VOL 2 N.1 - 2017

ANAIS DO SER: II Seminário Internacional de Estatística com R

ISSN: 2526-7299

Luciane Ferreira Alcoforado, Orlando Celso Longo, José Rodrigo de Moraes e Ariel Levy
UNIVERSIDADE FEDERAL FLUMINENSE

VOL 2 N.1 - 2017

SOBRE O EVENTO

O Seminário Internacional de Estatística com R foi uma iniciativa pioneira no Brasil, iniciada na Universidade Federal Fluminense, em Niterói, no ano de 2016. O software R vem de encontro as necessidades dos pesquisadores, das Universidades, do setor público e privado, por ser gratuito e contar com uma rede mundial de colaboradores. Há disponível diversos canais de articulação entre os *stakeholders* como é o caso da equipe de Estatística com R da UFF, diversos blogs e grupos em nível nacional e internacional, unindo-os em torno da temática do uso e aprendizado da linguagem R.

Espera-se que o evento traga uma relevante contribuição para a formação acadêmica e profissional quanto a técnicas e pacotes utilizados no desenvolvimento de pesquisas e propostas inovadoras em todos os campos de sua aplicação, Engenharias, Ciências Sociais Aplicadas, Saúde e áreas afins.

Público Alvo: Pesquisadores, professores, estudantes e profissionais do mercado interessados em compartilhar conhecimentos, aprender e se atualizar no uso da linguagem R.

FICHA TÉCNICA

Realização
UNIVERSIDADE FEDERAL FLUMINENSE

Comissão Organizadora
Luciane Ferreira Alcoforado - UFF – Presidente
Ariel Levy – UFF Vice-Presidente
Orlando Celso Longo - UFF
José Rodrigo de Moraes - UFF
Alex Laier Bordignon - UFF
Fabiano dos Santos Souza - UFF

Editoração dos Anais
Luciane Alcoforado

Apoio
Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro
Programa de Pós-Graduação em Engenharia Civil
Instituto de Matemática e Estatística
Escola Nacional de Ciências Estatísticas
Escola de Engenharia
Instituto de Matemática Pura e Aplicada
Programa de Pós-Graduação em Administração
Núcleo de Pesquisas, Informações e Políticas Públicas DATAUFF
Sociedade Brasileira de Matemática Pura e Aplicada
Pró-Reitoria de Extensão
Pró-Reitoria de Pesquisa e Inovação
Uniteve
Núcleo de Estudos em Biomassa e Gerenciamento de Águas

Medalhas
SBBNET

Realizado entre os dias 23 e 24 de maio de 2017, na Universidade Federal Fluminense, Niterói-RJ
Informações
ser.uff.br@gmail.com

ISBN:978-85-94029-02-7
ISSN:2526-7299

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Volume 2 – N.01 – ANO 2017

Artigos completos submetidos e aprovados pela comissão científica

COMISSÃO CIENTÍFICA

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e do Instituto de Computação da Universidade Federal Fluminense

S471 Seminário Internacional de Estatística com R: the world of big data analysis (2. : 2017 : Niterói, RJ).
Anais ... / II Seminário Internacional de Estatística com R : the world of big data analysis ; organizadores Luciane Ferreira Alcoforado, Ariel Levy, Orlando Celso Longo, José Rodrigo de Moraes, Alex Laier Bordignon, Fabiano dos Santos Souza. – Niterói : PROPPI, 2017.
2 v.
Conteúdo: n. 1. Artigos completos aprovados – n. 2. Resumos simples aprovados.
Evento realizado entre os dias 23 e 24 de maio de 2017.

1. Desenvolvimento de software. 2. Estatística. 3. Inovação tecnológica I. Alcoforado, Luciane Ferreira (org.). II. Levy, Ariel (org.). III. Longo, Orlando Celso (org.). IV. Moraes, José Rodrigo de (org.). V. Bordignon, Alex Laier (org.). VI. Souza, Fabiano dos Santos. (org.). VII. Título.

CDD 005.1 (21. ed)

ISBN:978-85-94029-02-7
ISSN:2526-7299

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Sumário

PROGRAMAÇÃO GERAL

PROGRAMAÇÃO

23 Maio/2017 (Terça-feira)

10h00 às 12h00	Recepção aos palestrantes	
12h30 às 13h30	Credenciamento (Auditório do NAB – Praia Vermelha)	
13h00 às 13h30	Espaço Blog: Paixão por Dados - Sillas Gonzaga	
13h30 às 14h30	Mesa de Abertura	
14h30 às 15h30	Conferência: Avaliação, comparação e seleção de modelos de previsão em R usando o pacote performance Estimation Prof. Luís Torgo - Un. do Porto-PT	
15h30 às 16h00	Coffee Break	
16h00 às 17h00	Conferências Time Series: Time Series with R Prof. Manuel Febrero - Un. Santiago de Compostela-ES Brazilian Economic Time Series (BETS) package Prof. Pedro Ferreira - FGV/IBRE	
17h00 às 18h30	Mesa Redonda: The World of Big Data Analysis	Mediador – Ariel Levy Eduardo Camilo – PPGAD/UFF Orlando Longo – PPGEC/UFF Jorge Zubelli - IMPA E convidados
18h30 às 19h00	Espaço de Confraternização	Experimentação do R para iniciantes com equipe de monitores

24 Maio/2017 (Quarta-feira)

09h00 às 12h00	**Oficinas em Laboratório de Informática** O1: Gráficos no R com ggplot2 - Luciane Alcoforado-UFF O2: Introdução à Programação em R - Felipe Ribeiro-UNIRIO O3: Inteligência Artificial com R - Alex Laier-UFF **Mini Cursos em sala de aula** M1: Modelos Lineares Generalizados com R - J. Rodrigo-UFF M2: Análise Multivariada Aplicada com R - Ludmilla-UFF M3: Como e onde começar com o R - Ariel Levy-UFF M4: Teoria da Resposta ao item com R - Leandro-Cesgranrio M5: Letramento Estatístico com o R: possibilidades para a Educação Básica - Alexandre-UNIRIO e Fabiano-UFF M6: Desenvolvimento de Dashboards interativos com o R - Steven Ross-UNIRIO M7: Análise de Séries Temporais utilizando o pacote BETS - Pedro Ferreira-FGV/IBRE
13h30	Sessão Pôster e Comunicação Oral
15h30	Coffee Break
16h00 às 19h00	**Palestras de 30 minutos** P1: Praticando Data Science com R nos dados de criminalidade do Rio de Janeiro - Prof. Hélio Lopes-PUC P2: Comex Vis: visualizações interativas dos dados do comércio exterior brasileiro - Saulo Guerra-MDIC P3: O papel do R no ensino de economia e áreas correlatas - Vítor Wilher - Análise Macro P4: Data Analytics com R e Banco de Dados - SQL e NOSQL - Flávio Brito - Fundação CECIERJ P5: Aplicação dos modelos de regressão censurados em Estatísticas Públicas - Gustavo Rocha-ENCE P6: Explorando e modelando dados de temperatura máxima nas regiões sul e sudeste do Brasil - Renata Bueno-ENCE
19h00	Cerimônia de premiação melhor pôster e melhor artigo
19h15	Fechamento Oficial
19h30 às 20h15	Espaço de Confraternização/Network

Coordenação Geral: Luciane Alcoforado

Inscrições online - Vagas Limitadas

www.ser.uff.br ou http://ser2017.weebly.com

Público Alvo: Interessados na linguagem R e suas aplicações

Local das Palestras: Auditório do NAB - Praia Vermelha, Niterói - RJ

Siga-nos no facebook: www.facebook.com/eventoser.uff

Apoio: Parceria:

Discurso de Abertura do II SER – profa. Luciane Alcoforado

Boa tarde a todos, sejam muito bem-vindos ao II Seminário Internacional de Estatística com R, o II SER.

É com grande alegria que dou início a este importante evento que não seria possível sem o apoio e participação de um grupo de entusiastas e apaixonados pela análise de dados e que fazem uso da linguagem R.

Quem faz o SER são vocês, o desejo é de vocês em ir de encontro ao que o mundo do R pode proporcionar e a Comissão apenas organiza e apoia.

Confesso que não foi fácil organizar este II evento, especialmente por não ter tido o esperado apoio financeiro das agências de fomento que estavam com editais abertos. A FAPERJ foi a única agência de fomento a nos apoiar, porém ainda não nos repassou o financeiro.

Desse modo o evento só se concretizou pelo interesse e adesão dos participantes que se inscreveram e dos palestrantes que se prontificaram em dividir suas experiências e conhecimentos. A estes, meu especial agradecimento. Aos participantes que vieram da Bahia, do Distrito Federal, do Espirito Santo, de Goiás, de Minas Gerais, da Paraíba, do Paraná, do Pará, do Piauí, do Rio de Janeiro, do Rio Grande do Norte, do Rio Grande do Sul, de Santa Catarina, de São Paulo, de Portugal, da Espanha e da Itália.

Gostaria de agradecer ao NAB por nos apoiar mais uma vez cedendo este espaço que muito nos encanta. Agradeço também aos palestrantes internacionais que atravessaram o oceano para nos trazer novos conhecimentos, a todos que irão proferir palestras e ministrar minicursos e oficinas, aos que submeteram seus trabalhos (desejo sucesso, teremos uma sessão de premiação ao final), à sbbnet que nos oferta todas as medalhas de premiação. Aos professores e alunos da comissão

organizadora, em especial à Elizete do Datauff, a Noelli e Raquel da Faculdade de Turismo e Hotelaria, aos professores da Comissão Científica que foram eficazes em todas as etapas que precederam este momento. Em especial aos professores Orlando, Ariel, José Rodrigo, Alex Laier e Fabiano que atenderam aos meus chamados em todos os momentos críticos da etapa de Organização.

Meu agradecimento especial à nossa querida ENCE, parceira desde o início, à SBMAC que nos apoia pela primeira vez e ao IMPA que novamente nos apoia, em especial ao prof. Zubelli sempre pronto a nos indicar o melhor caminho. Também meu agradecimento ao Instituto de Matemática e Estatística, ao departamento de Estatística, à Escola de Engenharia, aos Programas de Pós-Graduação da Engenharia Civil e da Administração, às Pró-reitorias PROPPI, PROGRAD, PROEX e à UFF, essa grande instituição com enorme potencial de unir diversas áreas do conhecimento como é o caso que ora iremos presenciar.

Voltando um pouco no tempo percebo que tudo tem o momento certo de ocorrer, em 2015 iniciei o projeto deste evento, era eu e um pequeno grupo de estudantes de Estatística que tinham um desejo enorme de aprender e se aprimorar na linguagem R.

Foram inúmeros encontros, até a realização do primeiro evento em maio de 2016. E o tempo vai passando e os alunos vão se desenvolvendo de uma forma surpreendente. Com certeza este evento propicia um olhar da academia para o mercado e vice-versa. Como tem que ser, afinal este é o papel de uma instituição como a UFF e os demais apoiadores.

Lembro a vocês que em junho de 2015 era anunciado que o R se encontrava na 6ª. posição de um ranking de linguagens mais populares do mundo segundo a IEEE Spectrum; na realização do primeiro evento em maio de 2016 não podíamos imaginar que o R subiria este ranking, pensávamos que já estava no topo. E não é

que logo após o I SER, em junho de 2016 o R subia para a 5ª. posição! Vamos aguardar o próximo ranking.

Não é por acaso que a profissão de Estatístico vem se valorizando no mundo, acredito que o R contribui fortemente para que os profissionais possam desenvolver-se, utilizando-se desta ferramenta computacional imprescindível. E não apenas os Estatísticos, mas todos os profissionais que fazem uso da análise de dados.

Meus caros, desejo que vocês aproveitem tudo que este evento tem a oferecer, foi feito com muito carinho pensando principalmente na troca que iremos experimentar. Teremos espaço para contribuições dos participantes para aprimorarmos o evento nas próximas edições, por isso não deixem de responder ao questionário de avaliação do evento que enviaremos oportunamente por e-mail.

Observem ainda que o folder foi projetado para ser um marcador de livros com vários códigos do Rmarkdown.

Para encerrar minha fala, reforço meu agradecimento a todos os participantes e às autoridades presentes.

Um bom SER a todos! Muito Obrigada.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 12 - DETERMINAÇÃO DAS MEDIDAS DE DESEMPENHO DE UMA FILA *M/M/1* ATRAVÉS DE UMA ABORDAGEM BAYESIANA.

Nilson Luiz Castelucio Brito[1]
Celimar Reijane Alves Damasceno Paiva[2]
Pedro Humberto de Almeida Mendonça Gonzaga[3]
Rodrigo Fonseca Santana Costa[4]

Resumo

O uso de modelos estocásticos para análise de filas reais baseia- se na premissa fundamental de que os parâmetros são desconhecidos, precisando, portanto, ser estimados. Os modelos de simulação são utilizados para determinar estes parâmetros por se apresentar como uma técnica relativamente barata e eficiente quando executadas várias vezes. Este trabalho tem por objetivo utilizar o software R para estimar os parâmetros de desempenho de filas markovianas infinitas com um único servidor *M/M/1* utilizando uma abordagem Bayesiana. Sob o enfoque Bayesiano, deve-se obter distribuições a priori e a posteriori para os parâmetros de interesse. Foram simulados dados do número de clientes no sistema no momento da partida e feitas 5000 replicações Monte Carlo com amostras de tamanhos n=10, 100 e 200 para valores da intensidade de tráfego ρ=0.2, 0.5 e 0.9. O algoritmo mostrou-se robusto, pois foram obtidas estimativas muito próximas dos valores teóricos, principalmente quando se aumenta o tamanho da amostra.

Palavras-Chave: Filas, Inferência Bayesiana,Simulação.

Abstract

The use of stochastic models for the analysis of real queues is based on the fundamental premise that the parameters are unknown, and therefore need to be estimated. Simulation models are used to determine these parameters because they are presented as a relatively inexpensive and efficient technique when executed several times. The purpose of this work is to use R software to estimate the performance parameters of infinite Markovian queues with a single *M/M/1* server using a Bayesian approach. Under the Bayesian approach, one must obtain a priori and a posteriori distributions for the parameters of interest. We simulated data on the number of clients in the system at the time of departure and made 5000 Monte Carlo replicates with samples of sizes n = 10, 100 and 200 for traffic intensity values ρ = 0.2, 0.5 and 0.9. The algorithm was robust, since estimates were obtained very close to the theoretical values, especially when the sample size was increased

Keywords: Queues, Bayesian Inference, Simulation.

[1] Unimontes, casteluciobrito@gmail.com
[2] IFNMG, ceelimarreijane@yahoo.com.br
[3] Unimontes, pedrogonza13@gmail.com
[4] Unimontes, rodrigo-fsc@hotmail.com

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

Um sistema de filas pode ser resumidamente descrito como usuários chegando para receber um serviço e, devido à impossibilidade de atendimento imediato, são alocados em uma fila de espera. Uma fila não precisa ser necessariamente formada por pessoas, como em uma fila de banco, por exemplo. Ela pode ser formada por estações de trabalho tentando acessar uma rede de computadores. As principais características de uma fila são: o *processo de chegada*, que descreve como os usuários procuram o serviço; o *tempo de serviço*, a *disciplina de atendimento*, referente à maneira como os usuários recebem o serviço, sendo no caso mais comum o regime *FCFS,* first come, first served e a *capacidade do sistema*, que está associada à limitação física da "sala de espera", ou seja, diz respeito ao número de usuários que podem ali permanecer.

Uma forma simples de descrever um modelo de fila é através da notação de Kendall [8], cujo padrão é *A/B/X/Y/Z*, em que *A* indica a distribuição dos tempos entre chegadas, *B* indica a distribuição do tempo de serviço, *X* é o número de servidores em paralelo, *Y* é a restrição na capacidade do sistema e *Z*, a disciplina de atendimento. Por exemplo, a notação *M/D/2/∞/FCFS* indica um processo de fila com tempo entre chegadas exponencial, tempo de serviço determinístico, dois servidores em paralelo, sem restrição no tamanho máximo da capacidade do sistema e disciplina de fila "primeiro a chegar, primeiro a ser atendido". Quando são omitidos os símbolos *Y* e *Z* na notação de Kendall, entende-se que a fila tem capacidade infinita e disciplina *FCFS*. Por exemplo, a fila *M/M/1* tem chegadas exponenciais, serviço com distribuição exponencial, um único servidor, não há limite na capacidade do sistema e o atendimento é por ordem de chegada. Pode parecer estranho utilizar o símbolo *M* para a distribuição exponencial, em vez do usual *E*. A razão é para evitar confusão com E_k, símbolo utilizado para a distribuição *Erlang-k.* O símbolo *M* utilizado para designar a distribuição exponencial tem origem na falta de memória (memoryless) dessa distribuição.

