017.

Mando de Campo e Gol Qualificado - Análise da Vantagem na Copa do Brasil

Alice Paul Waquil [1]

Eduardo Horta[2]

Jean Carlo Moraes[3]

Resumo: No futebol, acredita-se que em confrontos de mata-mata – isto é, disputas eliminatórias com jogos de ida e volta – o time que faz o segundo jogo em seu estádio teria uma vantagem. Essa crença vem do fato, amplamente reconhecido na literatura científica, de que o fator local é uma vantagem numa partida de futebol. Logo, muitos pensam que fazer o segundo jogo com essa vantagem traria uma maior chance de classificação. Quando os confrontos estão empatados em número de pontos, precisa-se de um critério para definir o vencedor; os três mais usados são o *saldo de gols*, o *gol qualificado* (em que o vencedor do confronto será o time que marcar mais gols enquanto joga como visitante) e a *disputa de pênaltis*.

Esse estudo traz evidência de que decidir um confronto mata-mata em casa é um benefício quando olhado de forma geral, pois o mandante se classifica em aproximadamente 65% das disputas. Porém quando a decisão se dá por gol qualificado ou pênaltis o percentual de classificação é 20% menor, ou seja, esses critérios beneficiam o time visitante, se não dando a vantagem, ao menos equiparando as chances das duas equipes. Além disso identificou-se que a probabilidade de classificação está relacionada com a diferença de qualidade entre os times.

Palavras-chave: *estatística esportiva, métodos estatísticos aplicados à análise futebolística*.

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *alice.waquil@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *eduardo.horta@ufrgs.br*
[3]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *jean.moraes@ufrgs.br*

1 Introdução

O futebol é um esporte de alcance mundial. Estima-se que esse esporte movimente anualmente entre 480 e 600 bilhões de reais, valor maior que o PIB de vários países. Com tanto dinheiro envolvido, o futebol se aliou à tecnologia e à ciência para, cada vez mais, entregar um produto de qualidade para seus espectadores. A cobertura esportiva investe em tecnologia que permite interação entre o público e a partida, os times investem em medicina avançada para prevenir e combater lesões e em métodos estatísticos para decidir escalações e esquemas táticos. Entretanto, apesar desses avanços, há ainda alguns "mitos" no futebol, mas que não necessariamente encontram suporte na literatura científica, muitas vezes pela simples ausência de estudos que busquem avaliar tais questões.

Uma crença antiga é que no futebol, assim como em outros esportes, um time jogar uma partida "em casa", isto é, em seu estádio, representa uma vantagem. Essa vantagem do fator doméstico é essencialmente um consenso na literatura científica, amplamente apoiado pelo dados [6],[8]. Portanto, os artigos buscam, em geral, compreender suas possíveis causas. Os principais fatores considerados são: torcida; fadiga de viagens; familiaridade com o local; viés do árbitro; territorialidade; táticas especiais; regras; fatores psicológicos.

A crença de que existem vantagens associadas ao mando de campo também ocorre quando se considera confrontos de eliminatórias simples, nos quais são disputados dois jogos, ocorrendo um na casa de cada time [3]. Neste caso, acredita-se que cada equipe terá uma vantagem quando jogar na sua casa, mas que o time mandante da segunda partida terá uma vantagem maior no total do confronto. Quando os confrontos estão empatados em número de pontos, utiliza-se de um critério para definir o vencedor do confronto, sendo os mais comuns o *saldo de gols*, o *gol qualificado* (em que o vencedor do confronto será o time que marcar mais gols enquanto joga como visitante) e a *disputa de pênaltis*.

O presente estudo visa analisar a vantagem de decidir, no Brasil, um confronto de eliminatórias simples como mandante, e principalmente a influência da regra do gol qualificado sobre isso, pois esse critério de desempate é atualmente utilizado nos principais campeonatos com sistema de eliminatórias no mundo inteiro. Todavia, ao se considerar que ao término do primeiro jogo esse o resultado está fixo, ou seja, o time que jogou como visitante não pode mais alterar o número de gols marcados fora de casa, então no segundo jogo apenas um time pode modificar esse critério. Posto isso, decidir um confronto de eliminatórias simples como mandante, sob a regra do gol qualificado é, de fato, um benefício?

2 Metodologia

A revisão de literatura baseia-se em uma série de dez artigos de periódicos nacionais e internacionais que abordam principalmente a vantagem do mando de campo em jogos de futebol e as possíveis causas

desse fenômeno. Os dados, contendo os resultados dos confrontos já disputados pela Copa do Brasil, de 1989 a 2017, foram coletados nos sites Wikipédia e Bola Na Área, em que a informação estava disponível para todos os anos. As análises foram feitas por meio de estatística descritiva, testes de hipóteses e regressão logística utilizando o software R Studio [12].

2.1 Formulação do Índice de Qualidade

Acredita-se que o principal componente explicando o desfecho de um confronto em duas partidas seja a diferença entre as qualidades dos times participantes do confronto. Foi criada, portanto, com base nos atuais critérios da CBF, utilizados desde 2014, uma variável instrumental que mede a qualidade dos times em cada ano. Os times recebem uma pontuação de acordo com suas classificações no campeonato brasileiro e Copa do Brasil, além das participações nas copas Sul-Americana ou Libertadores, caso não tenham participado da Copa do Brasil. A pontuação de um ano então é calculada como uma média ponderada dos cinco anos anteriores. Note que a pontuação anual contemporânea ao confronto não entra no cômputo do índice, no intuito de evitar problemas de endogeneidade.

Como o número de times participantes no campeonato brasileiro costumava variar entre os anos, a convenção da CBF prevê que, a partir do vigésimo-terceiro colocado, todos os times recebem a mesma pontuação para que se mantenha o critério de que todos os participantes de uma série têm a pontuação sempre superior a do primeiro colocado da série imediatamente inferior.

Os anos anteriores a 1994 não possuem o ranking completo, pois não havia Copa do Brasil antes de 1989, logo a pontuação dos times é mais baixa já que não considera esse campeonato, então, para manter o padrão foi decidido que estes dados não seriam utilizados nas análises.

No modelo, foi utilizada a variável que representa a diferença entre a pontuação padronizada dos dois times. Percebe-se que essa variável além de representar a diferença de qualidade, também capta a variação entre as fases disputadas na competição, já que, geralmente, quanto mais final for a fase, menor é a diferença na qualidade. Dessa forma, ao incluir a qualidade no modelo, considerou-se desnecessária a inclusão da fase em disputa.

2.2 Regressão Logística

A variável resposta nesse estudo é a classificação do visitante, que é uma variável dicotômica, assumindo valor 1 se o time visitante obteve a classificação (sucesso) e 0 caso contrário. Dessa forma, optou-se por utilizar modelos de Regressão Logística, que são adequados para descrever o tipo de problema abordado pois fazem um ajuste de modo à estimar a probabilidade de sucesso da resposta com base nas variáveis explicativas.

3 Resultados e Discussões

Na Copa do Brasil, considerando-se todos os confrontos, a proporção média de classificação do visitante é de 36,78%. Indicando que, como era esperado pela crença popular, há uma vantagem em decidir um confronto mata-mata como mandante. Porém, cerca de 35,41% do total de confrontos terminam empatados em pontuação e, portanto, precisam de um critério de desempate.

Quando o saldo de gols define o confronto o percentual de visitantes classificados é 37,76%, próximo ao geral. No entanto, os casos em que foram utilizados gol qualificado ou disputa de pênaltis esse percentual é de, respectivamente, 55,55% e 51,90%, ou seja, a vantagem passa a ser do time visitante. Mesmo que, nesses casos, as proporções estejam próximas de 0,5, e, portanto, não representem uma grande vantagem ao visitante, o fato importante é o aumento significativo em relação às outras possibilidades.

O modelo de regressão logística foi ajustado utilizando-se a classificação do visitante como variável resposta e como regressores as seguintes variáveis: diferença de qualiadade e três variáveis indicativas, uma para cada tipo de critério de desempate. Também foram incluidas no modelo as interações de cada dummie com a a diferença de qualidade. A Figura 1 representa o resultado do modelo, as probabilidades preditas de classificação do visitante para cada tipo de definição.

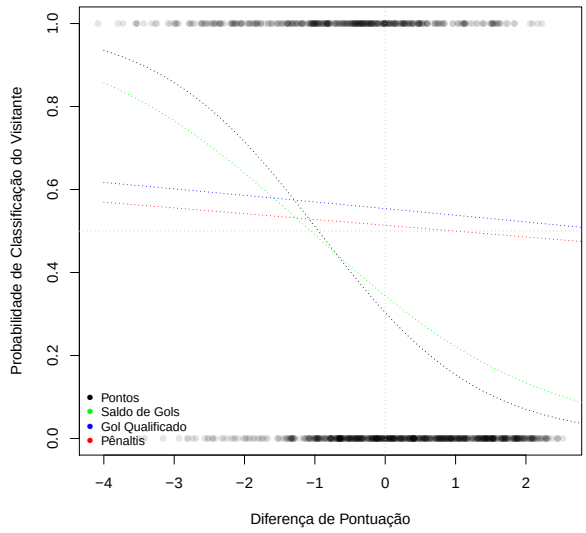


Figura 1: Probabilidades preditas de classificação do visitante em função da diferença de qualidade

Percebe-se na Figura 1 que, em todos os casos, a probabilidade de classificação do time visitante diminui a medida que a diferença de qualidade fica mais positiva, ou seja, quanto melhor é o mandante e pior é o visitante. Nos casos em que um dos times se classificou por pontuação ou por saldo de gols, os times têm iguais probabilidades de classificação quando o time visitante é melhor, em torno de um

desvio padrão.

Por outro lado, as decisões utilizando Gol Qualificado ou Pênaltis não apresentam uma diminuição tão significativa em relação ao aumento na diferença de qualidade em favor do mandante. Para o visitante só é melhor decidir por pontuação ou saldo de gols quando ele é mais do que um desvio padrão melhor do que o mandante.

4 Conclusão

Encontrou-se evidências de que, incondicionalmente, o fator doméstico constitui de fato uma vantagem nos confrontos de mata-mata, pois o mandante se classifica em aproximadamente 63% das disputas (significativamente maior que 0,5, p-valor<0,0001). Porém quando a decisão é por gol qualificado ou pênaltis o percentual de classificação é cerca de 20% menor que o percentual geral, (p-valor<0,0001 e p-valor=0,0113 respectivamente), ou seja, esses critérios beneficiam significativamente o time visitante, se não dando a vantagem, ao menos equiparando as chances das duas equipes (os dois critérios levam à proporções que não são diferentes de 0,5, p-valores 0,2229 e 0,7381).

Pode-se apontar que o uso de Gol Qualificado ou Pênaltis como critério de desempate aumenta a probabilidade de classificação do time visitante, considerando os efeitos isoladamente ou interagindo com a diferença de qualidade entre os times. Já o aumento na diferença de qualidade tem o efeito inverso, pois, quanto mais positiva é a diferença, melhor é o time mandante, e menor é a chance de o time visitante ser o vencedor do confronto.

Referências

[1] SEÇKÍN, A.; POLLARD, R. *Home Advantage in Turkish Professional Soccer*. *Perceptual and Motor Skills*. v. 107, p. 51-54, 2008.

[2] GELADE, G. A. *National Culture and Home Advantage in Football*. *Cross-Cultural Research*. v. 49, n. 3, p. 281-296, 2015.