Objetivo

Utilizar o software R para estimar os parâmetros de desempenho de uma fila *M/M/1*, através de uma abordagem bayesiana.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Material e Métodos

Filas *M/M/1*: Serão utilizadas distribuições exponenciais com parâmetros λ e μ para a taxa de chegada e taxa de atendimento, respectivamente. Considerando que a fila esteja em equilíbrio, a intensidade de tráfego ρ=λ/μ é menor do que 1, o que implica a condição λ<μ. Caso contrário, a fila "explode", isto é, o tamanho da fila cresce continuamente. Nestas condições, a distribuição do número de usuários no sistema é dada pela equação 1:

$$\boldsymbol{P(N=n)=p_n=\rho^n(1-\rho), 0<\rho<1, n\geq 0.} \qquad (1)$$

A equação (1) é a função de probabilidade de uma variável aleatória com distribuição geométrica.

Inferência bayesiana: O esquema de geração de dados é consistente com a distribuição definida pela equação (1). Suponha uma amostra do número de usuários no sistema dada por $x_1,\dots,x_n$. A função de verossimilhança é

$$\boldsymbol{L(\tilde{x}|\rho)=\prod_{i=1}^{n}\rho^{x_i}(1-\rho)=(1-\rho)^n\rho^{\sum x_i}.} \qquad (2)$$

O núcleo da função de verossimilhança é uma beta$(\boldsymbol{n+1;\sum_i x_i+1})$.

2.1 Distribuições a priori e a posteriori: Neste ponto, temos duas possibilidades de escolha para a distribuição a priori do parâmetro ρ. A primeira é a priori conjugada natural beta(a,b), visto que $0<\rho<1$. Neste caso, a posteriori tem função de probabilidade dada por:

$$\boldsymbol{\pi_{NC}(\rho|dados)=\frac{1}{B(a+\sum_i x_i;n+b)}\rho^{a+\sum_i x_i-1}(1-\rho)^{n+b-1}, 0<\rho<1} \qquad (3)$$

ou seja, $\boldsymbol{\pi_{NC}(\rho|dados)\sim beta(a+\sum_i x_i; n+b)}$.

A segunda priori possível para ρ é uma uniforme. Como $\rho\in(0,1)$, pode-se pensar em uma priori não informativa uniforme$(0,1)$. Entretanto, em muitas situações práticas, verifica-se que $c<\rho<d$, com $0<c<d<1$. Assim, pode-se trabalhar com a distribuição a priori uniforme truncada no intervalo (c,d). Neste caso, a posteriori tem função de probabilidade dada por:

$$\boldsymbol{\pi_{TU}(\rho|dados)=\frac{1}{B(c;d;\sum_i x_i+1;n+1)}\rho^{\sum_i x_i}(1-\rho)^n, 0<\rho<1,} \qquad (4)$$

ou seja, $\boldsymbol{\pi_{TU}\sim beta\ incompleta(c;d;\sum_i x_i+1;n+1)}$.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

A partir daí, todas as estimativas podem ser encontradas. Tomando-se a função de perda quadrática, temos que o estimador de Bayes (EB) é a esperança das distribuições. Temos, então, os seguintes estimadores de Bayes para a intensidade de tráfego ρ.

$\widehat{\boldsymbol{\rho}}_{\boldsymbol{NC}} = \frac{\boldsymbol{a}+\sum_i \boldsymbol{x_i}}{\boldsymbol{a}+\sum_i \boldsymbol{x_i}+\boldsymbol{n}+\boldsymbol{b}}$ $\quad (\mathbf{5})$ e $\widehat{\boldsymbol{\rho}}_{\boldsymbol{TU}} = \frac{\boldsymbol{B}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{2};\boldsymbol{n}+\boldsymbol{1})}{\boldsymbol{B}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{1};\boldsymbol{n}+\boldsymbol{1})}$ $\quad (\mathbf{6})$

Medidas de desempenho: A partir da distribuição estacionária do número total de usuários dada pela equação (1), podemos obter as medidas de desempenho da fila. Sejam N o número de usuários no sistema e $\boldsymbol{L} = \boldsymbol{E}(\boldsymbol{N}) = \sum_{\boldsymbol{n}=\boldsymbol{0}}^{\infty} \boldsymbol{n}.\boldsymbol{p_n}$ sua média. Temos:

$$\boldsymbol{L} = \frac{\boldsymbol{\rho}}{\boldsymbol{1}-\boldsymbol{\rho}} \qquad (\mathbf{7})$$

Sejam $\boldsymbol{N_q}$ o número de usuários na fila e $\boldsymbol{L_q} = \boldsymbol{E}(\boldsymbol{N_q}) = \boldsymbol{0}.\boldsymbol{p_0} + \sum_{\boldsymbol{n}=\boldsymbol{1}}^{\infty}(\boldsymbol{n}-\boldsymbol{1}).\boldsymbol{p_n}$ sua média. Temos:

$$\boldsymbol{L_q} = \frac{\boldsymbol{\rho^2}}{\boldsymbol{1}-\boldsymbol{\rho}} \qquad (\mathbf{8})$$

Os estimadores bayesianos para $\boldsymbol{L}$ e $\boldsymbol{L_q}$ são:

Utilizando a distribuição conjugada natural $\boldsymbol{beta}(\boldsymbol{a}+\sum_i \boldsymbol{x_i};\boldsymbol{n}+\boldsymbol{b})$, obtemos:

$\hat{\boldsymbol{L}}_{\boldsymbol{NC}} = \frac{\boldsymbol{a}+\sum_i \boldsymbol{x_i}}{\boldsymbol{n}+\boldsymbol{b}}$ $\quad (\mathbf{9})$ e $\hat{\boldsymbol{L}}_{\boldsymbol{q_{NC}}} = \frac{(\boldsymbol{a}+\sum_i \boldsymbol{x_i})^2}{(\boldsymbol{a}+\sum_i \boldsymbol{x_i}+\boldsymbol{n}+\boldsymbol{b})(\boldsymbol{n}+\boldsymbol{b})}$ $\quad (\mathbf{10})$

Utilizando a distribuição não informativa $\boldsymbol{beta\ incompleta}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{1};\boldsymbol{n}+\boldsymbol{1})$, obtemos:

$$\hat{\boldsymbol{L}}_{\boldsymbol{TU}} = \frac{\boldsymbol{B}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{2};\boldsymbol{n})}{\boldsymbol{B}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{1};\boldsymbol{n}+\boldsymbol{1})} = \widehat{\boldsymbol{\rho}}_{\boldsymbol{TU}} \qquad (\mathbf{11})$$

$$\hat{\boldsymbol{L}}_{\boldsymbol{q_{TU}}} = \left[\frac{\boldsymbol{B}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{2};\boldsymbol{n})}{\boldsymbol{B}(\boldsymbol{c};\boldsymbol{d};\sum_i \boldsymbol{x_i}+\boldsymbol{1};\boldsymbol{n}+\boldsymbol{1})}\right]^{\boldsymbol{2}} = \widehat{\boldsymbol{\rho}}_{\boldsymbol{TU}}^{\boldsymbol{2}} \qquad (\mathbf{12})$$

Resultados e Discussão

Foi utilizado o software R Studio para simular dados do número de clientes no sistema no momento da partida, com base na equação (1) e foram feitas 5000 replicações Monte Carlo, com amostras de tamanho $n = 10, 100, 200$ para $\rho = 0.2, 0.5, 0.9$. As distribuições a priori escolhidas foram $\boldsymbol{beta}(\boldsymbol{0.6},\boldsymbol{1.7})$ e $\boldsymbol{uniforme\ truncada}(\boldsymbol{0.05},\boldsymbol{0.95})$, de acordo com a sugestão por Choudhury e Borthakur [2]. A figura 1 mostra as densidades a priori beta e uniforme truncada.

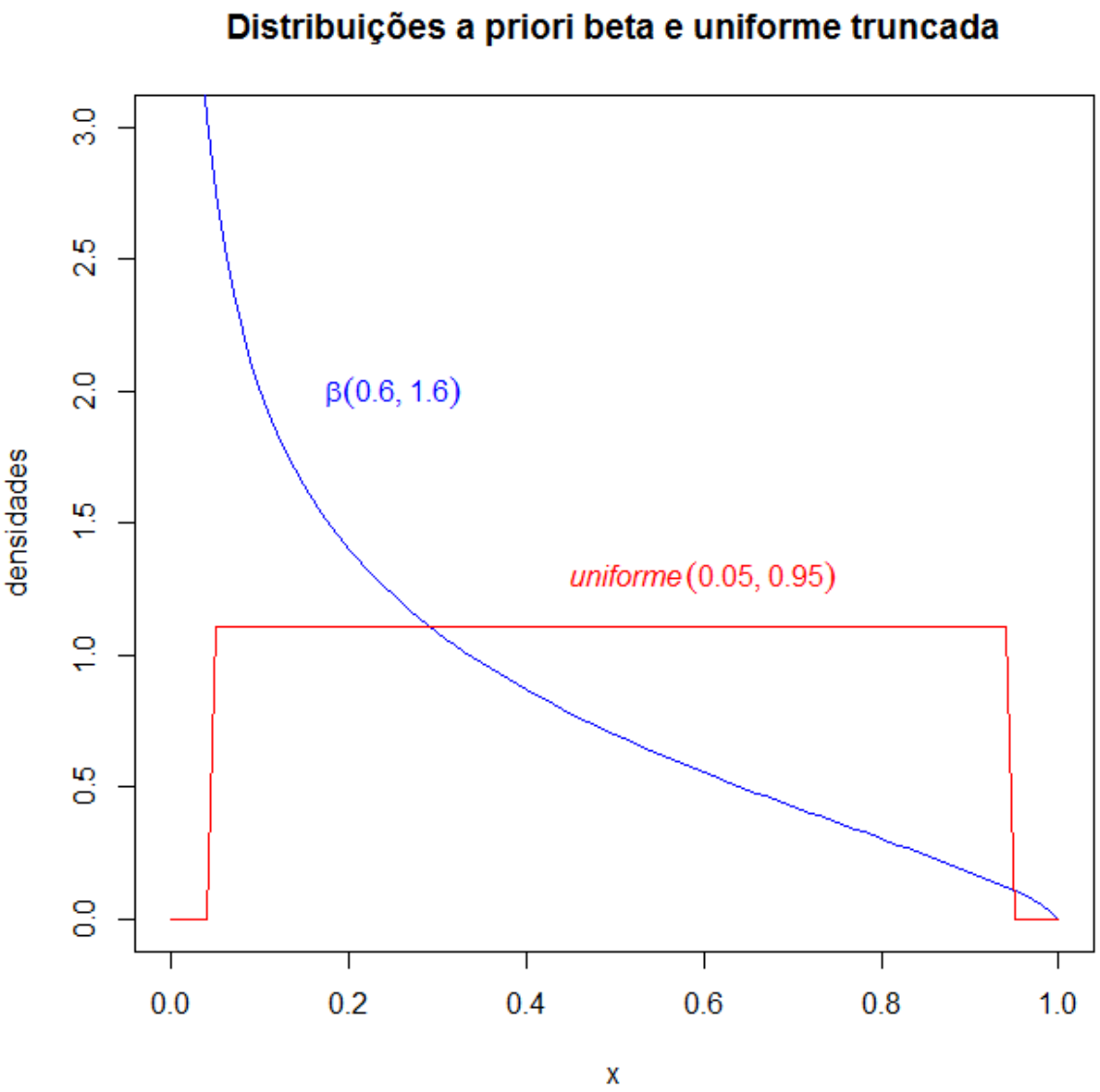


Figura 1: Distribuições a priori beta (0.6,1.7) e uniforme truncada (0.05,0.95) para a intensidade de tráfego.

A título de exemplificação, a figura 2 mostra os histogramas para a $\boldsymbol{beta(0.6,1,7)}$ **e** $\boldsymbol{uniforme\ truncada(0.05,0.95)}$ para $\boldsymbol{\rho = 0.2.}$

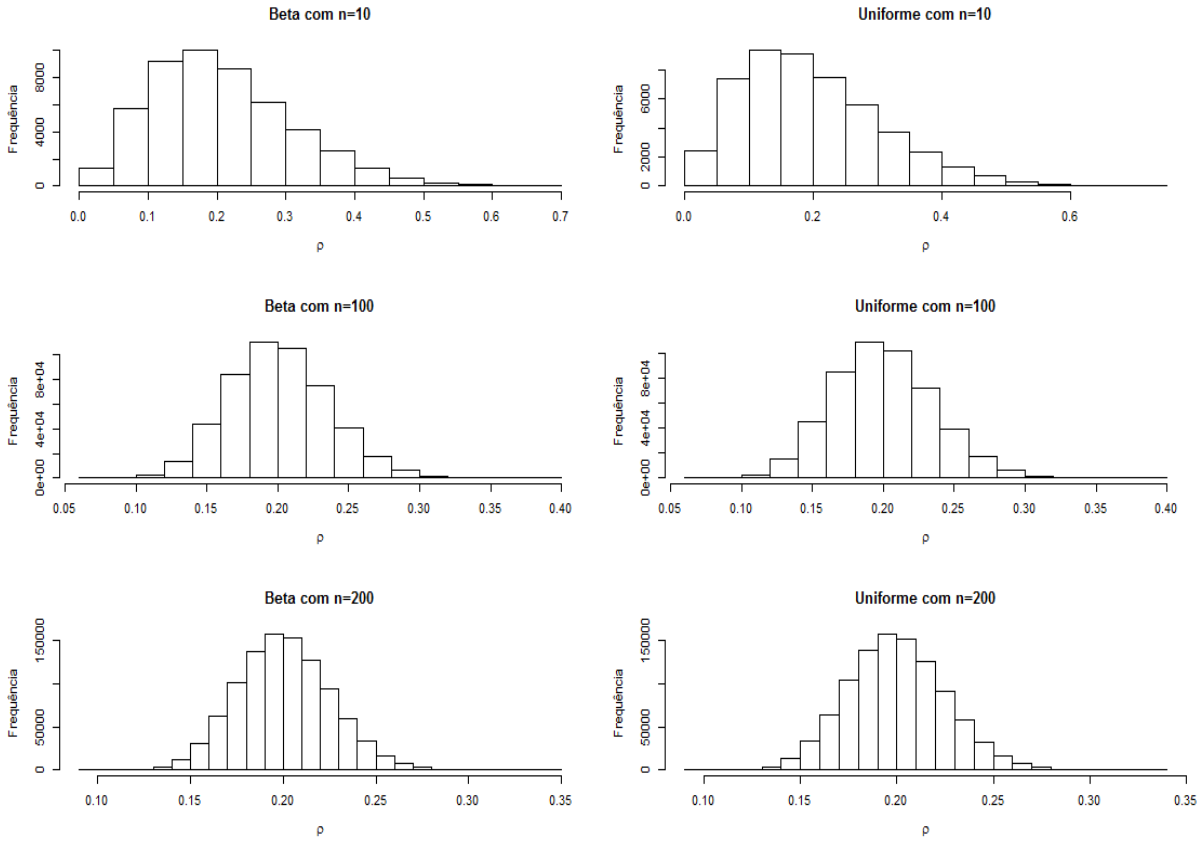


Figura 2: Histogramas para a $beta(0.6, 1.7)$ e $uniforme\ truncada(0.05, 0.95)$ considerando $\rho = 0.2$.

Posteriormente, foram obtidas as estimativas para L e L_q.

A tabela 2 mostra os valores teóricos e os resultados obtidos com amostras de tamanhos $n = 10, 100, 200$ para $\rho = 0.2, 0.5, 0.9$ para L e L_q.

Tabela 2: Valores exatos e estimativas para o número de clientes no sistema e tamanho da fila

				$n = 10$		$n = 100$		$n = 200$	
ρ	L	L_q	Distribuição	$\hat{L}$	$\hat{L}_q$	$\hat{L}$	L_q	$\hat{L}$	$\hat{L}_q$
0,2	0,25	0,05	Beta	0,2669	0,0562	0,2522	0,0508	0,2503	0,0501
			Uniforme	0,3523	0,1097	0,2605	0,0555	0,2545	0,0524
0,5	1,00	0,50	Beta	0,9047	0,4297	0,9877	0,4908	0,9937	0,4953
			Uniforme	1,0985	0,5089	1,0085	0,5089	1,0042	0,5044
0,9	9,00	8,10	Beta	7,7592	6,8733	8,8559	7,9573	8,9247	8,0255
			Uniforme	9,1182	8,2259	9,0104	8,1112	9,0026	8,1030

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

O algoritmo utilizado:

```
####################################
##rho = 0.2
##########################################
##CASO 1A: BETA(0.6,1.7)
##########################################
##amostras de tamanho 10
#########################
##posteriori beta(0.6+soma(dados);1.7+10)
##amostra1 recebe uma matriz de geometricas rho=0.2
##de tamanho 10
amostra10<-matrix(nrow=5000,ncol=10)
for(i in 1:5000){
amostra10[i,1:10]<- rgeom(10,0.8)
}
##matriz que recebe a soma das 5000 amostras de tamanho 10
somamostra10<-matrix(nrow=10,ncol=1)
for(i in 1:5000){
somamostra10[i]<-sum(amostra10[i,1:10])
}
mean(somamostra10)
##numero de clientes no sistema
lhatncA<-(0.6+mean(somamostra10))/(10+1.7)
lhatncA
##tamanho da fila
lqhatncA<-
((0.6+mean(somamostra10))^2)/((0.6+mean(somamostra10)+10+1.7)*
(10+1.7))
lqhatncA
########################################
########################################
##CASO 1B: BETA(0.6,1.7)
########################################
```

```
##amostras de tamanho 100
#########################
##posteriori beta(0.6+soma(dados);1.7+100)
##amostra1 recebe uma matriz de geometricas rho=0.2
##de tamanho 100
amostra100<-matrix(nrow=5000,ncol=100)
for(i in 1:5000){
  amostra100[i,1:100]<- rgeom(100,0.8)
}
##matriz que recebe a soma das 5000 amostras de tamanho 10
somamostra100<-matrix(nrow=100,ncol=1)
for(i in 1:5000){
  somamostra100[i]<-sum(amostra100[i,1:100])
}
mean(somamostra100)
##numero de clientes no sistema
lhatncB<-(0.6+mean(somamostra100))/(100+1.7)
lhatncB
##tamanho da fila
lqhatncB<-
((0.6+mean(somamostra100))^2)/((0.6+mean(somamostra100)+100+1.
7)*(100+1.7))
lqhatncB
########################################
##CASO 1C: BETA(0.6,1.7)
########################################
##amostras de tamanho 200
#########################
##posteriori beta(0.6+soma(dados);1.7+200)
##amostra1 recebe uma matriz de geometricas rho=0.2
##de tamanho 200
amostra200<-matrix(nrow=5000,ncol=200)
for(i in 1:5000){
  amostra200[i,1:200]<- rgeom(200,0.8)
```

```
}
##matriz que recebe a soma das 5000 amostras de tamanho 10
somamostra200<-matrix(nrow=200,ncol=1)
for(i in 1:5000){
  somamostra200[i]<-sum(amostra200[i,1:200])
}
mean(somamostra200)
##numero de clientes no sistema
lhatncC<-(0.6+mean(somamostra200))/(200+1.7)
lhatncC
##tamanho da fila
lqhatncC<-
((0.6+mean(somamostra200))^2)/((0.6+mean(somamostra200)+200+1.
7)*(200+1.7))
lqhatncC
####################
round(lhatncA,4)
round(lqhatncA,4)
round(lhatncB,4)
round(lqhatncB,4)
round(lhatncC,4)
round(lqhatncC,4)
##########################################
############################
##CASO 2: posteriori beta(soma;10)
############################
##amostras de tamanho 10
############################
#numero de clientes no sistema
lhattu<-(mean(somamostra10)+1)/(10)
lhattu
#tamanho da fila
```

```
lqhattu<-
((mean(somamostra10)+2)*(mean(somamostra10)+1))/(mean(somamost
ra10+12)*10)
lqhattu
##############################################
##amostras de tamanho 100
##############################################
#numero de cientes no sistema
lhattu100<-(mean(somamostra100)+1)/(100)
lhattu100
##tamanho da fila
lqhattu100<-
((mean(somamostra100)+2)*(mean(somamostra100)+1))/(mean(somamo
stra100+102)*100)
lqhattu100
##############################################
##amostras de tamanho 200
##############################################
#numero de clientes no sistema
lhattu200<-(mean(somamostra200)+1)/(200)
lhattu200
##tamanho da fila
lqhattu200<-
((mean(somamostra200)+2)*(mean(somamostra200)+1))/(mean(somamo
stra200+202)*200)
lqhattu200
round(lhattu,4)
round(lqhattu,4)
round(lhattu100,4)
round(lqhattu100,4)
round(lhattu200,4)
round(lqhattu200,4)
```

A partir daí, o algoritmo é recompilado substituindo rho pelos outros dois valores, quais sejam: 0.5 e 0.9.

Conclusão:

Foram utilizados métodos inferenciais sob uma abordagem bayesiana para estimar as medidas de desempenho de uma fila *M/M/1*. O modelo apresentado mostrou-se robusto para fazer predição, visto que o usuário pode atribuir seu conhecimento prévio sobre a operação de um sistema simples de filas. Mesmo sem o conhecimento dos valores da taxa de chegada e do tempo de serviço é possível inferir sobre o número de clientes no sistema no momento da partida, bem como do tamanho da fila. Através de simulação e utilizando duas formas de informação *a priori* foram obtidas estimativas bastante próximas dos valores teóricos, sobretudo quando se aumenta o tamanho da amostra. Probabilidades preditivas a posteriori do número de clientes no sistema e fator de Bayes estão em fase de estudo. Como recomendação para trabalhos futuros podem-se utilizar outras distribuições de probabilidade *a priori*, outros tipos de dados, bem como sistemas de filas mais gerais, tais como *M/G/1*.

Referências:

[1] C. Armero and M.J.Bayarri, Bayesian prediction in *MM/1* queues. Queueing Systems, vol. 14,401-417, (1994).

[2] A. Choudhury and A.C.Borthakur, Bayesian inference and prediction in the single server Markovian queue, Metrika, vol.67, 371-383,(2008).

[3] A.B. Clarke, Maximum likelihhod estimates in a simple queue. The Annals of Mathematical Statistics, vol.28, 1036-1040, (1957).

[4] F.R.B. Cruz and M. Almeida, Análise de Desempenho em Filas *M/M/1* Usando uma Abordagem Bayesiana. Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, vol. 2, N.2,(2015).

[5] D. Gamerman and H.F.Lopes, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman and Hall/CRC, London,UK, 2 ed.,(2006).

[6] Gross, D., Shortle and J.F. & Harris, C.M. *Fundamentals of Queueing Theory,* 4th ed, Wiley-Interscience, New York, USA. (2009)

[7] H. Jeffreys, The Theory of Probability, Oxford University Press, Oxford (1998).

[8] D.G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of embedded Marlov chains. *Annals Mathematical Statistics* 24: 338-354,(1953).

[9] C. D. Paulino, M. A. A. Turkman e B. Murteira, Estatística Bayesiana, Fundação Calouste Gulbenkian, Lisboa, (2003).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 15 - SELEÇÃO DE ATRIBUTOS ATRAVÉS DA REGRESSÃO LOGÍSTICA NO DRAFT DE QUARTERBACKS DA NFL

Brunno e Souza Rodrigues[5]
Carla Martins Floriano[6]
Valdecy Pereira [7]

Resumo

Este artigo tem por objetivo identificar quais são os principais atributos que maximizam a probabilidade de um *quarterback*, elegível ao *draft* da Liga Nacional de Futebol Americano (NFL – *National Football League*), ser ou não selecionado por uma equipe profissional. Serão, então, analisados os dados de todos os *quarterbacks* selecionados nos *drafts* do período de 1999 a 2016, considerando os resultados dos testes do *NFL Scouting Combine*, bem como as estatísticas dos jogos realizados por estes atletas durante as partidas do *College Football*. Através desses dados, será construído um modelo de regressão logística binária, de modo a identificar os atributos significativos. Assim, verificou-se que as variáveis independentes com maior importância para a seleção de *quarterbacks* foram: *Weight* (ligada ao aspecto físico do jogador); *Wonderlic* (resultado do teste de inteligência realizado no *Combine*); *40.Yard* (resultado do teste de esforço físico realizado no *Combine*); YdsG (dado estatístico de jogo); e TDG (dado estatístico de jogo).

Palavras-Chave: NFL, quarterback, logit.

Abstract

This article aims to identify which are the main attributes that maximize the likelihood of a quarterback, eligible for the NFL (National Football League) draft, be selected by a professional team or not. The data of all quarterbacks selected in the drafts from the period 1999 to 2016 will be analyzed, and it will be considered the results of the NFL Scouting Combine, as well as the statistics of the games performed by these athletes during the College Football. Through these data, a binary logistic regression model will be constructed in order to identify the most significant attributes. In this manner, it was verified that the independent variables with greater importance for the selection of quarterbacks were: Weight (linked to the physical aspect of the player); Wonderlic (result of intelligence test conducted in Combine); 40.Yard (result of physical test performed at Combine); YdsG (game statistics); And TDG (game statistics).

Keywords: NFL, quarterback, logit.

Introdução

É comum às ligas esportivas profissionais dos Estados Unidos a seleção anual dos jogadores universitários nos chamados *drafts*. Assim como nas demais ligas, esses atletas são convidados pela NFL (Liga Nacional de Futebol Americano) a demonstrarem suas habilidades e pontos fortes em uma série de testes físicos e

[5] UFF / e-mail: brunno.esr@gmail.com
[6] UFF / e-mail: carlafloria@gmail.com
[7] UFF / e-mail: valdecy.pereira@gmail.com

mentais que irão gerar dados estatísticos para o *NFL Scouting Combine*. É com base nestes dados e nas estatísticas dos jogos do *College Football* (campeonato universitário) que as franquias de futebol americano irão selecionar os jogadores que irão atuar pelos seus times. Essa intrincada combinação de dados e escolhas faz com que o time possa escolher um futuro astro do futebol americano ou um grande fracasso para a franquia.

Um time de futebol americano é formado por três unidades diferentes e cada unidade possui sempre 11 jogadores em campo: (i) ataque; (ii) defesa; e (iii) times especiais [NFL; 2016]. O *football* é uma modalidade esportiva que demanda grande especialização por posição e a principal posição de um time de futebol americano é a de *quarterback*. Este jogador funciona como o cérebro do time. Ele é quem recebe a bola no início das jogadas de ataque, é responsável pela organização do time, por distribuir os passes e, por vezes, realizar corridas. Dada a importância da posição, os *quarterbacks* estão sempre entre os primeiros a serem selecionados nos *drafts*.

Objetivo

Observado esse cenário, o objetivo central deste artigo é identificar quais são os principais atributos que maximizam a probabilidade de um *quarterback,* elegível ao *draft* da NFL, ser ou não selecionado.

Material e Métodos:

Para este artigo, serão analisados os dados de todos os 369 *quarterbacks* selecionados nos *drafts* de 1999 a 2016, com base nos resultados do *NFL Scouting Combine* [NFL SCOUTING COMBINE; 2016], bem como nas estatísticas dos jogos do *College Football* [NCAA; 2016]. Através desses dados, será construído um modelo de regressão logística binária, de modo a identificar os atributos significativos.

Vale ressaltar que o escopo da análise limitou-se ao período de 1999 a 2016, haja vista que os resultados do *NFL Scouting Combine* dos períodos anteriores a este não são públicos. Ademais, foram consideradas as estatísticas dos jogos do *College Football* nos anos de 1994 a 2015, período em que esses atletas atuaram como jogadores universitários.

Desta forma, conforme apresentado na Tabela 1, foram selecionadas para análise: 8 variáveis independentes do *NFL Scouting Combine*, 5 variáveis

independentes que remetem às estatísticas do *College Football* e uma variável dependente que aponta se o jogador foi selecionado ou não.

Tabela 1 – Sumário das variáveis analisadas

Variável	Fonte	Mo - Me [Min; Max]
Height	NFL Scouting Combine	74 - 75 [63; 80]
variável quantitativa pertinente à altura do atleta (dados em inches)		
Weight	NFL Scouting Combine	223 - 222 [171; 265]
variável quantitativa pertinente ao peso do atleta (dados em libras)		
Wonderlic	NFL Scouting Combine	24 - 26 [6; 48]
variável quantitativa pertinente à pontuação do atleta no teste de inteligência e raciocínio		
X40Yard	NFL Scouting Combine	4.84 - 4.83 [4.33; 5.37]
variável quantitativa pertinente à velocidade do jogador em uma corrida de 40 jardas em segundos		
VerLeap	NFL Scouting Combine	30.5 - 31.5 [21.5; 40]
variável quantitativa pertinente à altura máxima que o jogador consegue atingir em um pulo vertical em inches		
BroadJump	NFL Scouting Combine	110 - 110 [91; 128]
variável quantitativa pertinente à impulsão do jogador para saltar de um ponto a outro em inches		
Shuttle	NFL Scouting Combine	4.28 - 4.32 [3.87; 4.82]
variável quantitativa pertinente à velocidade de explosão do jogador e em distâncias laterais em segundos		
X3Cone	NFL Scouting Combine	7.17 - 7.17 [6.55; 7.97]
variável quantitativa pertinente à habilidade do jogador em mudar de direção quando sob alta velocidade em segundos		
Power Five	College Football	n/a
variável qualitativa que determina se o jogador disputou ou não seus jogos, em seu último ano de College, em uma das 5 mais importantes conferências (ACC, Big Ten, Big 12, Pac-12, SEC)		
G	College Football	45 - 46 [2; 61]
variável quantitativa pertinente ao total de jogos disputados pelo jogador durante o College		
PassCompAtt	College Football	64.1 - 59.2 [0; 75]
percentual de passes completados por passes tentados nos jogos do College		
YdsG	College Football	221.2 - 203.7 [0.5; 441.5]
total de jardas ganhas com passes, corridas e recepções por jogo, durante o College		
TDG	College Football	1.7 - 1.6 [0; 3.7]
touchdowns de passes, corridas e recepções por jogo durante o College		

Mo = moda; Me = média; Min = valor mínimo; Max = valor máximo.

Baseado nessas informações é possível aplicar um modelo de regressão logística binária, de modo a permitir a análise da variável de interesse (variável dependente) em função das demais variáveis (variáveis independentes). Com este modelo é possível identificar os fatores (*inputs*) que mais influenciam e determinam um resultado (*output*), mediante a aplicação de um modelo explicativo que os relaciona [MAKRIDAKIS et al.; 1983], conforme apresentado na Equação 1.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

$$Drafted \in \{0;1\}$$

$$Z_i = ln\left(\frac{p_i}{1-p_i}\right)$$

$$= \beta_0 + \beta_1 PowerFive + \beta_2 Height + \beta_3 Wonderlic + \beta_4 40Yard + \beta_5 VertLeap + \beta_6 BroadJump + \beta_7 Shuttle + \beta_8 3Cone + \beta_9 G + \beta_{10} PassCompAtt + \beta_{11} YdsG + \beta_{12} TDG \quad (1)$$

onde:

i = cada jogador da amostra;

Z_i = logit, regressão linear binária;

p_i = probabilidade do jogador ser draftado;

$1 - p_i$ = pronabilidade do jogador não ser draftado;

β_0 = constante;

β_k = coeficientes de regressão;

Haja vista que os parâmetros do modelo são não-lineares, fez-se necessário a aplicação do método da máxima verossimilhança. Com isso, as estimativas foram obtidas pelos valores dos parâmetros que maximizam a probabilidade do quarterback ser selecionado ($Y_i = 1$)). Todavia, para este artigo, será considerada a maximização da estimação do logarítimo da função de verossimilhança, conforme apresentado na Equação 2.

$$LL = \sum_{i=1}^{n} \left\{ \left[(Y_i) ln\left(\frac{e^{(Z_i)}}{1+e^{(Z_i)}}\right) \right] + \left[(1-Y_i) ln\left(\frac{1}{1+e^{(Z_i)}}\right) \right] \right\} = máx \quad (2)$$

A vantagem da aplicação do modelo de regressão logística binário é que o mesmo produz resultados que permitem interpretação dos efeitos preditórios das variáveis independentes sobre a variável dependente, bem como não é restringido por rígidos pressupostos iniciais [Pereira; 2016]. No entanto, faz-se necessário observar que, este modelo deve atender aos seguintes parâmetros: (i) a variável dependente deve ser dicotômica; (ii) as variáveis independentes devem ser dicotômicas ou métricas; (iii) ausência de colinearidade ou multicolinearidade; e (iv) ausência de *outliers*.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Resultados e Discussão:

Posto o modelo de regressão logística binária, através do *software* estatístico R 3.3.2 e Rstudio 0.99.903, foram estimados os coeficientes de regressão das variáveis independentes, conforme apresentado na Tabela 2.

Tabela 2 – Regressão logística binária dos quarterbacks elegíveis no draft da NFL

	Coeficiente Estimado	Erro Padrão	Z valor	p-Valor
(Intercept)	2.2213	2.0267	1.0960	0.2731
Pow er.Five	0.3267	0.2778	1.1760	0.2395
Height	0.0564	0.0616	0.9150	0.3604
Weight	0.0480	0.0136	3.5300	0.0004
Wonderlic	0.0779	0.0106	7.3340	0.0000
X40Yard	-4.2899	0.8139	-5.2710	0.0000
Vert.Leap	-0.0133	0.0202	-0.6610	0.5086
Broad.Jump	0.0015	0.0057	0.2580	0.7966
Shuttle	0.1890	0.2162	0.8740	0.3819
X3Cone	-0.0967	0.1206	-0.8020	0.4227
G	0.0223	0.0146	1.5340	0.1250
PassCompAtt	0.0139	0.0259	0.5380	0.5907
YdsG	0.0119	0.0049	2.4490	0.0143
TDG	-0.8325	0.4771	-1.7450	0.0810

Através desse resultado, é possível verificar que as variáveis estatisticamente significativas ao modelo e que, portanto, possuem p-valor inferior a 0.1 (dentro do intervalo de confiança de 90%) são: *Weight* (ligada ao aspecto físico do jogador); *Wonderlic* (resultado do teste de inteligência realizado no *Combine*); 40Yard (resultado do teste de esforço físico realizado no *Combine*); YdsG (ligada às estatística dos jogos disputados pelo atleta no *College Football*); e TDG (ligada às estatística dos jogos disputados pelo atleta no *College Football*).