[3] PAGE, L.; PAGE, K. *The Second Leg Home Advantage: Evidence from European Football Cup Competitions*. *Journal of Sports Sciences*. v. 25, n. 14, p. 1547-1556, 2007.

[4] GOUMAS, C. *Home Advantage in Australian Soccer*. *Journal of Science and Medicine in Sport*. v. 17, n. 2014, p. 119-123

[5] GÓMEZ, M. A.; POLLARD, R. *An Analysis of Home Advantage in the Top Two Spanish Professional Football Leagues*. *Perceptual and Motor Skills*. v. 108, p. 789-797, 2009.

[6] POLLARD, R. *Home Advantage in Soccer: a Retrospective Analysis*. *Journal of Sports Sciences*. v. 4, p. 237-248, 1986.

[7] POLLARD, R. *Home Advantage in Football: a Current Review of an Unsolved Puzzle*. *The Open Sports Sciences Journal*. v. 1, p. 12-14, 2008.

[8] POLLARD, R. *Worldwide Regional Variations in Home Advantage in Association Football*. *Journal of Sports Sciences*. v. 24, n. 3, p. 231-240, 2006.

[9] ALMEIDA, L. G.; OLIVEIRA, M. L.; SILVA, C. D. *Uma Análise da Vantagem de Jogar em Casa nas Duas Principais Divisões do Futebol Profissional Brasileiro*. *Revista Brasileira de Educação Física e Esporte*. v. 25, n. 1, p. 49-54, 2011.

[10] MEDEIROS N. C.; SILVA C. D.;POLLARD, R. *Home Advantage in Football in Brazil: Differences Between Teams and the Effect of Distance Traveled*. *The Brazilian Journal of Sports Sciences*. v. 1, n. 1, p. 3-10, 2008.

[11] HOSMER D. W.; LEMESHOW S. *Applied Logistic Regression*. 2. ed. 2000.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. . 2016. Disponível em: <https://www.R-project.org/>

Aplicação do método de séries temporais funcionais em linguagem R

Vitória Maria Martini Wendt [1]

Eduardo de Oliveira Horta[2]

Resumo: Dados funcionais estão cada vez mais em evidência principalmente no âmbito científico. Nesse sentido, propomos a implementação de uma ferramenta que padronize computacionalmente o uso de um método importante na área introduzido por Bathia et al (2010). Desta forma, foi desenvolvido o pacote em linguagem R ftsa2 que almeja tornar análises de séries temporais funcionais mais rápidas e universais, automatizando e melhorando processos.

Palavras-chave: *Dados funcionais, Série Temporal, Métodos Numéricos*.

1 Introdução

Séries temporais funcionais são sequências de dados funcionais ordenadas no tempo. Em muitos casos, uma série temporal funcional é obtida a partir de um processo estocástico a tempo contínuo, mediante uma quebra do processo original em uma sequência de processos concatenados. Por exemplo, o monitoramento contínuo da temperatura em uma estação metereológica induz uma sequência de gráficos anuais de trajetórias dessa variável. Em outras situações, uma série temporal funcional pode ser constituída de funções cujo domínio não é o tempo contínuo. Esse é o caso, por exemplo, de uma sequência de *kernel density estimates*, onde a dimensão temporal reside tão somente no ordenamento dessa sequência.

Esta nova forma de visualizar e analisar os dados está ganhando espaço nas mais diversas áreas da ciência, especialmente pelo fato de que sua aplicabilidade está diretamente relacionada ao aumento da capacidade computacional de processamento de dados. De fato, apenas recentemente encontra-se a implementação destes métodos em pacotes estatísticos como o R (pacote ftsa – functional time series analysis). Ainda assim, tal pacote não engloba uma importante contribuição metodológica feita por Bathia et al (2010). Nesse contexto, o presente trabalho tem como objetivo implementar e automatizar o uso computacional desta metodologia em linguagem R.

O principal método utilizado no contexto de análise de dados funcionais e que está presente no pacote ftsa é o método de componentes principais funcionais, que consiste em expandir cada curva observada em uma base associada à função de covariância correspondente. Este método apenas consegue fazer

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *vitoriawendt@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *eduardohorta@ufrgs.br*

inferência sobre os dados com certas restrições, a saber, de que não há presença de erros de medida. O modelo desenvolvido por Bathia et al (2010) vem justamente suprir esta necessidade de se trabalhar com dados funcionais realistas e que apresentam dependência entre suas observações e erros de medida.

Nota-se que é latente a necessidade de implementação computacional do modelo de séries temporais funcionais citado. Porém, é necessário que esta implementação ocorra de forma padronizada e respeitando boas práticas de programação já que o modelo prevê o uso de métodos computacionalmente custosos como o Bootstrap, além de envolver bases de dados usualmente grandes.

2 Metodologia

O método de análise de componentes principais funcionais é central no contexto de dados funcionais e séries temporais funcionais. Em suma, dado um conjunto de observações de dados funcionais (que podem ser uma série temporal) $x_1, x_2, \ldots, x_n$, a representação de Karhunen–Loève garante que, sob certas condições de regularidade, a seguinte representação é válida:

$$x_t(u) = \mathbb{E}x_1(u) + \sum_{j=1}^{d} Z_{t,j}\varphi_j(u), \tag{1}$$

onde $Z_{t,j}$ são variáveis aleatórias reais de média zero e variância λ_j, e onde φ_j são funções determinísticas que satisfazem a equação

$$\varphi_j(u) = (1/\lambda_j) \int \mathrm{Cov}(y_0(u),\, y_0(v))\, \varphi_j(v)\, dv.$$

Assim, por exemplo, a dinâmica dos dados funcionais $x_1, \ldots, x_n$ se resume à dinâmica do vetor aleatório $\boldsymbol{Z}_t = (Z_{t,1}, \ldots, Z_{t,d})$. Esse fato é importante pois permite ao analista modelar e predizer os dados funcionais através de métodos usuais de séries temporais multivariadas, como o modelo VAR.

Nesse contexto, uma importante contribuição foi dada por Bathia et al. (2010). Em muitos casos, os dados funcionais de interesse, x_t, são mensurados na presença de ruído ε_t, de forma que o estatístico tem acesso somente aos dados (y_t), onde

$$y_t(u) = x_t(u) + \varepsilon_t(u), \qquad \mathbb{E}(\varepsilon_t(u)) = 0. \tag{2}$$

Em um cenário desse tipo, torna-se impossível estimar a estrutura de componentes principais funcionais dada em (1). Os autores propõem uma representação alternativa,

$$x_t(u) = \mathbb{E}x_1(u) + \sum_{j=1}^{d} W_{t,j}\psi_j(u), \tag{3}$$

a qual pode ser recuperada a partir dos dados.

O procedimento de estimação baseia-se no cômputo das quantidades $\int y_t(u),\, y_s(u)\, du$, com $t, s = 1, \ldots, n$, a partir das quais é possível recuperar os demais estimadores da teoria. Esse fato mostra que a metodologia pode rapidamente demandar um elevado custo computacional.

3 Implementação computacional

Uma das principais ferramentas atualmente disponíveis para padronização do uso computacional de uma metodologia estatística é a linguagem R com sua rica estrutura de pacotes. Esta ferramenta possibilita o desenvolvimento e compartilhamento de funções de forma compacta e informativa, podendo conter exemplos e tutoriais explicativos para o usuário. Existem múltiplas formas de desenvolvimento de pacotes em liguagem R. No presente projeto, utilizou-se funcionalidades disponíveis na plataforma RStudio, as quais permitem a criação de pacotes sem a necessidade do uso de mais intervenientes.

A versão Beta do pacote ftsa2 está em uso atualmente para que suas características sejam testadas. Estas características referem-se principalmente a aplicabilidade do método quanto a armazenagem de dados e tempo de execução de funções. Em paralelo, técnicas estatísticas estão sendo testadas e desenvolvidas visando uma melhor aderência do pacote às premissas do modelo.

3.1 Armazenamento

A primeira função criada neste pacote permite que o usuário carregue seus dados sem precisar manipulá-los dentro de objetos. Esta é uma facilidade que se contrapõe à abordagem presente no pacote ftsa, onde o usuário deve utilizar como *input* um objeto de classe fts. Tal objeto trata-se de um array de dimensões (1,t_u,u), aonde t_u é o número de séries temporais de comprimento u.

No pacote ftsa2, o usuário carrega seus dados utilizando arquivos de extensão .txt ou .csv e uma função específica compacta estas informação dentro de uma matriz. A escolha por se trabalhar com matrizes ao invés de arrays ou outros tipos de objetos está ligada ao evidente ganho computacional na manipulação de dados utilizando este tipo em linguagem R. Comparando tempos de desempenho, o pacote ftsa2 reduz em quase 80% os tempos de execução quando comparado com a função do método de componentes principais funcionais do pacote ftsa.

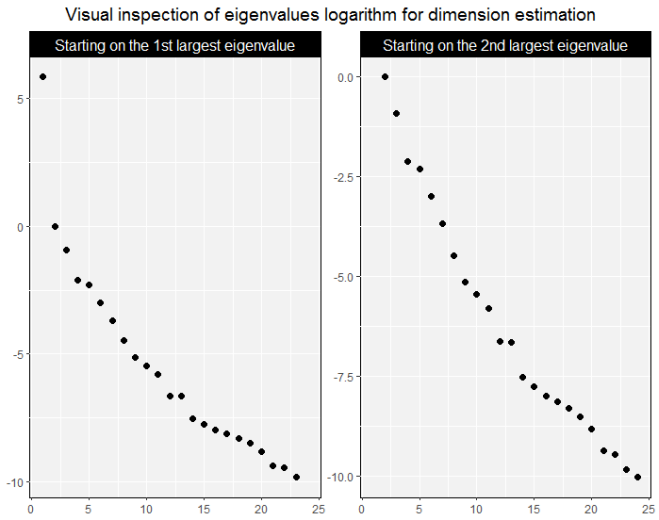
3.2 Estimação

Uma das principais premissas que o pacote ftsa2 carrega é o de automatizar o uso de séries temporais funcionais para usuários em geral, dentro e fora da academia. Seguindo a metodologia proposta em Bathia et al. (2010), o pacote apresenta a funcionalidade de estimação dos parâmetros do modelo especi-

ficado em (2) e (3), a saber, a função-média $\mathbb{E}x_1(u)$, a dimensão d e as auto-funções $\psi_j(u)$, $j = 1, \ldots, d$, além de retornar estimativas para as séries temporais latentes $W_{t,j}$, $j = 1, \ldots, d$.

3.3 Seleção de modelos

O parâmetro d representa a dimensão do modelo de séries temporais funcionais e, como citado anteriormente, também deve ser estimado. A forma mais intuitiva e que também está contemplada no pacote ftsa2 é estimação por inspeção visual dos autovalores.



Todavia, essa abordagem pode ser considerada *ad hoc*. Nesse sentido, o pacote oferece ao usuário um critério de seleção de d via método de bootstrap conforme apresentado em Bathia et al. (2010).

Outro método para a seleção de d que pode ser utilizado é o de validação cruzada por erro de previsão fora da amostra. Esta técnica ainda encontra-se em período de testes no pacote, pois requer um banco de dados significativamente grande, além de ser extremamente custosa do ponto de vista computacional.

3.4 Predição

O principal objetivo do método de séries temporaris funcionais introduzido por Bathia et al (2010) é fazer predições acuradas sobre os dados. Deste modo, é crucial a presença de uma função que a partir dos dados seja capaz de fazer previsões respeitando as premissas de aplicabilidade de métodos de *forecasting*.