Com isso, os coeficientes foram novamente calculados considerando apenas as variáveis significantes, conforme apresentado na Tabela 3.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Tabela 3 – Regressão logística binária dos *quarterbacks* elegíveis no *draft* da NFL, desconsiderando as variáveis não significativas

	Coeficiente Estimado	Erro Padrão	Z valor	p-Valor
(Intercept)	3.9858	1.9186	2.0780	0.2731
Weight	0.0527	0.0114	4.6100	0.0000
Wonderlic	0.0742	0.0101	7.3460	0.0000
X40Yard	-3.7236	0.6083	-6.1210	0.0000
YdsG	0.0135	0.0046	2.9380	0.0033
TDG	-0.7868	0.4598	-1.7110	0.0810

Os coeficientes estimados nas Tabelas 2 e 3 podem ser interpretados como o efeito parcial da variável independente no resultado da variável dependente do modelo, mantendo constantes as demais variáveis. Os coeficientes negativos (40Yard e TDG) indicam que valores mais altos atribuídos a essas variáveis predizem uma maior chance de escolha no *draft*, ou seja, através do modelo é possível inferir que jogadores que obtiveram altos resultados no teste de corrida (40Yard), ou que realizaram mais *touchdowns* por jogo (TDG), possuem maiores chances de serem escolhidos no *draft*.

Mediante o exponenciamento dos coeficientes obtidos, na Tabela 4, é apresentada a probabilidade dos jogadores serem selecionados ao alterar a entrada de uma variável independente, mantendo as demais constantes.

Tabela 4 – Variação das chances de seleção dos *quarterbacks* no *draft*

Variável	Coeficiente estimado	Coeficiente exponeciado	Interpretação
Weight	0.0527	$e^{0.0527} = 1.0541$	A chance do jogador ser escolhido no *draft* aumenta em 5.41% se o peso for aumentado em 1 libra.
Wonderlic	0.0742	$e^{0.0742} = 1.0769$	A chance do jogador ser escolhido no *draft* aumenta em 7.69% se o resultado do teste de inteligência for aumentado em 1 ponto.
40.Yard	- 3.7236	$e^{-3.7236} = 0.0241$	A chance do jogador ser escolhido no *draft* decresce em 97.69% se o resultado do teste corrida de 40 jardas for aumentado em 1 segundo.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

YdsG	0.0135	$e^{0.0135} = 1.0136$	A chance do jogador ser escolhido no *draft* aumenta em 1.36% se o mesmo realizar 1 jarda a mais por jogo disputado.
TDG	- 0.7868	$e^{-0.7868} = 0.4553$	A chance do jogador ser escolhido no *draft* decresce em 54.47% se o mesmo realizar 1 touchdown a menos por jogo disputado.

Conclusão:

Neste artigo, buscou-se identificar quais são os principais atributos que determinam se um *quarterback* vai ser ou não selecionado durante o *draft* da NFL. Através da aplicação de um modelo de regressão logística binária, verificou-se que as variáveis independentes com maior significância foram: *Weight* (ligada ao aspecto físico do jogador); *Wonderlic* (resultado do teste de inteligência realizado no *Combine*); *40Yard* (resultado do teste de esforço físico realizado no *Combine*); YdsG (dado estatístico de jogo); e TDG (dado estatístico de jogo).

Isto posto, encontrou-se um modelo, baseado nas cinco supracitadas variáveis, cuja acurácia foi de 77.5%. Desta forma, *quarterbacks* postulantes a contratos da NFL que apresentam alto desempenham nesse conjunto de variáveis possuem maiores chances de serem selecionados nos *drafts* regulares. A título de ilustração, caso um jogador apresente dados iguais à média de todos os *quarterbacks* analisados (222lbs *Weight*; 26pts *Wonderlic*; 4.83s *40Yard*; 203.7 Yds/G; 1.6 *Touchdowns* por jogo), a probabilidade do mesmo ser selecionado em um *draft* é de 75.6%.

Referências:

BERRI, David J.; BROOK, Stacey L.; FENN, Aaju J. From college to the pros: Predicting the NBA amateur player draft. **Journal of Productivity Analysis**, v. 35, n. 1, p. 25–35, 2011.

MAKRIDAKIS, Spyros; WHEELWRIGHT, Steven C.; HYNDMAN, Rob J. **Forecasting: Methods and Applications**. John Wiley & Sons, Inc., 3rd ed.,1983.

MULHOLLAND, Jason; JENSEN, Shane T. Predicting the draft and career success of

tight ends in the National Football League. **Journal of Quantitative Analysis in Sports**, v. 10, n. 4, p. 381–396, 2014.

NCAA. **National Collegiate Athletic Association**. Web page. http://www.ncaa.com. Acessado: 2016-12-04.

NFL. **National Football League**. Web page. http://www.nfl.com .Acessado: 2016-12-02.

NFL SCOUTING COMBINE. **National Football League Scouting Combine**. Web page. http://www.nflcombineresults.com. Acessado: 2016-12-06.

PEREIRA, Valdecy. (2016). Project: Multivariate Data Analysis. File: R-MVDA-08-LR-B.pdf. GitHub repository: <https://github.com/Valdecy/Multivariate_Data_Analysis>

WOLFSON, Julian; ADDONA, Vittorio; SCHMICKER, Robert H. The Quarterback Prediction Problem: Forecasting the Performance of College Quarterbacks Selected in the NFL Draft. **Journal of Quantitative Analysis in Sports**, v. 7, n. 3, p. 1–19, 2011.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 17 - ANÁLISE DA SUSTENTABILIDADE DE UNIDADES DE GASEIFICAÇÃO POR MEIO DE TÉCNICA PARA AVALIAR O DESEMPENHO DE ALTERNATIVAS ATRAVÉS DE SIMILARIDADE COM A SOLUÇÃO IDEAL

Gloria Maria Alves Ney[8]
Luiz Octávio Gavião [9]
Gilson Brito Alves Lima[10]
Márcio Zamboti Fortes[11]

Resumo

A crescente contribuição de gases no impacto da camada de ozônio e no aquecimento global mostra a relevância na adoção de um sistema de geração de energia para que haja a redução de emissões. O processo de gaseificação se resume em realizar uma oxidação parcial com o objetivo de se obter gás de síntese ao término do processo. Com o objetivo de avaliar a eficiência desses processos, o artigo analisa unidades de produção que aplicaram diferentes tecnologias de processo de gaseificação. A modelagem explorou o referencial da sustentabilidade proposta pelo *Triple Bottom Line* (TBL). Como resultado, foi obtida uma ordenação das unidades de produção, sob a perspectiva da eficiência econômica, social e ambiental. Para a avaliação da sustentabilidade das unidades de gaseificação foi utilizado a técnica TOPSIS que é utilizada para apoio à decisão multicritério, desenvolvida por Hwang & Yoon (1981).

Palavras-Chave: Gaseificação, Comissionamento, Sustentabilidade, TOPSIS

Abstract

Due to the increasing contribution of gases in the impact of the ozone layer and in global warming shows the relevance in the adoption of a system of generation of energy for the reduction of emissions. The gasification process is summarized in performing a partial oxidation in order to obtain synthesis gas at the end of the process. In order to evaluate the efficiency of these processes, the article analyzes production units that applied different gasification process technologies. The modeling explored the sustainability referential proposed by the Triple Bottom Line (TBL). As a result, an ordering of production units was obtained from the perspective of economic, social and environmental efficiency. For the evaluation of the sustainability of the gasification units, the TOPSIS technique was used to support multicriteria decision, developed by Hwang & Yoon (1981).

Keywords: Gasification, Commissioning, Sustainability, TOPSIS

Introdução

A crescente contribuição de gases no impacto da camada de ozônio e no aquecimento global mostra a relevância na adoção de um sistema de geração de

[8] Universidade Federal Fluminense (UFF), glorianey1612@gmail.com
[9] Universidade Federal Fluminense (UFF), luiz.gaviao67@gmail.com
[10] Universidade Federal Fluminense (UFF), glima@id.uff.br
[11] Universidade Federal Fluminense (UFF), mzf@id.uff.br

energia que possa ser realizado em ciclo fechado ou com alternativa de diferentes combustíveis para que haja a redução de emissões. A gaseificação, utilizando líquidos e gases como combustíveis, foi desenvolvida no final da década de 1940 pela TEXACO e no início da década de 1950 pela SHELL (HIGMAN E BURGT, 2011). Este processo se resume em uma oxidação parcial, que converte materiais ricos em carbono em gás de síntese (THYSSENKRUPP UHDE, 2012), contribuindo para uma eficiência térmica maior e consequente redução dos gases de efeito estufa. Neste aspecto, o artigo analisa unidades de produção que aplicaram diferentes tecnologias de processo de gaseificação, com o objetivo de avaliar a eficiência desses processos. A modelagem explorou o referencial da sustentabilidade proposta pelo *Triple Bottom Line* (TBL), com cálculos suportados pelo pacote "topsis" do software "R" (R-Core Team, 2016). Como resultado, foi obtida uma ordenação das unidades de produção, sob a perspectiva da eficiência econômica, social e ambiental.

Objetivo

O artigo apresenta o estudo de unidades de produção que aplicaram diferentes tecnologias de processo de gaseificação, objetivando a avaliação da eficiência dos processos, a partir da aplicação da Técnica de Similaridade com Solução Ideal (TOPSIS), no contexto da abordagem da sustentabilidade proposta pelo TBL.

Material e Métodos:

Para avaliar a sustentabilidade das unidades de gaseificação foi utilizado o TOPSIS que é uma técnica de apoio à decisão multicritério, desenvolvida por Hwang & Yoon (1981). De maneira geral, o método estabelece uma ordenação das alternativas, com base no princípio de que a melhor alternativa é aquela que, simultaneamente, possua a menor distância para a referência positiva (PIS) e a maior distância para a referência negativa (NIS). Essas referências podem ser consideradas alternativas fictícias, elaboradas a partir das melhores e piores avaliações x_{ij} da i-ésima alternativa do j-ésimo critério. A lógica da modelagem TOPSIS, de forma intuitiva, pode ser aplicada em diferentes áreas de conhecimento.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, \text{ i=1,...,m; j=1,...,n} \quad (1)$$

$$v_{ij} = w_j r_{ij}, \text{ i=1,...,m; j=1,...,n} \quad (2)$$

$$PIS = \{v_1^*, \dots, v_j^*, \dots, v_n^*\} = \{(max_j\ v_{ij} | j = 1, \dots, n) \mid i = 1, \dots, m\} \quad (3)$$

$$NIS = \{v_1^-, \dots, v_j^-, \dots, v_n^-\} = \{(min_j\ v_{ij} | j = 1, \dots, n) \mid i = 1, \dots, m\} \quad (4)$$

$$D_i^* = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^*)^2}, \text{ i=1,...,m} \quad (5)$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2}, \text{ i=1,...,m} \quad (6)$$

$$C_i^* = S_i^- / (S_i^* + S_i^-), \text{ i=1,...,m} \quad (7)$$

A modelagem da técnica TOPSIS é desenvolvida em cinco etapas: na primeira, deve-se normalizar a base de dados, tendo em vista a diferença de escala entre as avaliações nos critérios da matriz de decisão. Os valores x_{ij} são transformados em r_{ij}. Usualmente é empregada a equação (1), segundo Lu, Zhang, Ruan, & Wu (2007). Na segunda etapa, os valores normalizados são multiplicados pelos pesos, previamente estabelecidos, w_j, conforme a Equação (2). Os valores r_{ij} são ponderados e transformados em v_{ij}. Na terceira etapa são identificados os vetores PIS e NIS, conforme as Equações (3) e (4). Na quarta etapa são calculadas as distâncias D* e D^- de cada alternativa para a PIS e a NIS, conforme as Equações (5) e (6), respectivamente. Para finalizar, o coeficiente de similaridade C é obtido a partir da Equação (7). A ordem de preferência é estabelecida a partir do maior C, que caracteriza a melhor alternativa.

Duas informações adicionais são necessárias para a modelagem: o impacto e os pesos dos critérios. O impacto se refere à interferência positiva ou negativa de um critério na tomada de decisão, com o aumento da avaliação. Os critérios com impacto negativo requerem a inversão dos operadores “max-min” das Equações (3) e (4), respectivamente, para determinar os vetores PIS e NIS. Os pesos dos critérios devem ser previamente obtidos, de forma a serem utilizados na segunda etapa do TOPSIS.

Resultados e Discussão:

Para o estudo foram consideradas as unidades de gaseificação da China, disponibilizadas pelo *US Department of Energy* (2016). As emissões decorrentes de

cada usina foram obtidas em Higman & Burgt (2011). A Tabela 1 apresenta a base de dados, composta por 29 usinas, com seus respectivos desempenhos quantitativos nas dimensões econômica, ambiental e social.

Tabela 1 – Dados de entrada

Unidades	Tecnolog	**Econômico**	**Ambiental**					**Social**
		Produção	Composição do gás emitido (t/dia)					Nr. Funcionários
		(US$)	CO2	CO	H2	CH4	N + Ar	
			Número de risco segundo CETESB					
			20	263	23	23	20	
Pesos		0,333	0,019	0,251	0,022	0,022	0,019	0,333
USINA 1	GE	531225,00	23,40	315,00	549,90	2,70	9,00	7500
USINA 2	GE	436785,00	19,24	259,00	452,14	2,22	7,40	1000
USINA 3	GE	129369,86	5,70	76,71	133,92	0,66	2,19	3900
USINA 4	GE	485136,99	21,37	287,67	502,19	2,47	8,22	4503
USINA 5	GE	590250,00	26,00	350,00	611,00	3,00	10,00	300
USINA 6	GE	227246,25	10,01	134,75	235,24	1,16	3,85	1300
USINA 7	GE	436785,00	19,24	259,00	452,14	2,22	7,40	21771
USINA 8	GE	531225,00	23,40	315,00	549,90	2,70	9,00	39485
USINA 9	GE	753424,66	35,62	479,45	836,99	4,11	13,70	94600
USINA 10	GE	904109,59	42,74	575,34	1004,38	4,93	16,44	94600
USINA 11	GE	550000,00	26,00	350,00	611,00	3,00	10,00	95498
USINA 12	GE	904109,80	42,74	575,34	1004,38	4,93	16,44	10000
USINA 13	LURGI	313643,84	7,93	121,95	9514,52	0,38	3,84	6702
USINA 14	LURGI	404280,82	24,79	381,10	29732,88	1,20	11,99	11000
USINA 15	LURGI	180821,92	11,90	182,93	14271,78	0,58	5,75	6702
USINA 16	LURGI	813698,63	53,56	823,17	64223,01	2,59	25,89	10000
USINA 17	SHELL	153465,00	4,45	90,71	159,64	2,60	2,60	1864
USINA 18	SHELL	153465,00	4,45	90,71	159,64	2,60	2,60	2498
USINA 19	SHELL	323424,66	9,37	191,18	336,44	5,48	5,48	4867
USINA 20	SHELL	485136,99	14,05	286,77	504,66	8,22	8,22	3426
USINA 21	SHELL	808561,64	23,42	477,95	841,10	13,70	13,70	5400
USINA 22	SHELL	485136,99	14,05	286,77	504,66	8,22	8,22	11000
USINA 23	SHELL	447903,00	5,35	109,21	192,18	3,13	3,13	1300
USINA 24	SHELL	301369,86	9,37	191,18	336,44	5,48	5,48	864
USINA 25	SHELL	753424,66	23,42	477,95	841,10	13,70	13,70	200
USINA 26	SHELL	452054,79	14,05	286,77	504,66	8,22	8,22	988
USINA 27	SHELL	753424,66	23,42	477,95	841,10	13,70	13,70	11000
USINA 28	SHELL	55000,00	1,71	34,89	61,40	1,00	1,00	46555
USINA 29	SHELL	2994410,96	21,55	439,71	773,81	12,60	12,60	24279

Fontes: *US Department of Energy* (2016) e Higman & Burgt (2011).

Para a modelagem do processo de gaseificação foi aplicada a técnica TOPSIS a partir de uma adequação dos parâmetros à sustentabilidade econômica, social e ambiental. Neste sentido, como vetor econômico foi considerado a "produção das unidades". Como vetor social, utilizou-se o "número de trabalhadores da unidade de produção". Para o vetor ambiental, utilizou-se as "emissões geradas pelas unidades de produção", tendo, respectivamente, a direção dos três vetores os sentidos: positivo (i.e., quanto maior, melhor), positivo, negativo (i.e., quanto menor melhor).

A modelagem do TOPSIS requer a atribuição de pesos aos critérios. Por coerência ao referencial teórico, no que se refere ao equilíbrio das dimensões do TBL (econômico, social e ambiental), os pesos foram distribuídos equitativamente para as dimensões, equivalendo ao valor de 0,333 para cada item. Tendo em vista a composição da dimensão ambiental com cinco subcritérios, foi necessário redistribuir o peso entre os diferentes tipos de emissões. Nesse caso, as emissões foram ponderadas segundo o grau toxicidade, seguindo os parâmetros indicados pela Companhia Ambiental do Estado de São Paulo (CETESB, 2016). Na Tabela 1, os valores da linha "Peso" indicam os valores finais da ponderação de cada critério e subcritério.

Por requisito de confidencialidade, os nomes das usinas foram preservados. O código em linguagem "R" está descrito no Apêndice. O resultado gerado pelo TOPSIS está apresentado na Tabela 2.

Tabela 2 – Resultado TOPSIS

Unidade	Tecnologia	Coeficiente TOPSIS	Rank
USINA 1	GE	0,237839	20
USINA 2	GE	0,233892	23
USINA 3	GE	0,260332	10
USINA 4	GE	0,235262	22
USINA 5	GE	0,228525	25
USINA 6	GE	0,249128	14
USINA 7	GE	0,269715	8
USINA 8	GE	0,321571	6
USINA 9	GE	0,498636	3
USINA 10	GE	0,509879	2
USINA 11	GE	0,486829	4
USINA 12	GE	0,250060	13
USINA 13	LURGI	0,263377	9
USINA 14	LURGI	0,206420	28

Unidade	Tecnologia	Coeficiente TOPSIS	Rank
USINA 15	LURGI	0,237395	21
USINA 16	LURGI	0,199078	29
USINA 17	SHELL	0,256180	12
USINA 18	SHELL	0,256727	11
USINA 19	SHELL	0,244749	15
USINA 20	SHELL	0,233867	24
USINA 21	SHELL	0,240562	17
USINA 22	SHELL	0,244697	16
USINA 23	SHELL	0,273286	7
USINA 24	SHELL	0,239085	19
USINA 25	SHELL	0,224595	27
USINA 26	SHELL	0,227743	26
USINA 27	SHELL	0,239859	18
USINA 28	SHELL	0,348554	5
USINA 29	SHELL	0,636019	1

Fonte: Autores

Realizando uma comparação entre as posições extremas das unidades analisadas, tem-se a Usina 29 como primeira e a Usina16 como a última. Em relação à dimensão econômica, é possível verificar que a Usina 29 apresenta a maior produção, enquanto a Usina 16 é a quarta mais lucrativa, com uma diferença de US$ 2.180.712,33 nesse critério. Em relação ao número de funcionários, a Usina 29 ocupa a sexta posição, enquanto a Usina 29 ocupa a décima primeira. Em relação à dimensão ambiental, verifica-se que a Usina 16 apresenta elevados índices de emissão, principalmente em volumes de monóxido de carbono (CO) e hidrogênio (H_2). Dessa forma, a Usina 16 obteve a pior classificação, indicando um desequilíbrio em relação às três dimensões do TBL. Por outro lado, a Usina 29 apresentou o melhor equilíbrio em relação às dimensões econômica, ambiental e social, atendendo em melhores condições o requisito da sustentabilidade, conforme estabelece o referencial teórico do TBL.

Conclusão:

Com o objetivo de realizar uma análise mais consistente sobre a sustentabilidade de uma unidade de gaseificação considerando a metodologia TBL, é preciso avaliar os três diferentes critérios. Através do TOPSIS foi possível fazer uma análise dos resultados por ser uma ferramenta que avalia diferentes critérios

simultaneamente e a partir disto ranquear as unidades para que se conhecesse a melhor opção dentre as usinas estudadas.

O modelo utilizado anteriormente não permitia uma análise mais consiste, pois considerava apenas um único item (econômico) apresentando assim uma análise superficial do problema. O TOPSIS se propõe em encontrar a solução mais próxima a ideal avaliando diferentes critérios. Este modelo permitiu o apoio a decisão sobre os multicritérios econômico, social e ambiental, reduzindo a subjetividade do processo monocritério produção.

Referências:

COMPANHIA AMBIENTAL DO ESTADO DE SÃO PAULO (CETESB). **Consulta por nome ou sinônimo do produto**. Disponível em: <http://sistemasinter.cetesb.sp.gov.br/produtos/produto_consulta_nome.asp>. Acesso em: 21 dez. 2016.

HIGMAN, C.; BURGT, M. v. d. **Gasification**. Gulf Professional Publishing, p. 456, 2011.

HWANG, C. L.; YOON, K. **Multiple attribute decision making: methods and applications**. Springer-Verlag. 1981.

LU, J., ZHANG, G., RUAN, D., WU, F. **Multi-Objective Group Decision Making: Methods, Software and Applications with Fuzzy Set Techniques (With CD-ROM)**. v. 6. World Scientific. 2007.

R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing**. Disponível em: < https://www.R-project.org/>. Vienna, Austria, 2016.

THYSSENKRUPP UHDE. **Gasification Technologies**. 2012. Disponível em: <http://www.thyssenkrupp-industrial-solutions.com/fileadmin/documents/brochures/gasification_technologies.pdf>. Acesso em: 12 jul. 2016.

TSAUR, R. C. (2011). **Decision Risk Analysis for Interval TOPSIS Method**. Elsevier. v. 218, p. 4295 – 4304, 2011. DOI: 10.1016/j.amc.2011.10.001

U. S. DEPARTMENT OF ENERGY. **Gasification Plant Databases**. 2016. Disponível em: <https://www.netl.doe.gov/research/coal/energy-systems/gasification/gasification-plant-databases>. Acesso em: 18 dez. 2016.

ZELENY, M. **Multi criteria decision making**. New York: McGraw-Hills. 1982.

Apêndice

```
# TOPSIS USINAS
require(topsis)
require(xlsx)
require(readxl)
path = 'Diretório de trabalho do usuário'
setwd(path)
dados = read_excel("Nome do arquivo.xlsx", col_names = FALSE, sheet = 1)
dados = as.matrix(dados)
d = dados
w = c(0.333, 0.019083095, 0.250942693, 0.021945559, 0.021945559, 0.019083095,
0.333)  # pesos dos critérios
i = c("+","-","-","-","-","-","+")  # impactos positivos ou negativos dos critérios
Topsis = topsis(d, w, i)
# Resultados exportados para planilha do MS Excel
write.xlsx(Topsis, “Diretório de trabalho do usuário\\Topsis.xlsx”)
```

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 18 - PROGRAMAÇÃO LINEAR INTEIRA NA ORGANIZAÇÃO DE FÓRUNS EMPRESARIAIS: UM EXEMPLO DO USO COMBINADO DO R COM O EXCEL

José Francisco Moreira Pessanha[12]
Narcisa Maria Gonçalves dos Santos[13]

Resumo

Associações empresariais costumam promover fóruns entre empresários e empreendedores de diferentes segmentos. Uma estratégia para estabelecer contatos entre os participantes consiste em organizá-los em pequenos grupos e alocá-los em mesas, de tal forma que os participantes em um mesmo grupo possam divulgar suas empresas, apresentar projetos e trocar ideias e experiências. Após um determinado intervalo de tempo, uma nova sessão é iniciada e novos grupos de participantes são formados e alocados em outras mesas para que novos contatos sejam realizados. A repetição desta dinâmica até o encerramento do fórum permite que cada participante realize um grande número de contatos. Contudo, visando ampliar as redes de relacionamentos é fundamental evitar os reencontros de participantes ao longo das sessões do evento. Neste trabalho apresenta-se um modelo de Programação Linear Inteira capaz de organizar um roteiro com a sequência de mesas a serem visitadas por cada participante com o objetivo de evitar ou minimizar os reencontros. O modelo foi implementado em ambiente R e conta com uma interface em MS Excel para entrada de dados e apresentação de resultados. Os resultados gerados pelo modelo são ilustrados por meio de um experimento computacional.

Palavras-Chave: Programação Linear Inteira, fóruns empresariais, R, planilha

Abstract

Business associations often promote forums between entrepreneurs from different segments. One strategy for establishing contacts among the participants is to organize them into small groups and allocate them to tables so that participants in the same group can advertise their companies, present projects and exchange ideas and experiences. After a certain time interval, a new session is started and new groups of participants are formed and allocated at other tables in order to make new contacts. The repetition of this dynamic until the closure of the forum allows each participant to make a large number of contacts. However, in order to broaden the networks of relationships it is essential to avoid re-encounters in the new groups throughout the sessions of the event. In this work we present an Integer Linear Programming model capable of organizing the sequence of tables to be visited by each participant with the objective of avoiding or minimizing re-encounters. The model was implemented in R environment and it has an interface in MS Excel for input/output data. The results generated by the model are illustrated by a computational experiment.

Keywords: Integer Linear Programming, business forums, R, spreadsheet

[12]Universidade do Estado do Rio de Janeiro – Uerj, professorjfmp@hotmail.com
[13]Universidade do Estado do Rio de Janeiro – Uerj, narcisa@imagelink.com.br

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

Uma rede de contatos é um poderoso instrumento para alavancar o crescimento de qualquer negócio. A conectividade favorece o surgimento de ideias (Johnson, 2011) e iniciativas com forte potencial para promover a inovação e o crescimento econômico, beneficiando toda a sociedade. Por esta razão, não raro, as associações empresariais costumam promover fóruns que estimulem a conversação e a colaboração entre empresários e empreendedores de diferentes segmentos.

Uma estratégia para estabelecer contatos entre os numerosos participantes de um fórum de negócios consiste em organizá-los em pequenos grupos e alocá-los em mesas, de tal forma que os participantes em um mesmo grupo (mesa) possam divulgar suas empresas, apresentar projetos e trocar ideias e experiências com os demais participantes do grupo. Após um determinado intervalo de tempo, suficiente para que os membros dos grupos se conheçam, uma nova sessão é iniciada e novos grupos de participantes são formados e alocados em outras mesas para que novos contatos sejam realizados. A repetição desta dinâmica até o encerramento do fórum permite que cada participante realize um grande número de contatos aleatórios. Contudo para que esta estratégia seja exitosa em ampliar as redes de relacionamentos é fundamental evitar os reencontros de participantes na formação dos novos grupos ao longo das sessões do evento. Para evitar ou minimizar os reencontros é necessário organizar previamente um roteiro com a sequência de mesas a serem visitadas por cada participante.

A determinação de roteiros sem reencontros dos participantes guarda alguma semelhança com o problema do caixeiro viajante, um problema clássico da Programação Linear Inteira - PLI (Ragsdale, 2004) em que o objetivo consiste em estabelecer uma sequência de localidades a serem visitadas, de tal forma que cada localidade seja visitada apenas uma vez. Trata-se de um problema de difícil solução e que requer grande esforço computacional para ser resolvido. A PLI tem vasta aplicação prática e fornece desde soluções aos problemas de roteirização e localização de instalações (Ragsdale, 2004) até soluções para o quebra cabeça Sudoku (Bartlett et al, 2008) e a organização dos convidados à uma festa de casamento com base nas afinidades entre eles (Bellows & Peterson, 2012).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Objetivo

No presente texto descreve-se um modelo de Programação Linear Inteira concebido para estabelecer os roteiros das mesas a serem visitadas pelos participantes de um fórum de negócios, de tal forma que o número de reencontros ao longo do evento seja minimizado. O modelo foi implementado em linguagem R (R Core Team, 2014) e conta com uma interface para entrada de dados e visualização de resultados programada em planilha MS Excel (Ragsdale, 2004).

Material e Métodos:

Considere um evento com n participantes, organizados em m mesas e p pessoas por mesa. Uma condição necessária para que não haja reencontros é que o número de pessoas por mesa seja menor que o número de mesas ($p<m$). Um participante $i \ \forall \ i=1,n$ pode ocupar uma mesa $j \ \forall \ j=1,m$ e para representar esta possibilidade considera-se uma variável binária x_{ij} que assume valor igual a 1 se o participante i ocupa a mesa j, caso contrário x_{ij} é igual a 0. Para um evento com n participantes e m mesas há um total de n x m variáveis binárias $x_{ij} \in \{0,1\} \ \forall \ i=1,n$ e $j=1,m$. O número de participantes em uma mesa j deve ser igual a p, logo os valores de $x_{ij} \ \forall \ i=1,n$ devem satisfazer a seguinte condição em cada mesa j:

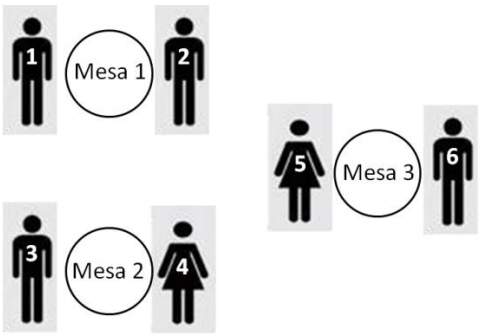
$$\sum_{i=1}^{n} x_{ij} = p \ \forall \ j=1,m \qquad (1)$$

Adicionalmente, cada participante i só pode ocupar uma mesa em cada sessão, logo os valores de $x_{ij} \ \forall \ j=1,m$ devem satisfazer a seguinte condição em cada participante:

$$\sum_{j=1}^{m} x_{ij} = 1 \forall \ i=1,\text{n} \qquad (2)$$

As mesas são numeradas de 1 a m e ao chegarem no evento os participantes escolhem as mesas aleatoriamente e definem uma configuração inicial. Dado que cada mesa comporta p pessoas, os participantes que ocupam a mesa 1 são identificados pelos números inteiros de 1 até p, os participantes da mesa 2 recebem os números (p+1) até (p+p) e assim sucessivamente até o n-ésimo participante na mesa m. Na Figura 1a tem-se uma ilustração da configuração inicial da alocação dos participantes nas mesas e ao lado, na Figura 1b, a representação matricial desta configuração.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017



a) n=6 participantes, m=3 mesas, p =2 lugares

1ª sessão	Cadeira 1	Cadeira 2
Mesa 1	1	2
Mesa 2	3	4
Mesa 3	5	6

x_{11}=1, x_{21}=1, x_{31}=0, x_{41}=0, x_{51}=0, x_{61}=0
x_{12}=0, x_{22}=0, x_{32}=1, x_{42}=1, x_{52}=0, x_{62}=0
x_{13}=0, x_{23}=0, x_{33}=0, x_{43}=0, x_{53}=1, x_{63}=1

b) Matriz de alocação

Figura 1 Configuração inicial da ocupação das mesas

Então, inicia-se a primeira sessão de contatos e após 20 minutos os participantes devem trocar de mesas, de tal forma que na nova configuração os reencontros sejam minimizados. Assim, a nova configuração de n participantes em m mesas é determinada pelo seguinte Problema de Programação Linear Inteira - PPLI, cujo objetivo consiste em minimizar o número de reencontros:

$$\underset{x,y}{Min} \sum_{k\in 1}^{m(m-1)} y_k \tag{3}$$

s.a.

$$\sum_{i=1}^{n} x_{ij} = p \;\forall\; j=1,m \tag{4}$$

$$\sum_{j=1}^{m} x_{ij} = 1 \;\forall\; i=1,n \tag{5}$$

$$\sum_{i\in mesa\ k} x_{ij} \le 1 + y_{l+(m-1)(k-1)} \;\; \forall\; k=1,m \text{ , } \forall j=1,m \text{ e } j\neq k, \;\forall\; l=1,(m-1) \tag{6}$$

$$x_{ij} \in \{0,1\} \;\forall\; i=1,n, \;\forall j=1,m \tag{7}$$

$$y_j \in \{0,1,2,3,...\} \;\; \forall j=1,m(m-1) \tag{8}$$

As m restrições na equação (4) significam que cada mesa deve ter p participantes. As n restrições em (5) indicam que cada participante pode ocupar apenas uma mesa em cada sessão. As $m(m-1)$ restrições em (6) impedem os reencontros de participantes que ocuparam a mesma mesa na configuração inicial. A restrição em (7) indica que as variáveis x_{ij} são binárias e a restrição em (8) informa que as variáveis y_k são inteiras. As variáveis binárias indicam a presença ou ausência do participante i na mesa j. Por sua vez, a variável y_k conta o número de reencontros e, portanto, a soma destas variáveis deve ser minimizada, conforme indicado na

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

função objetivo em (3). No caso com n=6 participantes, m=3 mesas e p=2, a nova configuração é determinada pelo seguinte PPLI:

$$\underset{x,y}{Min}\ y_1 + y_2 + y_3 + y_4 + y_5 + y_6 \quad (9)$$

s.a.