A primeira premissa que deve ser cumprida é a de que a série temporal Y trata-se de um processo estacionário. Mas testar estacionariedade de Y é o mesmo que testar a estacionariedade dos coeficientes latentes $\hat{W}_{tj}$. Deste modo, foi desenvolvido uma função no ftsa2 que utiliza testes de estacionariedade já implementados em outros pacotes de R, como o ADF, para averiguar a adequação das suposições acima mencionadas.

Na função de predição, o usuário pode escolher se deseja conduzir a modelagem das séries temporais latentes de forma autônoma ou automática. No segundo caso, o pacote gerará predições a partir de modelos ARIMA se $d = 1$ ou de modelos VAR no caso em que $d > 1$.

3.5 Visualização de dados funcionais

Uma questão inerente ao modelo de dados funcionais utilizado é: como melhor visualizar um conjunto de dados funcionais? Esta é uma questão ainda não explorada em linguagem R de forma geral. Pensando neste problema, o pacote ftsa2 trará em sua composição um conjunto de funções que permite esta visualização utilizando métodos diferentes.

Um destes métodos já implementados é o Waterfall Graphs e é baseado em uma publicação da área de econometria sobre a visualização de múltiplas densidades. Este método permite que o usuário veja o comportamento da série temporal Y, podendo até avaliar de forma visual a estacionariedade do processo.

4 Conclusão

Conforme mencionado, a inferência sobre conjuntos de dados de natureza funcional está a cada dia se difundindo mais na literatura estatística e em áreas afins. O paradigma teórico neste campo é o método de componentes principais. Tal abordagem apresenta um importante inconveniente no caso em que os dados funcionais são observados com erro de medida. No contexto de séries temporais funcionais, a metodologia proposta em Bathia et al. (2010) representa uma importante contribuição, a qual ainda não havia sido implementada em pacotes estatísticos de amplo uso. O pacote ftsa2 surge para preencher essa lacuna.

Os resultados computacionais obtidos pelo pacote são extremamente satisfatórios quando comparado aos métodos já implementados no pacote ftsa. Isto se dá principalmente pela aplicação de boas técnicas de programação como o uso de matrizes para armazenar os dados.

O método desenvolvido por Bathia et al (2010) prevê a estimação de parâmetros que nem sempre podem ser obtidos de forma analítica. Deste modo, o pacote ftsa2 procurou também aplicar métodos computacionais para certas resoluções, não se privando de utilizar técnicas já desenvolvidas em outro pacotes, aumentando assim também o ganho computacional.

O pacote ftsa2 for construído sob a plataforma RStudio utilizando a versão 3.4 da linguagem R e ainda está sendo testado e desenvolvido. Por enquanto, apenas o caso univariado do modelo foi aplicado. Portanto, faz parte das próximas fases de desenvolvimento implementar também o modelo para casos multivariados em que o custo computacional será potencialmente maior. Porém, as técnicas desenvolvidas até o momento pelos autores neste projeto de Iniciação Científica Voluntária, à qual este trabalho está vinculado, mostraram-se extremamente importantes para uma futura propagação computacional do

modelo como principal forma de análise de séries temporais funcionais.

Referências

[1] SMART, Francis. *Waterfall and 3D plotting exploration*. Disponível em: <https://github.com/EconometricsBySimulation/BivariateSlicer/blob/master/slicedens.R>. Acesso em: 27 jul. 2017.

[2] BATHIA, Neil; YAO, Qiwei; ZIELGELMANN, Flavio *IDENTIFYING THE FINITE DIMENSIONALITY OF CURVE TIME SERIES*. Disponível em: <https://projecteuclid.org/euclid.aos/1291126960>. Acesso em: 15 mai. 2017.

Classificação de Doenças Cardíacas Através de Eletrocardiogramas e Fonocardiogramas

Mikaela Baldasso[1]

Marcio Valk [2]

Airton Kist[3]

Resumo: Uma grande parcela da população sofre de problemas relacionados a doenças do coração que estão entre as principais causas de morte em todo o mundo. Em particular, 1-2% da população mundial sofre de algum tipo de arritmia cardíaca que pode afetar pessoas das mais variadas faixas etárias. Recentemente o "National Institute of General Medical Sciences" (NIGMS) lançou um desafio com o objetivo de estimular a proposição de técnicas para classificação dos diferentes tipos de arritmias baseados em eletrocardiogramas (ECG's) e fonocardiogramas (PCG's) que podem ser vistos como séries temporais em que a técnica de classificação e agrupamento baseada em U-estatísticas pode ser aplicada. A utilização dessas técnicas depende fundamentalmente de medidas de distâncias ou similaridade que sejam capazes de capturar diferenças entre dois ECG's (ou PCG's), quando elas existem. Abordagens muito comuns na análise de sinais, como a filtragem, que elimina os ruídos que possivelmente poderiam afetar a classificação, devem ser consideradas. A partir disso, pode-se utilizar ferramentas comuns na análise de séries temporais, como a autocorrelação que é característica definidora podendo ser usada na classificação dos diferentes tipos de arritmia. Por fim, neste estudo, os resultados são comparados aos disponibilizados pelo desafio sendo possível fazer uma comparação com outras técnicas propostas na literatura.

Palavras-chave: *Doenças Cardíacas, Classificação, Fonocardiograma.*

2 Introdução

As doenças cardiovasculares (DCV) continuam sendo a principal causa de morbidade e mortalidade no mundo todo Liu et al. (2016). Estima-se que 17,5 milhões de pessoas morreram de DCV em 2012, representando 31% de todas as mortes globais (OMS 2015). Um dos primeiros passos na avaliação do sistema cardiovascular é o exame físico. Auscultação dos sons do coração é parte essencial do exame físico e pode revelar muitas condições cardíacas patológicas,

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *mikaelabaldasso@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marcio.valk@ufrgs.br*
[3]UEPG - Universidade Estadual do Paraná. Email: *kist@uepg.br*

como arritmias, doença valvar, insuficiência cardíaca e muito mais. Os sons cardíacos fornecem importantes pistas iniciais na avaliação da doença, servem de guia para um exame diagnóstico posterior e, assim, desempenham um papel importante na detecção precoce de DCVs, (Liu et al., 2016).

Durante o ciclo cardíaco, o coração primeiro sente um impulso elétrico que leva a atividade mecânica sob a forma de contrações atriais e ventriculares. Isso, por sua vez, força o sangue entre as câmaras do coração e ao redor do corpo, como resultado da abertura e fechamento das válvulas cardíacas. Essa atividade mecânica e o início ou parada repentina do fluxo de sangue dentro do coração, dá origem a vibrações de toda a estrutura cardíaca (Liu et al., 2016). Essas vibrações são audíveis na parede torácica e escutas de sons cardíacos específicos podem dar uma indicação da saúde do coração. As gravações de áudio desses sons são armazenadas em forma de uma série temporal e a representação gráfica dos sons resultantes, obtidos a partir da superfície do tórax, é conhecida como fonocardiograma (PCG).

Quatro locais são mais utilizados para ouvir e transduzir os sons do coração, que são nomeados de acordo com as posições em que as válvulas podem ser melhor ouvidas (Springer et al., 2016): *Área aórtica*- centrada no segundo espaço intercostal direito. *Área pulmonar* - no segundo espaço intercostal ao longo da borda esternal esquerda. *Área tricúspide* - no quarto espaço intercostal ao longo da borda esternal esquerda. *Área mitral* - no ápice cardíaco, no quinto espaço intercostal na linha do meio da clavícula.

A Figura 1 apresenta um esquema de como o sinal *Fundamental heart sound* (FHS) é naturalmente dividido em partes, que se repetem com um comportamento cíclico. Como podemos ver na Figura 1, em um paciente saudável, primeiramente observa-se o "primeiro som" (S1) em seguida o "segundo som" (S2). Outros sons também são audíveis devido ao movimento natural do coração denominados (S3) e (S4) além dos chamados murmúrios cardíacos causados por um fluxo de sangue turbulento e de alta velocidade, entre outros ruídos. Sons externos, como a respiração, também podem afetar o PCG.

A análise automatizada do som cardíaco nas aplicações clínicas geralmente consiste em três passos; Pré-processamento, segmentação e classificação. Nas últimas décadas, métodos para segmentação automatizada e classificação de sons cardíacos foram amplamente estudados. Muitos métodos demonstraram potencial para detectar com precisão patologias em aplicações clínicas. Infelizmente, as comparações entre técnicas foram dificultadas pela falta de bases de dados de alta qualidade, rigorosamente validadas e padronizadas, de sons cardíacos obtidos a partir de uma variedade de condições saudáveis e patológicas. Em muitos casos, ambos os dados experimentais e clínicos são coletados a custos consideráveis, mas apenas analisados uma vez por seus

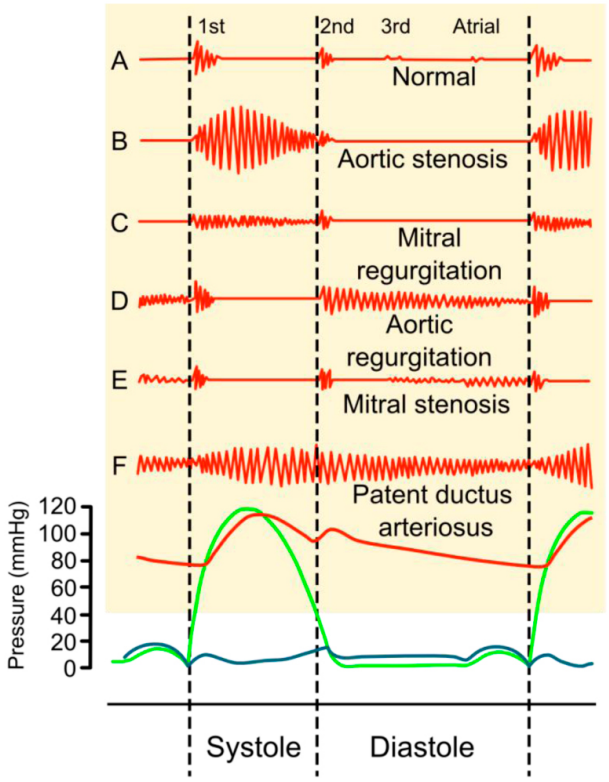


Figura 1: Fonocardiograma de sons cardíacos normais e anormais. Vermelho indica pressão aórtica, verde indica pressão ventricular e azul a pressão atrial. Fonte: Liu et al. (2016).

colecionadores e, em seguida, arquivados indefinidamente por variados motivos. Além disso, a energia necessária para documentar dados para uso externo, armazenar e compartilhar de forma semi-permanente raramente está disponível no final de um projeto de pesquisa Liu et al. (2016).

Outro aspecto importante na análise automatizada de sinais cardíacos é a segmentação que consiste basicamente na localização precisa de S1 e S2. Nesse trabalho, para classificar os sinais cardíacos, será empregada uma técnica que não faz a segmenteação dos sinais. Uma revisão metodológica pode ser vista em Liu et al. (2016).