$$x_{11} + x_{21} + x_{31} + x_{41} + x_{51} + x_{61} = 2 \quad \text{(mesa 1)} \quad (10)$$

$$x_{12} + x_{22} + x_{32} + x_{42} + x_{52} + x_{62} = 2 \quad \text{(mesa 2)} \quad (11)$$

$$x_{13} + x_{23} + x_{33} + x_{43} + x_{53} + x_{63} = 2 \quad \text{(mesa 3)} \quad (12)$$

$$x_{11} + x_{12} + x_{13} = 1 \quad \text{(participante 1)} \quad (13)$$

$$x_{21} + x_{22} + x_{23} = 1 \quad \text{(participante 2)} \quad (14)$$

$$x_{31} + x_{32} + x_{33} = 1 \quad \text{(participante 3)} \quad (15)$$

$$x_{41} + x_{42} + x_{43} = 1 \quad \text{(participante 4)} \quad (16)$$

$$x_{51} + x_{52} + x_{53} = 1 \quad \text{(participante 5)} \quad (17)$$

$$x_{61} + x_{62} + x_{63} = 1 \quad \text{(participante 6)} \quad (18)$$

$$x_{12} + x_{22} \le y_1 + 1 \quad \text{(evita reencontro entre participantes 1 e 2 na mesa 2)} \quad (19)$$

$$x_{13} + x_{23} \le y_2 + 1 \quad \text{(evita reencontro entre participantes 1 e 2 na mesa 3)} \quad (20)$$

$$x_{31} + x_{41} \le y_3 + 1 \quad \text{(evita reencontro entre participantes 3 e 4 na mesa 1)} \quad (21)$$

$$x_{33} + x_{43} \le y_4 + 1 \quad \text{(evita reencontro entre participantes 3 e 4 na mesa 3)} \quad (22)$$

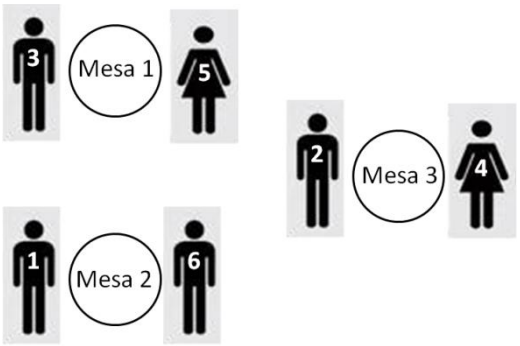
$$x_{51} + x_{61} \le y_5 + 1 \quad \text{(evita reencontro entre participantes 5 e 6 na mesa 1)} \quad (23)$$

$$x_{52} + x_{62} \le y_6 + 1 \quad \text{(evita reencontro entre participantes 5 e 6 na mesa 2)} \quad (24)$$

$$x_{11} = 0, x_{21} = 0, x_{32} = 0, x_{42} = 0, x_{53} = 0, x_{63} = 0 \quad \text{(impede revisitar mesas)} \quad (25)$$

$$x_{ij} \in \{0,1\} \ \forall\ i=1,6\ ,\ \forall j=1,3 \quad y_j \in \{0,1,2,3\} \quad \forall j=1,6 \quad (26)$$

Vale ressaltar que o conjunto formado pelas restrições (19) até (24) foi gerado diretamente da solução ilustrada na Figura 1.b e tem por objetivo evitar os reencontros entre participantes na configuração inicial. A solução do PPLI acima resulta na solução ótima ilustrada na Figura 2. Na nova solução a soma y_1+...+y_6 é nula, logo, não houve reencontros.



a) n=6 participantes, m=3 mesas, p =2 lugares

2ª sessão	Cadeira 1	Cadeira 2
Mesa 1	3	5
Mesa 2	1	6
Mesa 3	2	4

x_{11}=0, x_{21}=0, x_{31}=1, x_{41}=0, x_{51}=1, x_{61}=0
x_{12}=1, x_{22}=0, x_{32}=0, x_{42}=0, x_{52}=0, x_{62}=1
x_{13}=0, x_{23}=1, x_{33}=0, x_{43}=1, x_{53}=0, x_{63}=0
y_1=0, y_2=0, y_3=0, y_4=0, y_5=0, y_6=0

b) Matriz de alocação

Figura 2 Configuração na segunda sessão

Após 20 minutos mais uma troca de mesas deve ser realizada e uma nova configuração deve ser encontrada. Novamente o PPLI deve ser resolvido, porém com um conjunto de restrições adicionais que evitem os reencontros com todas as configurações anteriores. Tal conjunto de restrições é semelhante ao conjunto em (6) contendo $m(m\text{-}1)$ restrições, porém é derivado diretamente da configuração da sessão imediatamente anterior. Assim, ao longo da evolução do evento o conjunto de restrições do PPLI é acrescido de $m(m\text{-}1)$ restrições a cada sessão. Adicionalmente, todas as variáveis x_{ij} que assumiram valores unitários nas sessões anteriores são fixadas em zero nas novas sessões para que cada participante visite cada mesa uma única vez. Para obter a configuração da terceira e última sessão do caso com n=6 participantes, m=3 mesas e p=2 basta observar a configuração ilustrada na Figura 2 para gerar o seguinte conjunto de restrições a serem adicionadas ao PPLI definido a partir da função objetivo em (9):

$$x_{32} + x_{52} \leq y_7 + 1 \text{ (evita reencontro entre participantes 3 e 5 na mesa 2)} \quad (27)$$
$$x_{33} + x_{53} \leq y_8 + 1 \text{ (evita reencontro entre participantes 3 e 5 na mesa 3)} \quad (28)$$
$$x_{11} + x_{61} \leq y_9 + 1 \text{ (evita reencontro entre participantes 1 e 6 na mesa 1)} \quad (29)$$
$$x_{13} + x_{63} \leq y_{10} + 1 \text{ (evita reencontro entre participantes 1 e 6 na mesa 3)} \quad (30)$$
$$x_{21} + x_{41} \leq y_{11} + 1 \text{ (evita reencontro entre participantes 2 e 4 na mesa 1)} \quad (31)$$
$$x_{22} + x_{42} \leq y_{12} + 1 \text{ (evita reencontro entre participantes 2 e 4 na mesa 2)} \quad (32)$$

Por meio do R foi possível construir um programa capaz de automatizar a montagem do PPLI em cada sessão do evento e resolve-lo com o pacote *Rsymphony* (Harter et al, 2016), uma interface de alto nível entre o R e o COIN-OR *Symphony solver*. Todo o código em R encontra-se em um único arquivo texto com extensão Rexec. Algumas manipulações no *path* do sistema possibilitaram associar a extensão Rexec com o interpretador do R (Rscript.exe). O procedimento para associar a extensão Rexec ao Rscript encontra-se detalhado no *post Making R Files Executable* (*under Windows*) disponível no R-bloggers (Meissner, 2015). Assim, o arquivo com o código em R funciona como um arquivo executável. Os dados de entrada/saída do programa são armazenados em arquivos textos. Os dados de entrada incluem o nº de participantes do evento, o nº de pessoas por mesa e o nº de sessões. Já os resultados são gravados em arquivos .XLSX e incluem a sequência de mesas visitadas por cada participante (roteiros), um quadro com as alocações dos participantes em cada sessão

e o nº de reencontros. Para facilitar a utilização do programa foi construída uma planilha MS Excel com botões e macros (Jellen & Syrstad, 2008) programados para preparar o arquivo de entrada de dados, disparar a execução do arquivo Rexec e importar os resultados gerados pelo R para dentro da planilha. A comunicação entre o R e o MS Excel foi realizada por meio de arquivos .TXT (dados) e .XLSX (resultados) e contou com a interface proporcionada pelo pacote xlsx (Dragulescu, 2014). Adicionalmente, a planilha MS Excel disponibiliza recursos para a impressão de etiquetas com os roteiros a serem percorridos por cada participante, conforme ilustrado pelo botão "gera etiquetas" na Figura 3. Além do moderador, algumas mesas podem contar com um convidado altamente experiente e qualificado (âncora), cujo objetivo consiste em orientar e estimular a troca de ideias. O moderador e o âncora não trocam de mesas.

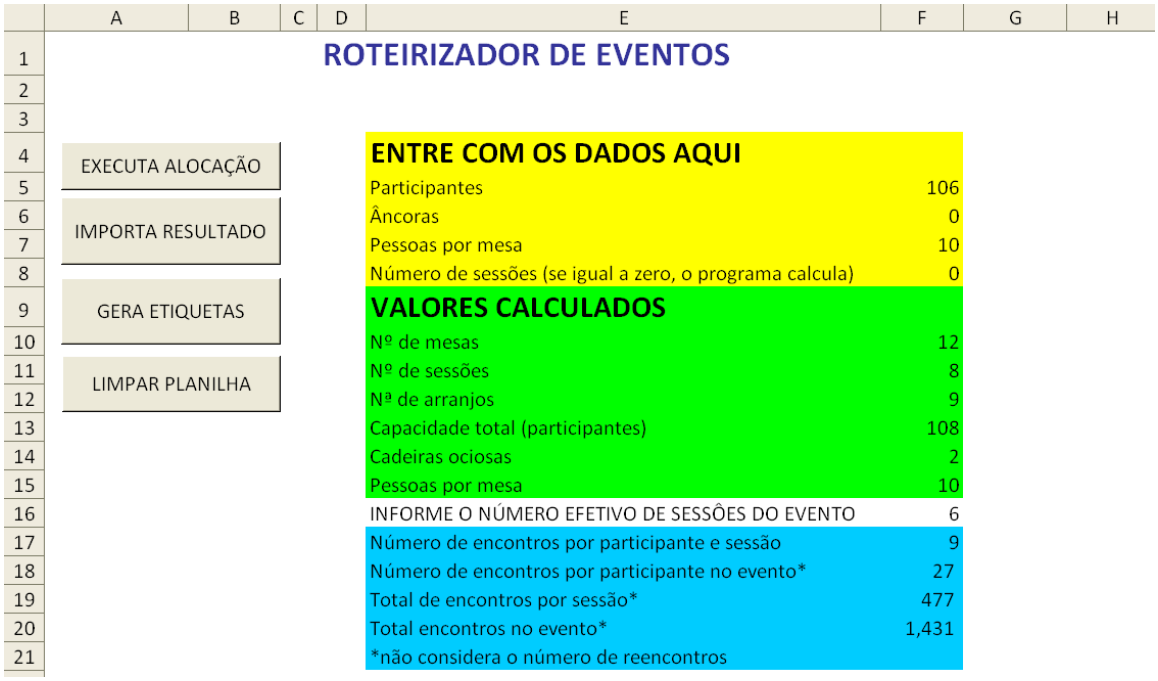


Figura 3 Planilha MS Excel para entrada de dados e visualização de resultados

Resultados e Discussão:

Para ilustrar os resultados gerados pelo aplicativo considere um evento com 106 participantes e 12 mesas com 10 lugares cada uma. Ressalta-se que em cada mesa há um moderador que pertence à organização do evento, assim, cada mesa tem apenas 9 lugares disponíveis para os participantes. O evento é programado para uma duração total de cerca de 2 horas (120 minutos) e cada sessão tem uma duração

de cerca de 20 minutos. Assim, as simulações realizadas consideraram um total de 6 seções, ou seja, cada participante percorre 6 mesas.

Na Figura 4 apresentam-se as alocações dos participantes nas 12 mesas ao longo das 6 sessões. Nos quadros ilustrados na Figura 4 são indicadas apenas as 9 cadeiras livres em cada mesa. Vale ressaltar que o resultados gerados pelo programa permitem acomodar até 108 pessoas, logo há duas cadeiras ociosas. Os quadros permitem que os organizadores do evento visualizem rapidademente as composições das mesas ao longo das sessões.

Sessão 1

	cadeira.1	cadeira.2	cadeira.3	cadeira.4	cadeira.5	cadeira.6	cadeira.7	cadeira.8	cadeira.9
mesa 1	1	2	3	4	5	6	7	8	9
mesa 2	10	11	12	13	14	15	16	17	18
mesa 3	19	20	21	22	23	24	25	26	27
mesa 4	28	29	30	31	32	33	34	35	36
mêsa 5	37	38	39	40	41	42	43	44	45
mesa 6	46	47	48	49	50	51	52	53	54
mesa 7	55	56	57	58	59	60	61	62	63
mesa 8	64	65	66	67	68	69	70	71	72
mesa 9	73	74	75	76	77	78	79	80	81
mesa 10	82	83	84	85	86	87	88	89	90
mesa 11	91	92	93	94	95	96	97	98	99
mesa 12	100	101	102	103	104	105	106	107	108

Sessão 2

	cadeira.1	cadeira.2	cadeira.3	cadeira.4	cadeira.5	cadeira.6	cadeira.7	cadeira.8	cadeira.9
mesa 1	36	45	54	63	72	81	90	97	108
mesa 2	27	44	53	62	71	80	89	96	107
mesa 3	29	43	52	61	70	79	88	95	106
mesa 4	6	18	51	60	69	78	87	94	105
mêsa 5	3	17	26	59	68	77	86	93	104
mesa 6	9	16	25	28	67	76	85	92	103
mesa 7	8	15	24	35	42	75	84	91	102
mesa 8	7	14	23	34	38	50	74	83	101
mesa 9	5	13	19	33	41	46	58	82	100
mesa 10	2	10	20	32	37	49	55	66	98
mesa 11	1	12	22	31	40	48	57	65	73
mesa 12	4	11	21	30	39	47	56	64	99

Sessão 3

	cadeira.1	cadeira.2	cadeira.3	cadeira.4	cadeira.5	cadeira.6	cadeira.7	cadeira.8	cadeira.9
mesa 1	18	19	34	44	48	76	84	93	106
mesa 2	7	29	41	55	72	75	86	92	105
mesa 3	8	17	30	37	53	60	81	83	103
mesa 4	4	12	20	54	58	68	80	95	101
mêsa 5	21	36	50	61	69	73	85	91	100
mesa 6	6	10	24	31	38	70	77	99	108
mesa 7	9	13	27	32	43	47	65	87	97
mesa 8	2	16	22	33	39	51	63	89	102
mesa 9	3	15	23	28	40	56	71	90	98
mesa 10	14	25	35	46	57	64	79	96	104
mesa 11	5	11	42	49	59	67	78	88	107
mesa 12	1	26	45	52	62	66	74	82	94

Sessão 4

	cadeira.1	cadeira.2	cadeira.3	cadeira.4	cadeira.5	cadeira.6	cadeira.7	cadeira.8	cadeira.9
mesa 1	11	23	29	37	51	62	68	96	100
mesa 2	5	26	40	54	60	79	85	99	102
mesa 3	6	41	47	67	73	80	84	98	104
mesa 4	8	10	22	43	59	64	74	90	92
mêsa 5	16	27	31	46	56	66	75	83	106
mesa 6	3	13	21	35	44	63	87	95	101
mesa 7	2	12	19	30	45	50	71	78	86
mesa 8	1	18	25	32	53	58	88	91	108
mesa 9	14	36	42	48	55	70	89	94	103
mesa 10	4	28	38	52	65	81	93	105	107
mesa 11	9	15	20	34	39	61	72	77	82
mesa 12	7	17	24	33	49	57	69	76	97

Sessão 5

	cadeira.1	cadeira.2	cadeira.3	cadeira.4	cadeira.5	cadeira.6	cadeira.7	cadeira.8	cadeira.9
mesa 1	14	15	21	31	43	52	60	80	86
mesa 2	1	24	30	39	46	68	81	88	98
mesa 3	5	45	51	55	64	76	89	91	101
mesa 4	11	25	44	61	71	77	83	97	102
mêsa 5	7	10	22	28	47	58	78	82	96
mesa 6	2	17	34	41	56	65	79	94	107
mesa 7	18	26	33	38	66	73	90	95	103
mesa 8	4	13	36	40	49	62	75	92	104
mesa 9	8	20	29	50	57	67	87	93	108
mesa 10	9	12	23	42	53	63	69	99	106
mesa 11	3	16	32	54	70	74	84	100	105
mesa 12	6	19	27	35	37	48	59	72	85

Sessão 6

	cadeira.1	cadeira.2	cadeira.3	cadeira.4	cadeira.5	cadeira.6	cadeira.7	cadeira.8	cadeira.9
mesa 1	27	32	41	57	70	74	78	85	91
mesa 2	2	48	56	67	69	82	83	95	108
mesa 3	4	13	31	42	50	77	90	96	105
mesa 4	1	15	19	38	47	49	63	79	103
mêsa 5	8	33	52	55	71	72	84	99	101
mesa 6	18	20	30	40	59	75	89	97	100
mesa 7	7	16	37	64	80	81	88	94	104
mesa 8	9	12	21	26	29	46	76	98	107
mesa 9	10	25	35	39	54	62	65	86	106
mesa 10	3	11	24	34	45	51	60	73	92
mesa 11	6	17	23	36	43	44	58	66	102
mesa 12	5	14	22	28	53	61	68	87	93

Figura 4 Alocação dos participantes (cada número representa um participante)

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Na Figura 5 tem-se um extrato da tabela contendo os roteiros com a sequências de mesas a serem visitadas pelos 9 participantes que iniciaram na mesa 1. O programa também gera uma planilha com as etiquetas contendo os roteiros que devem ser seguidos pelos participantes do evento. Na Figura 6 tem-se um extrato da planilha com as etiquetas do três primeiros participantes. Conforme ilustrado na Tabela 1, para um evento com as mesmas dimensões do exemplo os reencontros começam a ocorrer a partir da quarta sessão.

Tabela 1 Número acumulado de reencontros

Sessão	1ª	2ª	3ª	4ª	5ª	6ª
Reencontros	0	0	0	5	22	64

participantes	sessão.1	sessão.2	sessão.3	sessão.4	sessão.5	sessão.6
1	1	11	12	8	2	4
2	1	10	8	7	6	2
3	1	5	9	6	11	10
4	1	12	4	10	8	3
5	1	9	11	2	3	12
6	1	4	6	3	12	11
7	1	8	2	12	5	7
8	1	7	3	4	9	5
9	1	6	7	11	10	8

Figura 5 Extrato da tabela de roteiros, cada linha mostra a sequência de mesas a serem visitadas por um participante

Participante 1			Participante 2			Participante 3	
Sessão 1 - Mesa	1		Sessão 1 - Mesa	1		Sessão 1 - Mesa	1
Sessão 2 - Mesa	11		Sessão 2 - Mesa	10		Sessão 2 - Mesa	5
Sessão 3 - Mesa	12		Sessão 3 - Mesa	8		Sessão 3 - Mesa	9
Sessão 4 - Mesa	8		Sessão 4 - Mesa	7		Sessão 4 - Mesa	6
Sessão 5 - Mesa	2		Sessão 5 - Mesa	6		Sessão 5 - Mesa	11
Sessão 6 - Mesa	4		Sessão 6 - Mesa	2		Sessão 6 - Mesa	10

Figura 6 Exemplos de etiquetas distribuídas aos participantes

Conclusão:

Apresentou-se um modelo de Programação Linear Inteira para organizar fóruns empresariais. O modelo proposto visa minimizar o número de reencontros de forma a contribuir para o maior número de contatos entre os participantes. O modelo foi implementado em R e uma planilha MS Excel serve de interface com o usuário. O aplicativo já foi utilizado em quatro eventos e a experiência nestas oportunidades tem recebido elogios dos participantes. Além de reduzir o número de reencontros, o

aplicativo agilizou a organização dos fóruns em que foi utilizado. Os resultados obtidos mostram que a metodologia proposta é promissora, porém mais investigações devem ser realizadas no sentido de ampliar a capacidade do programa visando a sua aplicação na organização de eventos com um grande número de participantes.

Referências:

BARTLETT, A.; CHARTIER, T.P.; LANGVILLE, A.N.: RANKIN, T.D. An Integer Programming Model for the Sudoku Problem, **Convergence,** v. 8, May, 2008.

BELLOWS, M. L.; PETERSON, D.L. Finding an optimal seating chart, **Annals of Improbable Research**, February, 2012.

DRAGULESCU, A. A. xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7., 2014. Disponível em: < http://CRAN.R-project.org/package=xlsx >. Acesso em: 20 jan. 2017.

HARTER, R.; HORNIK, K.; THEUSSL, L.; SZYMANSKI, C. Rsymphony: SYMPHONY in R. R package version 0.1-22, 2016. Disponível em: < http://CRAN.R-project.org/package=Rsymphony >. Acesso em: 20 jan. 2017.

JELLEN, B.; SYRSTAD, T. **VBA e Macros para Microsoft Office Excel 2007**, Prentice Hall do Brasil, 2008.

JOHNSON, Steven. **De onde vêm as boas ideias**. Rio de Janeiro: Zahar, 2011.

MEISSNER, P. Making R Files Executable (under Windows). **R-Bloggers**, Feb., 2015. Disponível em: < https://www.r-bloggers.com/making-r-files-executable-under-windows/ >. Acesso em: 20 jan. 2016.

R CORE Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. Disponível em: <https://www.R-project.org/>. Acesso em: 20 jan. 2017.

RAGSDALE, C. T. **Spreadsheet Modeling & Decision Analysis: A Practical Introduction to Management Science,** 4. ed. Mason: Southern-Western, 2004.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 19 - PREVISÃO DE SÉRIES TEMPORAIS DE ACIDENTES EM UMA CONCESSIONÁRIA DE RODOVIA BRASILEIRA POR MEIO DO SOFTWARE R

Carla Martins Floriano[14]
Brunno e Souza Rodrigues[15]
Valdecy Pereira[16]

Resumo

Uma das principais causas de mortes no mundo é o trânsito. Não obstante estão as rodovias brasileiras, onde a busca pela redução de acidentes constitui uma tarefa árdua e contínua pelos órgãos públicos e privados do país. Com base nesse cenário, este artigo tem por objetivo aplicar um modelo de previsão de séries temporais na estimativa do número de acidentes de trânsito de uma concessão rodoviária brasileira, no último semestre de 2015, e aferir se a previsão está em harmonia com os dados reais. Para tanto, os dados foram analisados através do pacote *forecasting* função ets do *software* estatístico R 3.3.2 e *Rstudio* 0.99.903. Com isso, a análise indicou que o modelo mais adequado aos dados é o Alisamento Exponencial Simples com Erro Multiplicativo e a previsão de acidentes mostrou-se bem semelhante à original.

Palavras-Chave: Acidente de trânsito, previsão, série temporal.

Abstract

One of the main causes of death in the world is traffic accidents. Nevertheless, there are the Brazilian highways, where the search for traffic accidents reduction is an arduous and continuous task persecuted by the public and private agencies of the country. Based on this scenario, this article aims to apply a time series forecast model to estimate the number of traffic accidents in a Brazilian road concession in the last half of 2015 and to assess if the forecast is in harmony with the actual data. For that, the data were analyzed through the forecasting function ets package of statistical software R 3.3.2 and Rstudio 0.99.903. Thus, the analysis indicated that the most adequate model to the data is the Simple Exponential Smoothing with Multiplicative Error and the traffic accidents forecasting was very similar to the original one.

Keywords: Traffic accident, forecast, time series.

[14] UFF, email carlafloria@gmail.com
[15] UFF, email brunno.esr@gmail.com
[16] UFF, email valdecy.pereira@gmail.com

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

Uma das principais causas de mortes no mundo é o trânsito, segundo relatório da Organização Mundial da Saúde (2015), estando entre as principais causas de morte quando a questão são os jovens (entre 15 e 29 anos).

A busca pela redução de acidentes nas estradas brasileiras é, sem dúvida, uma tarefa árdua e contínua pelos órgãos públicos e privados do país. De acordo com o relatório anual de 2015 da Associação Brasileira de Concessionárias de Rodovias (ABCR), o foco na redução de acidentes é bastante expressivo nos últimos anos, resultado de grandes investimentos em infraestrutura e de campanhas educativas nas estradas brasileiras.

Objetivo

Diante do exposto, o objetivo central deste artigo é aplicar um modelo de previsão de séries temporais para estimar o número de acidentes de trânsito para os últimos seis meses de 2015, baseado em dados de acidentes coletados entre 2010 e 2015 de uma importante concessão rodoviária brasileira (anônima). Ao final será realizada uma comparação entre o número de acidentes previsto e o real.

Material e Métodos:

Os dados de acidentes de trânsito utilizados levaram em consideração três tipos de acidentes: (i) acidentes sem vítimas; (ii) acidentes com vítimas e; (iii) acidentes com vítimas fatais. Tendo os mesmos ocorridos diariamente entre janeiro de 2010 e dezembro de 2015. Vale ressaltar que para a confecção deste artigo não foi feita qualquer distinção entre os tipos de acidentes, ou seja, os tipos de acidentes foram somados e tabulados em uma planilha *Microsoft Excel*®, apresentando o total de acidentes por mês e por ano (Tabela 1).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Tabela 1 – Número de acidentes

ACIDENTES	Ano						Total
Mês	2010	2011	2012	2013	2014	2015	
janeiro	147	203	173	117	99	60	799
fevereiro	157	173	147	119	107	60	763
março	185	173	198	136	116	80	888
abril	150	202	151	129	117	85	834
maio	140	231	180	132	102	99	884
junho	164	227	147	109	96	88	831
julho	154	242	164	145	95	93	893
agosto	196	210	155	156	100	70	887
setembro	195	249	162	134	113	103	956
outubro	183	228	153	125	103	93	885
novembro	225	206	169	133	72	103	908
dezembro	245	237	153	104	85	68	892
Total	2141	2581	1952	1539	1205	1002	10420

Os dados de janeiro de 2010 a junho de 2015 foram analisados pelo pacote *forecasting* função ets do *software* estatístico R 3.3.2 e *Rstudio* 0.99.903. Os dados de julho a dezembro de 2015 servirão como base de comparação ao final do artigo.
O modelo ETS (***E****rror,* ***T****rend,* ***S****easonal*), ou seja, erro, tendência e sazonalidade proposto por Hyndman et al. (2008), conforme Tabela 2, consiste em classificar a base de dados em um modelo temporal adequado. Para a série temporal em análise o resultado obtido foi o ETS (M, N, N).

Tabela 2 – Modelo ETS

Seasonal / Trend	N (None)	A (Additive)	M (Multiplicative)
N (None)	(N,N)	(N,A)	(N,M)
A (Additive)	(A,N)	(A,A)	(A,M)
Ad (Additive Damped)	(Ad,N)	(Ad,A)	(Ad,M)
M (Multiplicative)	(M,N)	(M,A)	(M,M)
Md (Multiplicative Damped)	(Md,N)	(Md,A)	(Md,M)

Desta forma, a primeira letra do resultado (M) significa o erro do modelo que neste caso é o multiplicativo, a segunda letra (N) é a tendência e a última (N) é a sazonalidade, que em ambos os casos é nula. Assim, o modelo mais adequado aos dados é o Alisamento Exponencial Simples com Erro Multiplicativo.

A análise dos dados iniciou-se com a transformação dos dados em uma série temporal e a partir daí, realizou-se uma exploração mais criteriosa dos mesmos. Foi

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

utilizado o comando *summary,* na qual são apontadas características como mediana, média e mínimo-máximo da amostra (Tabela 3).

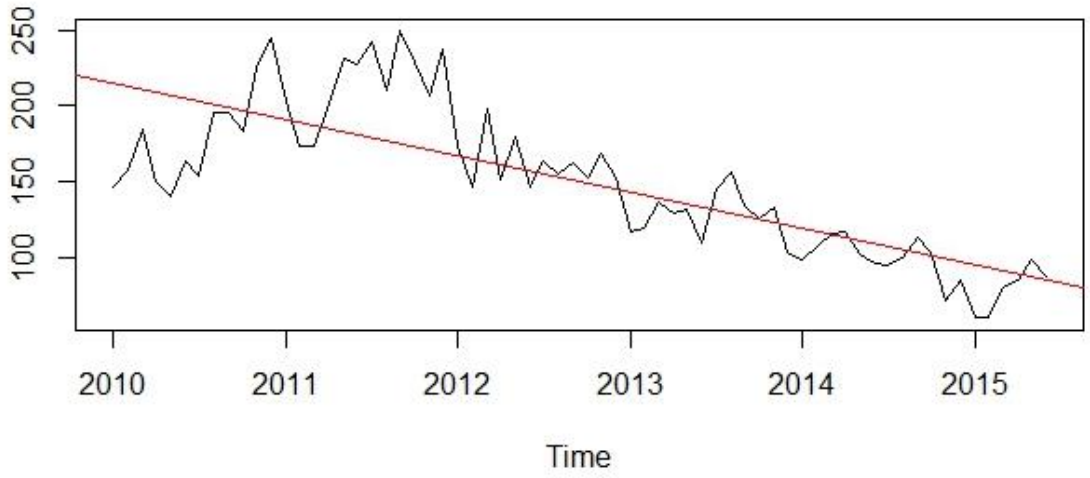
Tabela 3 – Sumário da amostra

Mediana	Média	[Min; Max]
148.5	149.8	[60; 2149]

Min=valor mínimo; Max=valor máximo

Em seguida, foi possível observar que a tendência de ocorrência de acidentes é negativa, posto que, a evolução dos acidentes vem demonstrando queda ao longo dos anos, conforme apresentado na Figura 1. Nesta Figura, a linha preta representa as ocorrências de acidentes ocorridas no período de janeiro de 2010 a junho de 2015 e a linha vermelha representa a tendência desses acidentes.

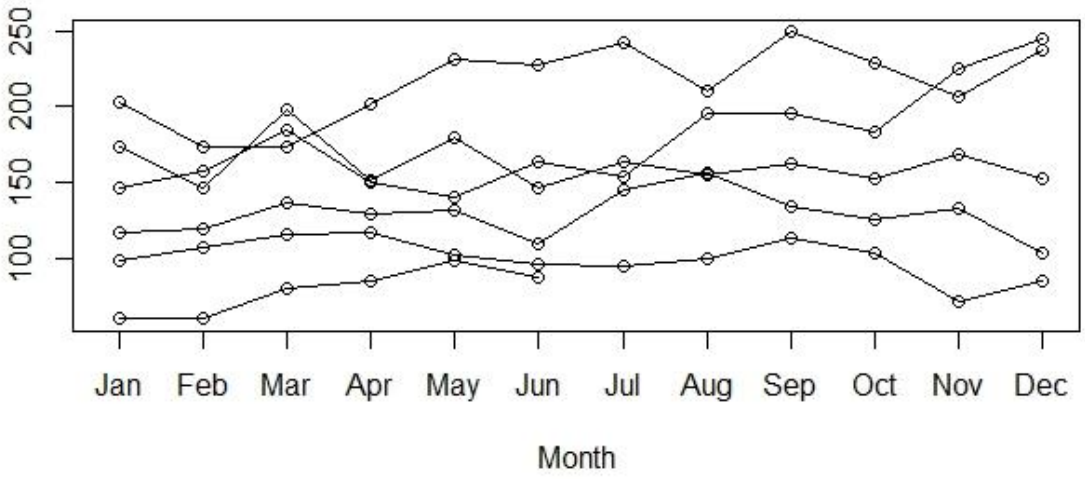
Figura 1 – Gráfico de séries temporais de acidentes de trânsito anual



A Figura 2 mostra o comportamento dos acidentes por mês ao logo dos cinco anos analisados. Nota-se que há uma tendência de queda no número de acidentes, conforme analisado no gráfico anterior (Figura 1).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figura 2 – Gráfico de séries temporais de acidentes de trânsito mensal



Resultados e Discussão:

Através da análise dos erros, segundo Tabela 4, podemos perceber que o Erro Médio (ME) dos dados mostrou-se bastante satisfatório, pois apresenta um valor bem pequeno (menor que zero). O Erro Médio Percentual (MPE) de -2,934% denota uma baixa incidência de erro quando comparamos os valores previstos com os valores reais. No entanto, o MPE pode estar mascarado, pois os valores negativos anulam os valores positivos. Porém, através do Erro Absoluto Médio Percentual (MAPE) é possível observar um erro de 12,538% o que ratifica o fato dos valores previstos estarem próximos dos valores reais.

Tabela 4 – Tabela de erros dos dados

ME	-1.581	Erro Médio
MPE	-2.934	Erro Médio Percentual
RMSE	22.054	Raiz do Erro Quadrático Médio
MAE	17.974	Erro Absoluto Médio
MAPE	12.538	Erro Absoluto Médio Percentual

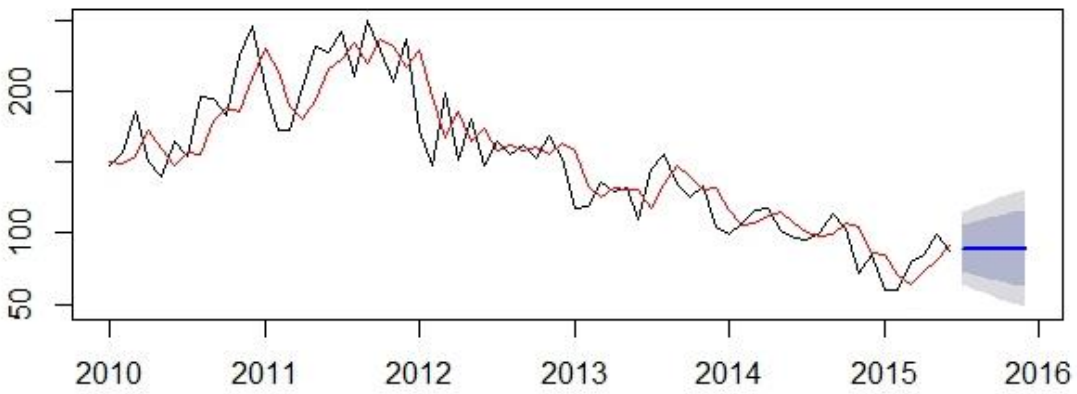
A Tabela 5 mostra as previsões de Julho de 2015 a Dezembro de 2015, assim como seus respectivos limites de confiança de 80% e 95%. A previsão pontual para cada mês foi de 89 acidentes.

Tabela 5 – Previsões e respectivos limites de confiança

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jul 2015	89.35541	73.0399	105.671	64.4029	114.308
Aug 2015	89.35541	70.376	108.335	60.3289	118.382
Sep 2015	89.35541	68.027	110.684	56.7364	121.974
Oct 2015	89.35541	65.898	112.813	53.4804	125.23
Nov 2015	89.35541	63.9336	114.777	50.4761	128.235
Dec 2015	89.35541	62.0981	116.613	47.6689	131.042

A Figura 3 exibe de forma gráfica a tabela acima. Na figura, a linha preta representa o número real dos acidentes, enquanto que a linha vermelha representa os valores ajustados, ou seja, previstos. Vale ressaltar, que ambas as linhas encontram-se próximas, o que aumenta a acurácia do modelo. A parte final do gráfico representa a previsão de acidentes de Julho a Dezembro de 2015, na qual a linha reta azul é a previsão pontual por mês e as áreas sombreadas são os intervalos de confiança de 80% (cinza mais forte) e 95% (cinza mais claro).