3 Metodologia

A metodologia adotada nesse trabalho segue basicamente aquela adotada em Deng e Han (2016). Inicialmente cada sinal x é convertido para uma frequência de 2kHz e então é filtrado através de um filtro Butterworth *band-pass*, *zero-phase* de ordem 6 (25Hz-900Hz) para eliminar ruídos que ultrapassem a banda. Depois disso o sinal resultante, $\hat{x}$, é normalizado por

$$\overline{x} = \frac{\hat{x}}{\max(|\hat{x}|)}. \tag{1}$$

Já convertido e normalizado, o sinal é decomposto em quatro níveis usando a transformada *wavelet* Daubechies de ordem 6 devido às suas semelhanças morfológicas com os componentes do som cardíaco. Os coeficientes wavelet de aproximação do quarto nível e os coeficientes wavelet de detalhes do segundo nível são selecionados para extrair os envelopes de *energia Shannon* de média normalizada (ASE), respectivamente. Usando uma *lag-window* de 30ms com uma mudança de quadro de 15ms, a ASE é calculada por

$$e(n) = -\frac{1}{N_w}\sum_{j=1}^{N_w} d^2(j)\log(d^2(j)), \tag{2}$$

onde N_w é o comprimento da janela e $d(j)_{j=1}^{N_w}$ são os coeficientes da sub-banda na n-ésima *lag-window*.

Os coeficientes de aproximação do quarto nível são correspondentes ao alcance de frequência de 0-125Hz que capta a maioria das informações sobre os sons fundamentais do coração. O conteúdo da frequência dos coeficientes wavelet de detalhes do segundo nível está principalmente entre 500-1000Hz, que representa principalmente a informação sobre ruídos. A partir da equação (2), pode-se calculara a Energia de Shannon $e_a(n)$ dos coeficientes de aproximação e $e_d(n)$ dos coeficientes de detalhes da transformada discreta de wavelet. Será considerada a informação desses dois vetores para caracterizar um sinal normal de um sinal de um paciente com arritmia.

Como o som cardíaco é quase periódico e é composto por uma série de ciclos cardíacos, muitas estruturas semelhantes emergem em cada ciclo, tais como os sons fundamentais do coração e alguns ruídos. Também existe uma quase periodicidade nos sinais da sub-banda da batida do coração e em seus envelopes. A função de autocorrelação da ASE dos coeficientes wavelet da sub-banda podem capturar de forma automática e indireta a periodicidade dos múltiplos ciclos cardíacos e, consequentemente, pode ser uma ferramenta capaz de identificar as características do som cardíaco, (Deng e Han, 2016). Devido à simetria da função de autocorrelação, a ACF de sub-banda somente é calculada para *lags* positivos e é definida por

$$r(m) = \frac{1}{r(0)}\sum_{n=0}^{N-m-1} e(n)e(n+m), \qquad m > 0, \tag{3}$$

onde $e(n+m)$ é a versão deslocada no tempo do sinal $e(n)$, com um intervalo de tempo m para $m = 0, 1, \cdots, M$.

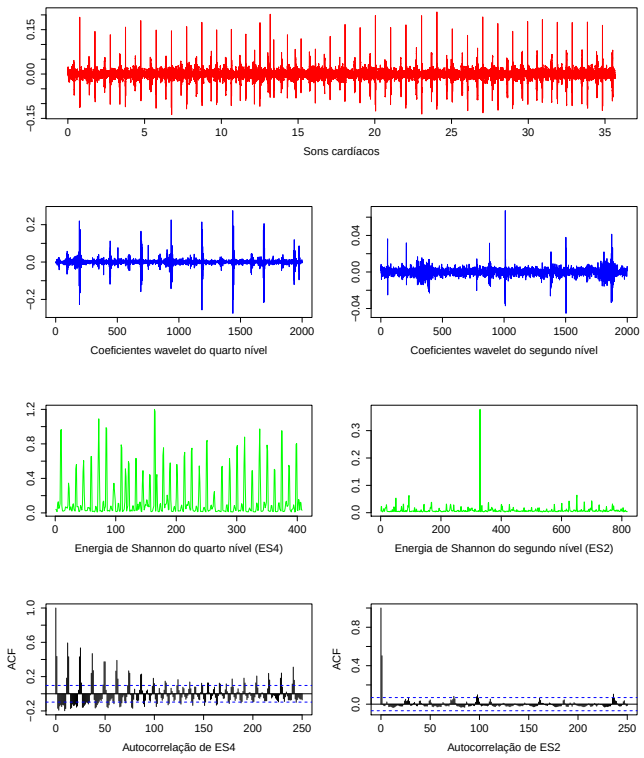


Figura 2: Autocorrelações dos coeficientes de aproximação e dos coeficientes de detalhes do som cardíaco com ruídos da sístole.

3.1 Função kernel gaussiano como medida de dissimilaridade

Em classificação e agrupamento de séries temporais é fundamental encontrar uma medida de dissimilaridade que seja capaz de detectar sinais com características distintas (ou processo gerador distinto). Em Deng e Han (2016) é utilizada a função kernel gaussiano como uma medida de dissimilaridade entre os FHS. O objetivo é que essa medida capture o padrão de comportamento de S1 e S2 e também os chamados murmúrios cardíacos. Devido à capacidade de capturar a estrutura temporal das informações, a ACF de sub-banda pode ser vista como a característica do sinal de sub-banda.

A característica ACF de sub-banda é construída colocando-se os M valores dos coeficientes ACF (ACFCs) como um vetor coluna $\mathbf{r} = [r(1), \cdots, r(M)]^T \in \mathcal{R}^M$. Os ACFCs dos coeficientes de aproximação e detalhes da transformada wavelet discreta (TDW), denotados, respectivamente, por aACFC $\mathbf{r}^a$ e dACFC $\mathbf{r}^d$, capturam a estrutura da informação temporal dos FHS e dos murmúrios, respectivamente (Deng e Han, 2016). Note que o valor de M deve ser maior que a sazonalidade do ciclo cardíaco para que os ACFCs contenham informação de todo o espaço de variação do som cardíaco. Em Deng e Han (2016) $\mathbf{r}^a$ e $\mathbf{r}^d$ são transformados em um só

vetor (adACFC) $\mathbf{r}^{ad} = [\mathbf{r}^a, \mathbf{r}^d] \in \mathcal{R}^{2M}$ para serem usados em classificadores (SVM, suport vector machine).

Sendo n o número de sinais e $\mathbf{r}_i$ o adACFC do sinal $i \in \{1, \cdots, n\}$, define-se a medida de similaridade baseada na função kernel gaussiano entre os sinais i e j por

$$\omega(i,j) = \exp\left\{-\frac{||\mathbf{r}_i - \mathbf{r}_j||^2}{\delta^2}\right\}, \tag{4}$$

em que δ^2 é a largura do kernel. A matriz de dissimilaridade W é composta pelas entradas $\omega(i,j)$.

4 Descrição dos dados

Em meados de 1999 foi criado o *PhysioNet Resouce* com o objetivo de estimular pesquisas atuais e novas investigações no estudo de sinais biomédicos e fisiológicos complexos. A partir disso, foi estabelecido pelo *Resource* o site *physionet.org* que é seu mecanismo de disseminação e intercâmbio livre e aberto de sinais biomédicos registrados e softwares de código aberto para analisá-los, fornecendo instalações para análise cooperativa de dados e avaliação de novos algoritmos propostos. O site é um serviço público do PhysioNet Research Resource for Complex Physiologic Signals, financiado pelo Instituto Nacional de Ciências Médicas Gerais (NIGMS) e pelo Instituto Nacional de Imagem Biomédica e Bioengenharia (NIBIB), que anualmente lança desafios para o público em geral que visa avanços na área da saúde.

Em 2016, foi proposto pelo PhysioNet o desafio do desenvolvimento de algum algoritmo capaz de classificar gravações do som do coração coletadas em uma variedade de ambientes, com o objetivo de identificar, a partir de uma única gravação curta (10-60s), quais sinais estão em bom estado e quais apresentam certo tipo de arritmia. Para isso, o site disponibilizou gravações de som do coração que foram obtidas de vários contribuidores em todo o mundo, coletados em um ambiente clínico ou não clínico, tanto de indivíduos saudáveis como de pacientes patológicos. O conjunto de treinamento disponível contém um total de 300 gravações de som de coração, que vão de 5 segundos a pouco mais de 120 segundos.

A Figura 3 mostra imagens de 4 sinais disponibilizados pelo *physionet.org*: dois de pacientes saudáveis e dois de pacientes com algum problema cardíaco. Essa representação gráfica dos ruídos cardíacos são chamados de *Fonocardiogramas* (PCG).

As gravações de som do coração foram coletadas de diferentes locais do corpo sendo quatro locais típicos: a área aórtica, a área pulmônica, a área tricúspide e a área mitral. Elas foram divididas em dois tipos: gravações de som cardíacas normais e anormais. As gravações normais eram

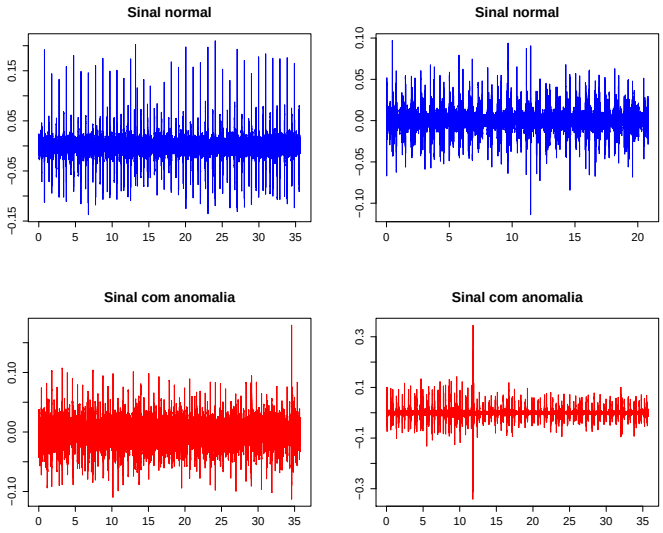


Figura 3: Exemplos de sinais cardíacos sem arritmia (parte superior) e sinais cardíacos com algum tipo de arritmia (parte inferior).

de indivíduos saudáveis e os anormais eram de indivíduos com diagnóstico cardíaco confirmado tais como pacientes que sofrem de uma variedade de doenças que geralmente são defeitos valvares cardíacos e pacientes com doença arterial coronariana. Os defeitos da valva cardíaca incluem o prolapso da valva mitral, regurgitação mitral, estenose aórtica e cirurgia valvular. Todas as gravações desses pacientes foram rotuladas como anormais.

Ambos os indivíduos saudáveis e pacientes patológicos incluem crianças e adultos. Cada um pode ter contribuído entre uma e seis gravações de som do coração que duram de vários segundos até mais de cem segundos. Todos os registros são baseados em 2000 observações por segundo e foram fornecidos em formato .wav.

5 Avaliação do Desempenho dos Métodos de Classificação

Muitas vezes as metodologias utilizadas para a classificação podem apresentar desempenhos diferentes quando comparadas em contextos distintos. Como um sujeito com a doença pode ser classificado como tendo a doença (verdadeiros positivos) ou pode ser classificado como não tendo a doença (falso positivo), também um indivíduo que não tem a doença pode ser classificados como sendo doente (falso negativo) ou não sendo doente (verdadeiro negativo). Assim, quantidade de verdadeiro positivos (VP), Negativos verdadeiros (VN), falso-positivo (FP) e falsos negativos (FN) cobrem todo o conjunto de possibilidades de classificação e os métodos podem podem apresentar diferenças, por exemplo, quando compara-se a proporção de FP e FN encontrados

por diferentes metodologias. Para uma comparação mais justa e precisa, utilizam-se algumas medidas que levam em consideração essas questões. As medidas mais usuais para avaliar a performance de um método de classificação são

$$\text{Sensitividade:} \qquad Se = \frac{VP}{VP + FN}$$

$$\text{Especificidade:} \qquad Es = \frac{VN}{FP + VN}$$

$$\text{Precisão/Acurácia:} \qquad Pr = \frac{VP + VN}{N}$$

6 Resultados

Podemos observar alguma separação na Figura 4, mais especificamente podemos observar 3 grupos. Como temos a possibilidade de não classificar sinais com muito ruído, podemos utilizar esse mapa para determinar os sinais "normais", "arritmia" e "não classifica".

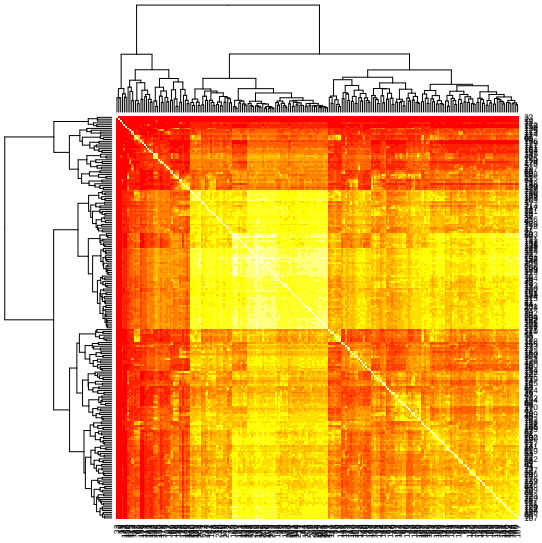


Figura 4: Mapa de calor com matriz de dissimilaridade obtida utilizando a função kernel gaussiano das autocorrelações da energia de Shannon calculada com coeficientes wavelet de quarto e segundo nível.

Os resultados obtidos nesse trabalho podem ser comparados, ou são comparáveis, com os resultados de Langley e Murray (2016) o qual apresenta uma precisão (score) de 78% e seu método

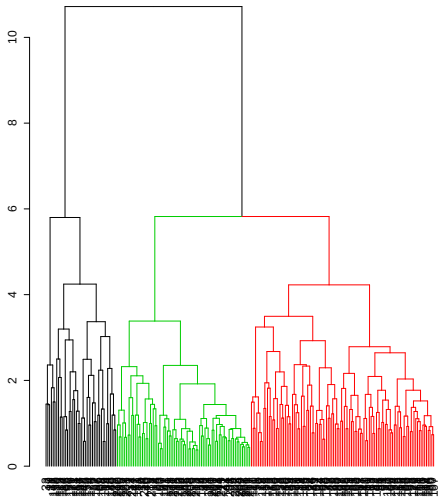


Figura 5: Dendrograma da Figura 4, com os grupos separados por cores. Preto (não classifica), vermelho (anormal) e verde (normal).

é baseado em uma medida de entropia dos coeficientes wavelets. Os resultados preliminares apresentam uma Sensitividade de 57%, Especificidade de 76.92% e uma precisão de 65.73%.

7 Conclusão

Nesse trabalho propomos uma abordagem para classificar sinais cardíacos em dois grupos: Normal e Anormal. Buscamos na literatura alguns métodos que se destacam na realização dessa tarefa e utilizamos algumas de suas ferramentas, como a decomposição wavelet dos sinais e a autocorrelação da energia de Shannon. Para a obtenção de uma matriz de dissimilaridade, foi utilizado a função kernel gaussiano. Então utilizamos um método de agrupamento hierárquico para obter os grupos. Observamos que existem 3 grupos bem definidos no conjunto de dados, o que é razoável já que especificamente para esses dados criou-se uma classe de sinais que eram muito ruidosos para serem classificados. Determinamos quais eram os grupos Normal, Anormal e Não Classifica. A precisão dessa abordagem foi de 65.73%, um pouco distante dos 78% encontrado por um dos participantes do desfio. No entanto, nesse trabalho não foi utilizado nenhuma informação a priori para determinar os grupos. Sendo assim, para a continuidade desse trabalho, pretendemos incorporar essa informação, determinando pequenos grupos de sinais normais e anormais para então classificar um novo elemento. Além disso, pretendemos utilizar as técnicas baseadas em U-estatísticas propostas por Cybis et al. (2017) para colocar inferência em classificação de sinais cardíacos.

Referências

Cybis, G. B., Valk, M., e Lopes, S. R. (2017). Clustering and classification problems in genetics through u-statistics. *Journal of Statistical Computation and Simulation*, pages 1–21.

Deng, S.-W. e Han, J.-Q. (2016). Towards heart sound classification without segmentation via autocorrelation feature and diffusion maps. *Future Generation Computer Systems*, 60:13–21.

Langley, P. e Murray, A. (2016). Abnormal heart sounds detected from short duration unsegmented phonocardiograms by wavelet entropy. In *Computing in Cardiology Conference (CinC), 2016*, pages 545–548. IEEE.

Leatham, A. (1975). Auscultation of the heart and phonocardiography.

Liu, C., Springer, D., Li, Q., Moody, B., Juan, R. A., Chorro, F. J., Castells, F., Roig, J. M., Silva, I., Johnson, A. E. W., Syed, Z., Schmidt, S. E., Papadaniil, C. D., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., Samieinasab, M., Samieinasab, M. R., Sameni, R., Mark, R. G., e Clifford, G. D. (2016). An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181.

Springer, D. B., Tarassenko, L., e Clifford, G. D. (2016). Logistic regression-hsmm-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, 63(4):822–832.

Avaliação da reconstrução de caractere em ancestral comum e estimação de correlações pelo modelo filogenético de variável latente

Lauren Alves Vieira [1] [3]

Gabriela Bettella Cybis[2] [3]

Resumo: O estudo de correlações entre variáveis fenotípicas ao longo da evolução é um dos problemas centrais da biologia evolutiva. O modelo filogenético de variável latente (Cybis et al. 2015) é uma opção para a estimação de tais correlações no contexto das filogenias bayesianas. O modelo permite a estimação simultânea de correlações entre variáveis contínuas, discretas ordenadas e discretas sem ordenamento, controlando para a história evolutiva compartilhada das amostras. Neste trabalho nós realizamos uma aplicação do modelo a um conjunto de dados de morcegos, que contem uma variável contínua e uma discreta não ordenada, na qual estimamos a correlação evolutiva entre as variáveis e reconstruímos o valor dessas variáveis no ancestral comum a todas as espécies em estudo. Como nos modelos ordenados a aplicação do modelo depende da escolha de um estado de referência, nós realizamos uma análise de sensibilidade, verificando que em geral estas estimativas são robustas à escolha do referencial.

Palavras-chave: *Variável latente, Filogenias, Inferência bayesiana*.

1 Introdução

O estudo das interações entre genótipos e fenótipos é um dos grandes focos da biologia evolutiva, com aplicações nas mais diversas áreas. Nesse contexto, uma questão de interesse é a estimação de correlações nos processos evolutivos de traços fenotópicos. Entretanto, para adequadamente estimar essas correlações, devemos separá-las das correlações induzidas pela história evolutiva compartilhada entre os indivíduos, que pode ser inferida a partir de dados genéticos. O modelo filogenético de variável latente (Cybis et al 2015) mostra-se como uma opção para estas análises, já que pode ser usado para estimar correlações entre diferentes tipos de dados fenotípicos, enquanto controla para a história evolutiva compartilhada dos indivíduos ou espécies em estudo.

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *laurendiasalves@gmail.com*

[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *gabriela.cybis@ufrgs.br*

[3]Um agradecimento especial a Gislene Lopes Gonçalves e Tiago Ferraz, do PPGBM da UFRGS pela disponibilização dos dados.

A separação de correlações inerentes aos processos de evolução dos fenótipos daquelas geradas pela história evolutiva é importante para a identificação de dois fenômenos de interesse biológico: ligação gênica e seleção natural. O estudo da evolução da resistência bacteriana a diferentes antibióticos é um exemplo de problema de interesse epidemiológico em que correlações na evolução de fenótipos são um indício de ligação gênica. De modo similar, pressões seletivas entre características com hábitos alimentares e traços morfológicos em grupos de mamíferos também podem ser estudadas por meio de correlações evolutivas.

Até onde temos conhecimento, o modelo filogenético de variável latente é o único modelo proposto que permite estimar correlações evolutivas entre traços contínuos, discretos binários e discretos com múltiplos estados ordenados ou não. Além disso, a metodologia de inferência bayesiana associada ao modelo e implementada na plataforma BEAST (software de inferência bayesiana para filogenias - Drumond et al. 2007) permite que estas correlações sejam estimadas, mesmo quando não se conhece a história evolutiva de modo preciso, fazendo uso direto de sequências de DNA. Adicionalmente, quando se considera o histórico de análises de correlações no contexto filogenético, nossa metodologia permite a análise de conjuntos de dados relativamente grandes.

Neste trabalho realizaremos a análise de um conjunto de dados de morcegos, disponibilizado por colaboradores (dados ainda não publicados), com o objetivo de estudar a correlação evolutiva entre o número de dentes e hábito alimentar destas espécies. Além disso, estimamos os valores dessa característica no ancestral comum do grupo de morcegos. Como a realização desta análise envolve algumas escolhas de referencial de parâmetros, realizamos uma pequena análise de sensibilidade para verificar o efeito destas escolhas sobre a estimação.

2 Metodologia

Modelo Filogenético de Variável Latente

Representamos a história evolutiva de um conjunto de N indivíduos por meio de um grafo acíclico denominado árvore filogenética F (ou filogenia). A árvore conta com N nós externos (vértices de grau 1), que representam os indivíduos da amostra no tempo presente, e uma raiz (vértice de grau 2), que representa o mais recente ancestral comum aos N indivíduos da amostra e o momento mais antigo no tempo representado na árvore. Além disso, há $N - 2$ nós internos (vértices de grau 3), que representam bifurcações evolutivas causadas pela separação de linhagens. O comprimento da arestas ligando esses nós representa a quantidade de tempo evolutivo até a ocorrência de uma bifurcação. A evolução das variáveis fenotípicas é modelada por um processo estocástico que inicia na raiz da árvore e evolui ao longo das arestas até os nós externos, onde o valor das variáveis nos N indivíduos da amostra é determinado. A

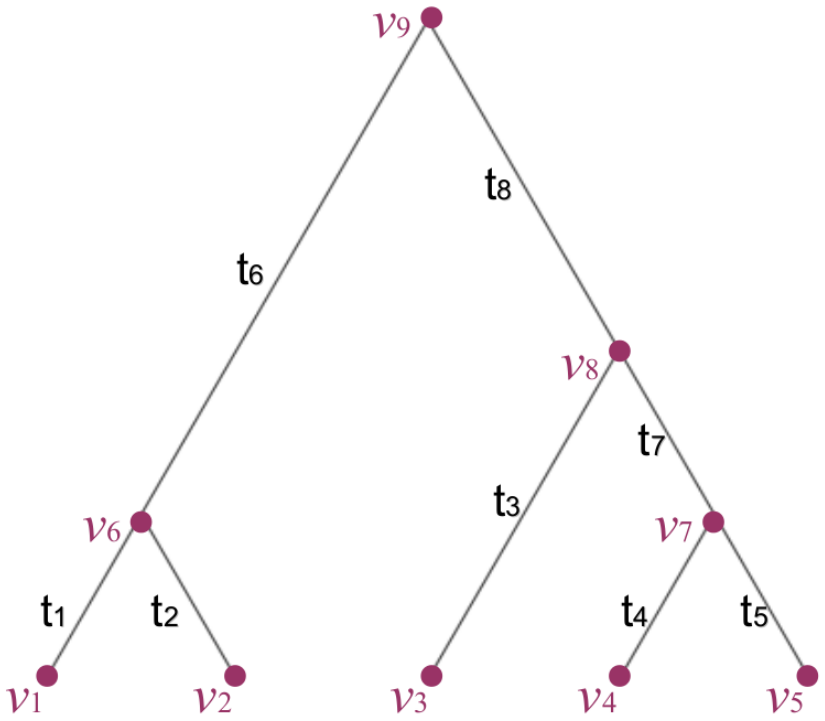


Figura 1: Exemplo de árvore filogenética com $N = 5$.

Figura 1 apresenta um exemplo de filogenia.

O modelo filogenético de variável latente descreve a evolução de uma variável latente X contínua, não observável, ao longo da árvore filogenética F, e a de uma variável de interesse observável Y. A evolução temporal da variável latente X segue um modelo de movimento browniano ao longo da filogenia. Ao final do processo, o valor da variável Y é determinado a partir de X por meio de uma função de ligação $g(X)$. Quando a variável Y é binária, por exemplo, seu valor é determinado pela posição de X em relação a um limiar, e quando Y é contínua temos $Y = X$. No caso de Y multivariado, cada componente de Y é determinada por mais de uma componente de X. A matriz de precisão Σ^{-1} do movimento browniano multivariado que descreve a evolução de X é utilizada como um proxy para estimar a correlação evolutiva entre as variáveis componentes de Y. Este modelo foi inspirado pelo modelo limiar filogenético (Felsenstein 2005).

Para calcular a função de verossimilhança para esse modelo, consideramos uma extensão dos dados de modo que $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, em que $\mathbf{Y} = (Y_0, ..., Y_N)$ são os valores da variável D-dimensional de interesse Y observados nos N nós externos da filogenia (amostra), e $\mathbf{X} = (X_0, ..., X_N)$ são os valores

da variável latente D-dimensional X nos mesmos nós. O movimento browniano ao longo da árvore F que descreve a evolução de X é um processo já bem explorado na literatura (Felsenstein, 1988), e sua densidade $P(\mathbf{X}|\Sigma^{-1}, F)$ pode ser calculada por meio de um algoritmo iterativo que computa uma série de convoluções de distribuições normais D-variadas ao longo das arestas de F. Desse modo, temos

$$P(X, Y|F, \Sigma^{-1}) = P(X|F, \Sigma^{-1})P(Y|X).$$

No caso de variáveis Y binárias, definimos $P(X|Y)$ como

$$P(Y|X) = \prod_{i=1}^{N}\prod_{j=1}^{D} \left(\mathbf{I}(y_{i,j} = 1)\mathbf{I}(x_{i,j} > 0) + \mathbf{I}(y_{i,j} = 0)\mathbf{I}(x_{i,j} \leq 0)\right),$$

em que $\mathbf{I}(A)$ é a função indicadora de A, e $x_{i,j}$ e $y_{i,j}$ são a j-ésima componente das respectivas variáveis no nó i. Ou seja, em cada coordenada, temos $Y = 1$ se a variável latente é maior do que zero, e $Y = 0$ caso contrário.

Quando Y é contínuo, tomamos $Y = X$, fixando o valor da variável latente nos nós externos. Já quando Y é uma variável categórica com k estados não ordenados, então a uma entrada de Y correspondem $k - 1$ variáveis latentes em X. O valor observado $y_{i,j}$ na componente j da observação i é determinado pela maior das variáveis latentes correspondentes $\{x_{i,j'}, \cdots x_{i,j'+k-2}\}$ de modo que a função link é dada por

$$y_{ij} = g(x_{i,j'}, \cdots x_{i,j'+k-2}) = \begin{cases} s_1 \quad se \quad 0 = sup(0, x_{i,j'}, \cdots x_{i,j'+k-2}) \\ s_l \quad se \quad x_{i,l} = sup(0, x_{i,j'}, \cdots x_{i,j'+k-2}), \end{cases}$$

em que, sem perda de generalidade, tomamos o primeiro estado s_1 como o estado de referência. Neste caso

$$P(Y|X) = \prod_{i=1}^{N}\prod_{j=1}^{D} \left(\mathbf{I}(y_{i,j} = g(x_{i,j'}, \cdots x_{i,j'+k-2})\right).$$

Também podemos naturalmente considerar a extensão em que alguns componentes de Y são discretos e outros contínuos.

Inferência nesse modelo é feita em uma perspetiva Bayesiana, de modo que calculamos a distribuição à posteriori como

$$P(\Sigma|X, Y, F) \propto P(X, Y|F, \Sigma^{-1})P(\Sigma) = P(Y|X)P(X|F, \Sigma^{-1})P(\Sigma),$$

em que utilizamos a distribuição Whishart para distribuição à priori $P(\Sigma)$. Para fazer inferência baseada nesse modelo utilizamos um algorítimo de MCMC.

Para estimara o valor da variável de interesse Y no ancestral comum (raiz da filogenia), consideramos a extensão $\mathbf{Z}^* = (\mathbf{Y}, \mathbf{X}^*)$, em que $\mathbf{X}^* = (X_0, ..., X_{2N-1})$ são os valores da variável latente D-dimensional X em todos os nós da árvore. O algoritmo de MCMC é utilizado para obter a distribuição à posteriori de X_{2N-1}, e a função de ligação $g(X)$ é utilizada para recuperar os valores correspondentes de Y.

3 Resultados

Aplicação

Consideramos um conjunto de dados de 41 diferentes espécies de morcegos cedido por colaboradores (dados ainda não publicados). Os dados consistem de uma árvore filogenética que relaciona as espécies e, para cada espécie, informações sobre o número de dentes, variando entre 20 e 36. Além disso temos dados sobre os hábitos alimentares das espécies, divididos em k=6 categorias frugívoro (11 espécies), insetívoro (10), onívoro (6), carnívoro (3), nectarívoro (10) e hematófago (3). Como não há ordenamento inerente entre estas categorias, empregamos $k - 1 = 5$ dimensões da variável latente X para determinar o hábito alimentar e uma para modelar o número de dentes. Consideramos o hábito frugívoro como o referencial, e realizamos a análise no BEAST para estimação de correlações e reconstrução da raiz.

Tomamos como critério de significância para uma correlação que o seu intervalo de credibilidade 95% (IC) não inclua o zero. Isto é equivalente à probabilidade à posteriori de a covariância ser positiva (sig) ser superior a 0.975 ou inferior à 0.025. Apenas duas das entradas da matriz Σ foram consideradas significativas, as respectivas estimativas para a correlação e valor de sig estão apresentados na primeira linha da Tabela 1. Na reconstrução do número de dentes no ancestral comum, a média a posteriori foi 32.5 com IC de [29.8; 34.3]. Já o hábito alimentar estimado para o ancestral comum, com uma probabilidade à posteriori de 0.765, foi insetívoro.

A Figura 2 apresenta a reconstrução do número de dentes, estimada pela média da distribuição à posteriori, e do hábito alimentar, segundo o estado com maior probabilidade à posteriori, em toda a árvore filogenética destas espécies. Notamos que embora boa parte da evolução destas espécies provavelmente tenha ocorrido no estado frugívoro, há alta probabilidade à posteriori que o ancestral comum (raiz no centro da figura) seja insetívoro.

Análise de Sensibilidade

Notamos que o modelo de variável latente, no caso de variáveis com múltiplos estados não ordenados não é perfeitamente simétrico em relação a todos os estados. O estado de referência tem uma probabilidade à priori superior aos outros estados quando $k = 6$. Além disso, a interpretação de corre-

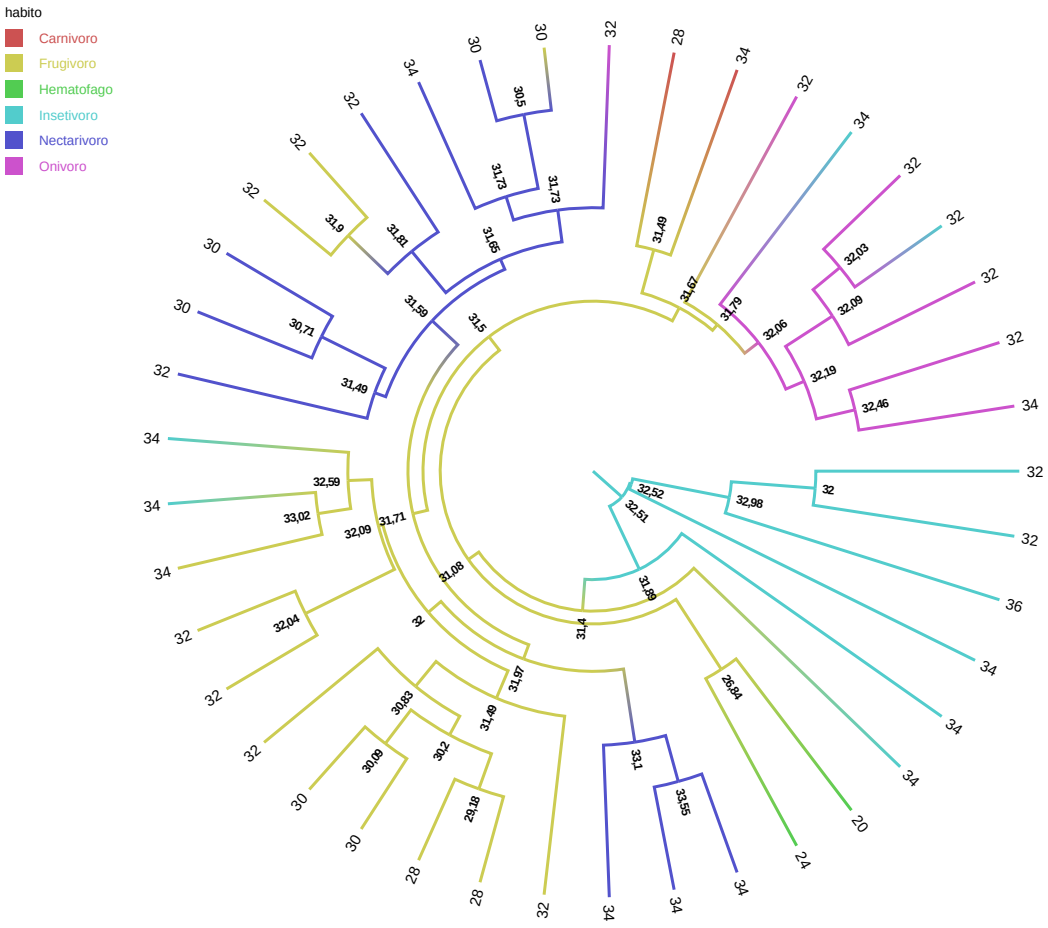


Figura 2: Reconstrução de hábito alimentar (representados por cores) e número de dentes (representados por números na figura) sobre a árvore filogenética para as 41 espécies de morcego consideradas.

Tabela 1: Correlações relevantes entre os hábitos alimentares e o número de dentes

	Dentes×Insetívoro		Dentes×Hematófago	
	Cor	Sig	Cor	Sig
Frugívoro	0.461	0.979	−0.501	0.007
Insetívoro	Referência		−0.575	0.002
Onívoro	0.476	0.960	−0.578	0.145
Nectarívoro	0.471	0.964	−0.548	0.012
Hematófago	0.533	0.994	Referência	

Tabela 2: Estimação do número de dentes e do hábito alimentar dos morcegos no ancestral comum situado na raiz da filogenia.

	Dentes		Hábito Alimentar						
				Probabilidade à Posteriori					
Referência	Média	MAP	MAP	Frugívoro	Insetívoro	Onívoro	Carnívoro	Nectarívoro	Hematófago
Frugívoro	32.502	31.403	Frugívoro	0.122	0.765	0.055	0.012	0.044	0.001
Insetívoro	32.476	30.673	Frugívoro	0.053	0.837	0.039	0.014	0.051	0.006
Onívoro	32.413	32.059	Insetívoro	0.062	0.759	0.123	0.009	0.043	0.003
Carnívoro	32.505	32.737	Insetívoro	0.060	0.794	0.057	0.032	0.053	0.003
Nectarívoro	32.449	31.935	Nectarívoro	0.044	0.757	0.057	0.012	0.125	0.004
Hematófago	32.473	33.529	Insetívoro	0.059	0.821	0.050	0.016	0.039	0.014

lações para este estado é indireta. Assim, buscamos responder à seguinte questão: A escolha do estado de referência afeta a estimação? Para tanto, repetimos a análise deste conjunto de dados considerando os outros estados como referenciais. A Tabela 1 apresenta as estimativas de correlações significativas na análise original, considerando os diferentes estados como referencial. A tabela apresenta as estimativas à posteriori para a correlação (cor) e o valor de sig. Notamos que tanto as estimativas de correlação quanto os valores de sig em geral são consistentes para os diferentes modelos. Uma exceção é a estimativa da correlação entre dentes e insetívoro quando o estado de referência é hematófago. Em todos os outros modelos, há correlação importante entre dentes e hematófago. Como neste caso não há uma variável latente específica ligada ao estado hematófago, o efeito desta correlação neste modelo acaba sendo manifestado em todos os outros estados.

Já a Tabela 2 apresenta as reconstruções para a raiz da filogenia em cada um destes casos. Para o número de dentes, apresentamos dois métodos de estimação a média à posteriori (Média) e o máximo a posteriori (MAP). Notamos que as estimativas pela média são muito mais consistentes considerando os diferentes referenciais. Para o hábito alimentar, apresentamos as estimativas por MAP e considerando a probabilidade à posteriori de cada estado. Pelo método do MAP, notamos variação na estimativa para a riz. Já quando a probabilidade à posteriori é considerada, concluímos que o estado da raiz é insetívoro com probabilidade superior a 0.7 para todos os referenciais. Assim, observamos que o método MAP parece não ser robusto para este tipo de estimação. Para os outros métodos, notamos que a escolha do estado de referência tem pouquíssimo efeito sobre a estimação da raiz.

4 Conclusões

Neste trabalho utilizamos a análise de um conjunto de dados de morcegos para verificar o comportamento da estimação das correlações evolutivas e reconstrução de caractere ancestral com o modelo filogenético de variável latente. Ao realizar a análise de dados considerando diferentes estados de referência para a variável com múltiplos estados não ordenados, percebemos que tanto as estimações de correlações quanto as do valor da variável na raiz aparentam ser robustas à escolha do referencial. Este

é um resultado importante pois destaca a confiança das estimativas obtidas, independente da escolha de estado de referência, que frequentemente é feita de forma arbitrária.

Referências

[1] CYBIS, G.B.; SINSHEIMER, J.S.; BEDFORD, T.; MATHER, A.E., LEMEY; P. and SUCHARD, M.A. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model.*The Annals of Applied Statistics* , v. 9(2), p969-991, 2015.

[2] DRUMMOND AJ; RAMBAUT A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, v. 7, p. 214, 2007.

[3] FELSENSTEIN J. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, v. 1, p. 445-71, 1988.

[4] FELSENSTEIN J. Using the Quantitative Genetics Threshold Model for Inferences Within and Between Species. *Philosophical Transactions of the Royal Society B*, v. 360, p. 1427-1434, 2005.

[5] KINGMAN, J. F. C. The coalescent. Stochastic processes and their applications, v. 13(3), p. 235-248, 1982.

Estimativa de casos de salmonelose humana atribuída às fontes de alimento de origem animal

Waldemir Santiago Neto, Luís Gustavo Corbellini[1]
Vanessa Bielefeldt Leotti[2]
Tine Hald[3]

Resumo: Tem se estimado que alimentos contaminados estejam relacionados com diversas doenças infecciosas e sejam responsáveis por 2,2 milhões de mortes ao redor do mundo anualmente. *Salmonella enterica* é considerada uma das principais causas de gastroenterites e bacteremias e a maioria de seus subtipos é encontrada em animais de sangue quente. Dados do Brasil apontam que as salmonelas são as principais causas de toxinfecção alimentar. A fim de obter compreensão da dinâmica de infecções por salmonela em humanos, um modelo bayesiano comparando a ocorrência de sorovares de *Salmonella* em animais e humanos foi utilizado para atribuir casos de salmonelose a frangos de corte, perus, porcos, galinhas poedeiras e surtos no Rio Grande do Sul (RS). Dados de salmonela para animais e seres humanos, cobrindo o período de 2000 a 2015, foram obtidos principalmente de estudos e relatórios publicados pela Secretaria de Vigilância em Saúde do Ministério da Saúde. A disponibilidade de fontes de alimento para consumo foi derivada dos dados de produção do Instituto Brasileiro de Geografia e Estatística. A principal fonte de salmonelose humana no RS foi estimada como sendo galinhas poedeiras, com 92,1% [3963 casos, intervalo de credibilidade de 95% (ICr95%) 3734-4159] de casos, seguido de 5,6% atribuídos a suínos de fora do RS (242 casos, ICr 95% 122-409). dos quais foi causada por *S. Enteritidis*. Este trabalho possibilita destacar diferenças na epidemiologia da *Salmonella*, foco de vigilância e hábitos alimentares no estado.

Palavras-chave: *Inferência bayesiana, salmonelose, Monte Carlo via Cadeias de Markov, avaliação de risco quantitativa, vigilância.*

[1] UFRGS - Universidade Federal do Rio Grande do Sul, Faculdade de Veterinária. Email: wal_sanet@*hotmail.com*
[2] UFRGS - Universidade Federal do Rio Grande do Sul, Instituto de Matemática e Estatística. Email: vleotti@*gmail.com*
[3] DTU – Universidade Técnica da Dinamarca

1 Introdução

A fim de se priorizar intervenções efetivas em sanidade alimentar é crucial determinar questões no âmbito da Saúde Única (do termo em inglês, *One Health*). Por exemplo, no que tange o impacto de diferentes doenças na saúde pública. Tem se estimado que alimentos contaminados estejam relacionados com diversas doenças infecciosas e sejam responsáveis por 2,2 milhões de mortes ao redor do mundo anualmente (WHO, 2007). *Salmonella enterica* é considerada uma das principais causas de gastroenterites e bacteremias (SCALLAN et al., 2011, HENDRIKSEN et al., 2011) e a maioria de seus subtipos é encontrada em animais de sangue quente. Faz-se necessário identificar as fontes das doenças e suas rotas de transmissão.

O termo 'fonte' tem seu significado de acordo com a etapa do caminho de transmissão, como por exemplo, na origem, os animais reservatórios (frangos, suínos, etc.), no processamento, por meio de veículos ou exposições cruzadas, até o destino final, incluindo diferentes alimentos específicos de origem animal. Embora as primeiras fontes sejam reconhecidas (BAKER et al. 2007; O´REILLY et al. 2007), a transmissão de *Salmonella* aos humanos ocorre na última, principalmente através do consumo e manipulação de alimentos contaminados (ACHA E SZYFRES, 2001). Os alimentos implicados são comumente carnes de suíno, gado e frango, produtos lácteos e ovos. Há evidências científicas de transmissão de cepas de reservatórios animais através da cadeia de alimentos à população humana (NEWELL et al. 2010).

Na Europa, a predominância de sorovares variou entre países em todas as fontes animais (carcaças de frango, linfonodos de suínos, poedeiras e perus) (EFSA, 2010). Em suínos, os sorovares isolados predominantemente foram *S.* Typhimurium e *S.* Derby. Em poedeiras, *S.* Enteritidis e *S.* Infantis (DE KNEGT et al., 2015a). No Rio Grande do Sul, *S.* Enteritidis também é o sorovar mais isolado em frangos de corte (RIBEIRO et al. 2007), ao passo que em suínos e seus derivados cárneos, *S.* Typhimurium, *S.* Panama e *S.* Bredeney foram isolados (CASTAGNA et al., 2004). Dado que há uma distribuição heterogênea de diferentes sorovares de *Salmonella* spp. em reservatórios animais distintos, assume-se que o risco de contaminação pela população se dá em função desta prevalência e pela quantidade de alimentos consumida. Além disto, há o pressuposto de que tanto os sorovares possuem características intrínsecas distintas de patogenicidade quanto os alimentos possuem capacidades díspares de veicular tais microrganismos. O conhecimento microbiológico em diferentes pontos da cadeia de alimentos servirá para o direcionamento de medidas de mitigação de riscos de transmissão de doenças transmitidas por alimentos.

Existem duas abordagens diferentes para a inferência estatística, as quais têm diferentes concepções a bases filosóficas e poderão levar a resultados distintos (DOHOO et al., 2009), as inferências clássica e Bayesiana. A análise Bayesiana tem ganhado popularidade recentemente, e tem sido aplicada a problemas complexos em epidemiologia veterinária como avaliação de risco (HALD et al., 2004),

comparação de testes diagnósticos sem padrão-ouro (BRANSCUM et al., 2005), e na análise de dados hierárquicos (DOHOO et al., 2001).

A metodologia Bayesiana deve seu nome ao papel fundamental do uso do teorema de Bayes. Na lógica Bayesiana, incertezas são atribuídas aos parâmetros, modelados por distribuições, enquanto os dados amostrados são mantidos como quantidades fixas uma vez coletados. Antes da coleta de qualquer dado, o conhecimento sobre os parâmetros desconhecidos de um problema é expresso na distribuição *a priori* para os parâmetros. Assim que coletados os dados, a distribuição *a priori* e os dados são combinados para gerar a distribuição *a posteriori* para os parâmetros. Esta, por sua vez, resume o conhecimento a respeito dos parâmetros depois de observar os dados.

Este projeto proporciona uma avaliação de risco microbiológica de comprovada importância para a tomada de decisões frente à segurança dos alimentos, transformando em informação os dados coletados e descritos de maneira sistemática. Prevê-se que este esforço permitirá que políticos e outras partes interessadas definam as prioridades adequadas, baseadas em evidências na área da segurança alimentar.

2 Materiais e Métodos

O presente estudo visa utilizar os dados secundários de prevalência em quatro reservatórios animais, de dentro e fora do estado (“importados”, os quais são transportados para dentro), bem como dados de investigação de surtos e da vigilância sanitária de hospitais vinculados ao *Global Salmon Surveillance* (GSS). Além destas fontes, informações de produção de produtos de origem animal disponibilizadas pelo Instituto Brasileiro de Geografia e Estatística foram acessadas como estimativa de risco de exposição.

A coordenação estadual alimentou um banco paralelo ao SINAN em Excel, desde 2000. Entre 2000 e 2012, foram notificados 2371 surtos de DTA, afetando 28.401 pessoas que adoeceram e causando quatro óbitos (FIGUEIREDO et al., 2013). Dos surtos notificados, 1492 (62,9%) foram investigados e destes, 979 (65,6%) confirmados. *Salmonella* spp. desponta como a maior causadora de surtos alimentares até 2011. Desde 2007, o VE-DTA implantou o GSS no Estado, que atende a compromissos internacionais com a Organização Mundial de Saúde (OMS) e Organização Pan-Americana de Saúde (OPAS). Desde então, 11 hospitais com Núcleo de Vigilância Epidemiológica Hospitalar (NVEH) fazem parte do programa, os quais remetem suas amostras às sessões de microbiologia e bacteriologia do Instituto de Pesquisas Biológicas do Laboratório Central do Estado (IPB-Lacen) da Fundação Estadual de Produção e Pesquisa em Saúde (FEPPS/RS). No Sistema de Informações Hospitalares (SIH/2011) consta que 3200 pessoas foram internadas por diarreia e gastroenterite de origem infecciosa e fonte presumível.

A fim de obter uma compreensão da distribuição de ocorrência de infecções por *Salmonella* em humanos, foi abordada a metodologia de Hald et al. (2004), os quais compararam a ocorrência de

sorovares de *Salmonella* spp. em animais e humanos por modelagem Bayesiana. O princípio do modelo de atribuição a fontes por microbiologia consiste em comparar o número de casos humanos causados por diferentes sorovares de um patógeno com a distribuição dos mesmos sorovares em diferentes alimentos de origem animal. O modelo é construído com coletâneas de isolados relacionados no tempo e no espaço de diversas fontes de alimentos e de humanos e o montante de alimentos disponíveis para a população é considerado (HALD et al., 2004; DE KNEGT et al., 2015b).

O escopo da inferência Bayesiana tem aumentado consideravelmente pela invenção e avanços recentes de ferramentas baseadas em simulação para inferência estatística, especialmente o método de simulação de Monte Carlo via cadeias de Markov (MCMC). A análise de muitos dos modelos complexos com abordagem Bayesiana é baseada em métodos MCMC (DOHOO et al., 2009). A maioria das análises Bayesianas requerem softwares especializados, e a escolha pode variar entre os programas livres WinBUGS, desenvolvido pelo *Medical Research Council Biostatistics Unit*, de Cambridge, e o pacote rjags do R. BUGS é a sigla em inglês para *Bayesian analysis using Gibbs sampling*, ao passo que JAGS refere-se a *Just Another Gibbs Sampler*, que é um tipo particular de algoritmo MCMC.

A equação utilizada para estimar o número esperado de casos humanos por alimento e sorovar é definida a seguir:

a_j ~ Exponencial (0.002),

q_i ~ Uniforme (0, 100),

o_i ~ Poisson ($\Sigma_j \lambda_{ij}$),

$$\lambda_{ij} = M_j p_{ij} q_i a_j$$

onde λ_{ij} é o número esperado de casos do sorovar *i* do alimento *j*; M*j* a quantidade de alimento *j* disponível para consumo; p_{ij}, a prevalência do sorovar *i* no alimento *j*; q_i, o fator relacionado ao sorovar *i*; e a_j, o fator relacionado ao alimento *j*.

O modelo atribui casos domésticos esporádicos a alimentos de origem animal. Um caso esporádico é definido como um sujeito para qual não foi possível associar a um surto de DTA reconhecido. Casos relacionados a surtos são adicionados aos resultados finais do modelo, e atribuídos ao alimento implicado no surto, caso conhecido. Caso contrário, estes são considerados surtos sem fonte conhecida. Como os subtipos de *Salmonella* são distribuídos por clones entre os hospedeiros animais (HALD et al., 2004), o modelo atribui casos a reservatórios animais. Isto significa assumir que casos causados por carne de porco são atribuídos a suínos, ovos a poedeiras, carne de frango a frangos, e assim por diante. Mas, caso uma carne de porco seja contaminada durante o preparo com um subtipo originalmente encontrado em frangos, os casos resultantes são atribuídos a frangos, não suínos.

Devido ao problema de sobreparametrização do modelo original (HALD et al., 2004), alternativas foram adotadas com base nos trabalhos de Mullner et al. e David et al. (MULNNER et al, 2009; DAVID et al. 2013).

3 Resultados

A principal fonte de salmonelose humana no RS foi estimada como sendo galinhas poedeiras (ou seja, ovos), com 92,1% [3963 casos, intervalo de credibilidade de 95% (ICr95%) 3734-4159] de casos, seguido de 5,6% atribuídos a suínos de fora do RS (242 casos, ICr 95% 122-409).

O sorovar mais importante contribuindo a salmoneloses humanas dos reservatórios animais foi *S.* Enteritidis (3265 casos, ICr 95% 3154-3378). Dentre todas as infecções por *S.* Enteritidis, 99,2% (3240 casos, ICr 95%) foram atribuídas a galinhas poedeiras, enquanto 77,3% de S. Typhimurium teve origem em suínos de fora do RS (187 casos, CrI 95% 87-321). Dentre os sorovares de importância intermediária estiveram S. Panama e S. Infantis.

4 Discussão

Este estudo representa a primeira tentativa de conduzir atribuição de fontes de salmonelose humana no Rio Grande do Sul. Os resultados sugerem que galinhas poedeiras foram a fonte mais importante no estado no período estudado, sendo responsáveis por praticamente a totalidade de infecções por *Salmonella*. Suínos criados em Santa Catarina tiveram uma importância relativamente menor. Outras fontes contribuíram com menos de 1% cada. A identificação das fontes mais importantes de salmonelose é uma etapa para a priorização de ações e intervenções direcionadas a reduzir doenças de importância em saúde pública. Estas estimativas de atribuição levaram em conta o montante de alimento produzido e transportado entre alguns estados e o RS (por exemplo, aqueles que possuíam dados de prevalência nos reservatórios estudados). O pressuposto foi que tais dados refletem o fluxo real de alimentos e a exposição consequente no estado.

5 Conclusão

O modelo apresentou estimativas em concordância com os casos observados.

Referências

[1] ACHA, P. N., SZYFRES, B. *Zoonosis y enfermedades transmisibles comunes al hombre y a los animales*. 3ª ed. Washington: Organización Panamericana de la Salud. p. 233-238. (Publ. Cient. nº 580), 2011.

[2] BAKER, M. G., THORNLEY, C. N., LOPEZ, L. D., GARRETT, N. K., NICOL, C. M. A recurring salmonellosis epidemic in New Zealand linked to contact with sheep. *Epidemiology and Infection* 135, 76-83, 2007.

[3] BRANSCUM, A. J., GARDNER, I. A., JOHNSON, W. O. Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine*, 68(2), 145-163, 2005.

[4] CASTAGNA, S. M. F., SCHWARZ, P., CANAL, C. W., CARDOSO, M. R. I. Prevalência de suínos portadores de *Salmonella* sp. ao abate e contaminação de embutidos tipo frescal. *Acta Scientiae Veterinariae*. 32(2): 141- 147, 2004.

[5] DAVID, J. M., SANDERS, P., BEMRAH, N., GRANIER, S. A., DENIS, M., WEILL, F. X., GUILLEMOT, D., WATIER, L. Attribution of the French human Salmonellosis cases to the main food-sources according to the type of surveillance data. *Preventive Veterinary Medicine*, 110 (1), 12-27, 2013.

[6] DE KNEGT, L. V., PIRES, S. M., HALD, T. Using surveillance and monitoring data of different origins in a Salmonella source attribution model: a European Union example with challenges and proposed solutions. *Epidemiology and Infection*, 143, 1148 – 1165, 2015a.

[7] DE KNEGT, L. V., PIRES, S. M., HALD, T. Attributing foodborne salmonellosis in humans to animal reservoirs in the European Union using a multi-country stochastic model. *Epidemiology and Infection*, 143, 1175 – 1186, 2015b.

[8] DOHOO, I. R., MARTIN, S. W., STRYHN, H. *Veterinary Epidemiologic Research*, pp. 589. VER Incorporated, Charlottetown, 2009.

[9] DOHOO, I. R., TILLARD, E., STRYHN, H., FAYE, B. The use of multilevel models to evaluate sources of variation in reproductive performance in dairy cattle in Reunion Island. *Preventive Veterinary Medicine*, 50(1), 127-144, 2001.

[10] EFSA – European Food Safety Authority. The community summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in the European Union in 2008. *EFSA Journal*; 8: 1496, 2010.

[11] FIGUEIREDO, D., TIM, L. N., CECCONI, M. C. P., BOTH, J. M. C., SOEIRO, M. L. T., RAMOS, R. C., HAAS, S., LONGARAY, S. M. Programa de Vigilância Epidemiológica das Doenças de Transmissão Hídricas e Alimentares – VE-DTHA. *Boletim Epidemiológico do Centro Estadual de Vigilância em Saúde*, Secretaria da Saúde, v. 15 (3), 5 – 8, 2013.

[12] HALD, T., VOSE, D., WEGENER, H.C., KOUPEEV, T. A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. Risk Analysis, 24 (1), 255 – 269, 2004.

[13] HENDRIKSEN, R. S., VIEIRA, A. R., KARLSMOSE, S., LO FO WONG, D. M., JENSEN, A. B., WEGENER, H. C., AARESTRUP, F. M. Global monitoring of Salmonella serovar distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: results of quality assured laboratories from 2001 to 2007. *Foodborne pathogens and disease*, 8(8), 887-900, 2011.

[14] MULLNER, P., JONES, G., NOBLE, A., SPENCER, S.E., HATHAWAY, S., FRENCH, N.P. Source attribution of food-borne zoonoses in New Zealand: a modified Hald model. *Risk Analysis* 29 (7): 970-84, 2009.

[15] NEWELL, D. G., KOOPMANS, M., VERHOEF, L., DUIZER, E., AIDARA-KANE, A. SPRONG, H., OPSTEEGH, M., LANGELAAR, M., THREFALL, J., SCHEUTZ, F., VAN DER GIESSEN, J., KRUSE, H. Food-borne diseases — The challenges of 20 years ago still persist while new ones continue to emerge. *International Journal of Food Microbiology*, 139, S3–S15, 2010.

[16] O'REILLY, C. E., BOWEN, A. B., PEREZ, N. E., SARISKY, J. P., SHEPHERD, C. A., MILLER, M. D., HUBBARD, B. C., HERRING, M., BUCHANAN, S. D., FITZGERALD, C. C., HILL, V., ARROWOOD, M. J., IAO, L. X., HOEKSTRA, R. M., MINTZ, E. D., LYNCH, M. F. A waterborne outbreak of gastroenteritis with multiple etiologies among resort island visitors and residents: Ohio, 2004. *Clinical Infectious Diseases* 44, 506-512, 2007.

[17] RIBEIRO, A. R., KELLERMANN, A., SANTOS, L. R., BESSA, M. C., NASCIMENTO, V.P. *Salmonella* spp. in raw broiler parts: occurrence, antimicrobial resistance profile and phage typing of the *Salmonella* Enteritidis isolates. *Brazilian Journal of Microbiology*, vol.38, n.2, pp. 296-299, 2007.

[18] SCALLAN, E., HOEKSTRA, R. M., ANGULO, F. J., TAUXE, R. V., WIDDOWSON, M. A., ROY, S. L., JONES, J. L., GRIFFIN, P. M. Foodborne illness acquired in the United States—major pathogens. *Emerging Infectious Diseases*, 17: 7-15, 2011.

[19] WORLD HEALTH ORGANIZATION. *The world health report 2007*. Disponível em: http://www.who.int/whr/2007/en/index.html. Geneva, Suíça: World Health Organization. Acessado em 7 de Julho 2015.