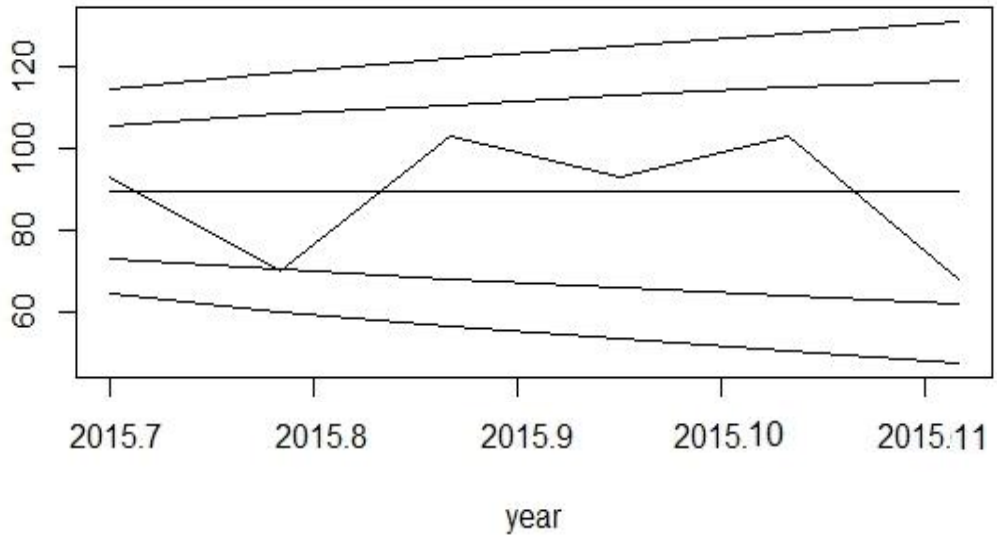
Figura 3 – Previsão para o ETS (M,N,N)



Após as análises acima, se faz necessária uma comparação entre os dados reais e os previstos pelo pacote *forecasting* função ets, conforme apresentada na Figura 4. Nela é possível observar que os dados reais estão contidos no intervalo de confiança de 80% previsto pelo modelo.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figura 4 – Real x Previsto (Jul- Dez 2015)



Conclusão:

O artigo foi capaz de atingir o objetivo central ao qual se propôs a alcançar: aplicar um modelo de previsão de séries temporais para estimar o número de acidentes de trânsito dos últimos seis meses de 2015 e aferir se a previsão está harmônica com a realidade.

Através das análises e dos gráficos notou-se que o modelo ajustou-se bem próximo à realidade. A Figura 4 prova esta afirmação, já que a reta dos dados reais de acidentes limitou-se ao intervalo de confiança calculado pelo modelo de Alisamento Exponencial Simples com Erro Multiplicativo.

Referências:

ABCR. Associação Brasileira de Concessionárias de Rodovias. Relatório anual 2015. Disponível em: <http://www.abcr.org.br/RelatoriosAnuais/Digital2015/>Acesso em: 10 jan. 2017

HYNDMAN, Rob J.; KOELHER, Anne B.; ORD, J.Keith; SNYDER, Ralph D. Forecasting with exponential Smoothing – The State Space Approach. Berlim: Springer, 2008.

MAKRIDAKIS, Spyros; WHEELWRIGHT, Steven C.; HYNDMAN, Rob J. Forecasting: Methods and Applications. John Wiley & Sons, Inc., 3rd ed.,1983.

OMS. Organização Mundial de Saúde. Relatório Global sobre o Estado da Segurança Viária 2015. Disponível em:

<http://www.who.int/violence_injury_prevention/road_safety_status/2015/Summary_GSRRS2015_POR.pdf?ua=1> Acesso em: 10 jan. 2017

PEREIRA, Valdecy. Project: Multivariate Data Analysis. File: R-MVDA-08-LR-B.pdf. GitHub repository: <https://github.com/Valdecy/Multivariate_Data_Analysis> . 2016

Script:

```
# Exponential smoothing state space model.

# A função realiza a previsão da série temporal my_data

ets_ts<- ets(my_data)

# A função retorna média, mediana, medidas de erros, além de outras
informações como o modelo a ser aplicado, neste caso o ETS (M,N,N)

summary(ets_ts)

# A função realiza a previsão para 6 meses

predicted <- forecast(ets_ts, h=6)

# A função realiza a previsão ajustadapara todo período

my_ets_ts<- fitted(ets_ts)

# A função plota o gráfico com dados reais (jan 2010 a junho de
2015) e com a previsão(julho a dezembro de 2015)

plot(predicted)

# A função plota o gráfico com a previsão ajustada

lines(my_ets_ts, col=2)

# A função plota o gráfico comparando a série temporal prevista com
a série temporal real

ts.plot(predicted$mean,my_data_2, my_data_2_up,
my_data_2_upper, my_data_2_lo, my_data_2_lower, gpars =
list(xlab = "year", ylab = "accidents", lty = c(1:1)))
```

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 21 - UMA ANÁLISE MULTICRITÉRIO DOS INDICADORES ECONÔMICO-FINANCEIROS DE EMPRESAS DA CONSTRUÇÃO CIVIL

Alessandra Simão[17]
Luciane Ferreira Alcoforado[18]
Leonardo Filgueira[19]

Resumo

Os investidores utilizam de diversos métodos para avaliar o desempenho de uma empresa. Uma das formas mais comuns é a análise dos demonstrativos contábeis para obtenção dos indicadores econômico-financeiros e escolha das melhores empresas. Esse processo pode ser utilizado o Método AHP que hierarquiza as empresas. Como questão problema, levanta-se: Quais empresas do setor de construção civil listadas na BM&FBOVESPA obtiveram melhor desempenho econômico-financeiro no ano de 2015 utilizando o Método AHP? O trabalho objetiva: Hierarquizar as empresas brasileiras do setor de construção civil listadas na BM&FBOVESPA de acordo com seu desempenho econômico-financeiro no ano de 2015 utilizando o Método AHP. Metodologicamente, adota-se objetivo descritivo com abordagem quantitativa com aplicação do Método AHP e o Software R. A amostra consiste em 12 empresas do setor de construção civil listadas BM&FBOVESPA. São analisados 17 índices e como principal resultado verifica-se que de acordo com o Método AHP a empresa Construtora Adolpho Lindenberg, seguida pelas empresas Ez Tec Empreendimentos e Participações e Rodobens Negócios Imobiliários apresentam os melhores desempenhos. Os resultados apurados convergem com o que especialistas defendem na teoria, em que a empresa pode possuir bom desempenho em um indicador, contudo não necessariamente apresentará o mesmo potencial em outro indicador.

Palavras-Chave: Analytic Hierarchy Process, Ranking, Empresas construtoras

Abstract

Investors use a variety of methods to evaluate a company's performance. One of the most common forms is the analysis of the financial statements to obtain the economic-financial indicators and the choice of the best companies. This process can be used the AHP Method that hierarchizes the companies. As a problem question, the following stand out: Which companies in the civil construction sector listed on BM&FBOVESPA obtained the best economic and financial performance in the year 2015 using the AHP Method? The objective of this paper is to: Hierarchize the Brazilian civil construction companies listed on the BM&FBOVESPA according to their economic and financial performance in the year 2015 using the AHP Method. Methodologically, it adopts a descriptive objective with a quantitative approach with application of the AHP Method and Software R. The sample consists of 12 companies from the civil construction sector listed BM & FBOVESPA. We analyze 17 indices and as main result it is verified that according to the AHP Method, the company Construtora Adolpho Lindenberg, followed by the companies Ez Tec Empreendimentos e Participações and Rodobens Negócios Imobiliários present the best performances. The results obtained converge

[17] UFF – Programa de Pós Grad. Eng. Civil - alessandra_simao@id.uff.br
[18] UFF – Programa de Pós Grad. Eng. Civil / Dep. Estatística – lucianealcoforado@gmail.com
[19] UFF – Grad. Estatística – leo-filgueira@hotmail.com

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

with what experts argue in theory, where the company can perform well in one indicator, but it does not necessarily have the same potential in another indicator.

Keywords: Analytic Hierarchy Process, Ranking, Construction companies

Introdução

Os investidores utilizam de diversos métodos para avaliar o desempenho de uma empresa. Uma das formas mais comuns é a análise dos demonstrativos contábeis para obtenção dos indicadores econômico-financeiros e escolha das melhores empresas.

A análise das demonstrações contábeis de empresas de determinado segmento, possibilita conhecer a situação das mesmas, entretanto, é necessário realizar comparações entre os dados obtidos para o estabelecimento da classificação. Esse processo pode ser realizado por meio de um *ranking*. Para ser estabelecido, deve-se considerar alguns critérios, e, por meio deles, é possível identificar a importância de cada elemento em relação ao seu conjunto.

Para estabelecer um *ranking*, pode-se atribuir pesos, isto é, valores que caracterizam a importância de cada elemento dentro do conjunto analisado. Esta classificação pode ser realizada por meio da aplicação do método de Análise Hierárquica, conhecido como *Analytic Hierarchy Process* (AHP), desenvolvido por (SAATY, 1991).

Diante desse contexto, perante a importância de análise do desempenho econômico-financeiro, surge o seguinte questionamento: Quais empresas do setor de construção civil listadas na BM&FBOVESPA obtiveram melhor desempenho econômico-financeiro no ano de 2015 utilizando o Método AHP?

Objetivo

Hierarquizar as empresas brasileiras do setor de construção civil listadas na BM&FBOVESPA de acordo com seu desempenho econômico-financeiro no ano de 2015 utilizando o Método AHP.

Material e Métodos:

Este estudo pode ser classificado como descritivo, com abordagem quantitativa por aplicação do método AHP com o auxílio do software R. A amostra consiste em 12

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

empresas do setor da construção civil listadas no BM&FBOVESPA, utilizando como fonte o *ranking* de 2014 e 2015 (ITC,2016), conforme apresentada no Quadro 1.

Quadro 1 – Empresas selecionadas

Cód	Empresa	Cód	Empresa
E1	Brookfield Incorporações	E7	Ez Tec Empreendimentos e Participações S.A
E2	Construtora Adolpho Lindenberg S.A.	E8	Gafisa S.A.
E3	CR2 Empreendimentos Imobiliários S.A.	E9	MRV Engenharia e Participações S.A.
E4	Cyrela Brazil Realty S.A.	E10	Rodobens Negócios Imobiliários S.A.
E5	Direcional Engenharia	E11	Rossi Residencial S.A
E6	Even Construtora e Incorporadora S.A.	E12	Tecnisa S.A.

Fonte: Autores (2017)

A coleta de dados para análise foi realizada no site da CVM (Comissão de Valores Mobiliários) onde foram obtidos os relatórios financeiros do exercício de 2015 das empresas selecionadas e calculados os índices dos 17 indicadores utilizados nos critérios e subcritérios da modelagem AHP, a saber, Indicador de Liquidez (Geral, Corrente e Seca), Indicadores de Rentabilidade (Margem Bruta, Margem Líquida, ROA e ROE), Indicadores de Endividamento (Grau de Endividamento, Composição de Endividamento, Imobilização do PL e Imobilização dos Recursos não Correntes), indicadores de atividade (GAT, PMRE, PMRV, PMPC), e Indicadores de Valor de Mercado (Lucro por ação e Preço por lucro).

Foi aplicado um *survey* com 10 especialistas: engenheiros do setor de gestão de construção civil, e professores universitários (área financeira) de uma IES Federal. Os especialistas atribuíram notas de 1 a 5 de acordo com o grau de importância dada aos 5 critérios utilizados na avaliação do desempenho econômico-financeiro. Com base nestas notas foi possível estabelecer a matriz de comparação paritária entre os cinco critérios.

Para a construção da matriz de comparação paritária de posse das notas e dos índices calculados, utilizou-se a seguinte função que transforma a nota ou índice obtido em valores na escala de Saaty:

```
fnotamag = function (x) { #transforma o vetor notas dos especialistas
/índices (x)  em valores no intervalo [1,9]

  if (sum(x<=0) >0) {     x=x-min(x)+1}

      if (max(x)/min(x)<=9.5) {notamag=x/min(x)} else{

        notamag=((8*x-8*min(x))/(max(x)-min(x)))+1     }

return(notamag) }
```

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Para a obtenção dos índices de consistência de cada matriz de comparação, utilizou-se a seguinte função:

```
fconsistencia = function (A){  #testa a consistência da matriz de co
mparação A

lambda=as.numeric(eigen(A)$values[1]) #aqui pode usar tb lambda=Re(e
igen(A)$values[1]) para obter a parte real do autovalor

IC=(lambda-ncol(A))/(ncol(A)-1)

IR=c(1,1,0.58,0.9,1.12,1.24,1.32,1.41,1.45,1.49,1.51,1.48,1.56,1.57,
1.59)#dados tabelados

RC=IC/IR[ncol(A)]

require(knitr)

return(kable(data.frame(lambda,IC, RC)))}
```

Resultados e Discussão:

Neste estudo foram utilizados os indicadores mencionados acima para comporem os 17 subcritérios dos 5 critérios utilizados nesta pesquisa: Liquidez, Rentabilidade, Atividade, Endividamento e Valor de Mercado. Esses indicadores foram selecionados por serem considerados relevantes em termos econômico e financeiros para demonstrar a situação da empresa e auxiliar o processo de tomada de decisão da escolha da empresa que obteve melhor desempenho. Dessa forma, esquematizou-se a estrutura hierárquica do problema, conforme apresentado na Figura 1.

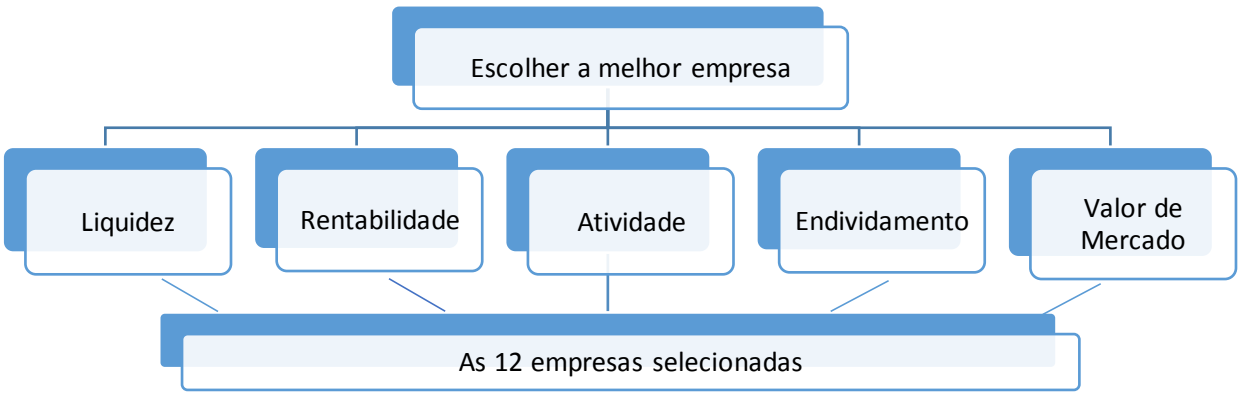


Figura 1 – Estrutura Hierárquica. Fonte: Autores (2017)

A matriz paritária do primeiro nível obtida através da função `fnotamag` para comparação dos 5 critérios pode ser vista na Figura 2, onde observamos que segundo

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

os especialistas, o critério Rentabilidade tem o dobro da importância do critério Atividade e consequentemente Atividade recebe nota 0.5 quando comparada com a Rentabilidade. Os demais critérios apresentam o mesmo grau de importância indicado pela nota 1.

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1  1.0    1    1    1
## [2,]    1  1.0    2    1    1
## [3,]    1  0.5    1    1    1
## [4,]    1  1.0    1    1    1
## [5,]    1  1.0    1    1    1
```

Figura 2 – Matriz paritária do grupo de indicadores. Fonte: Autores (2017)

Após a construção da matriz de comparação realiza-se o teste de consistência. O teste foi realizado com a função `fconsistencia` que calculado o maior autovalor da matriz de comparação, λ_{max} ; o índice de consistência, *IC* e a taxa de consistência, *CR*. Conforme (Shimizu,2006) “com uma taxa de consistência de 0,10 ou menos é considerada aceitável”.

$$IC = \frac{(\lambda\ max - n)}{n-1} \quad (1)$$

A matriz de comparação dos 5 critérios do primeiro nível apresentou consistência aceitável, conforme tabela 1.

Tabela 1 – Verificação da consistência da matriz

lambda	IC	CR
5.058618	0.0146544	0.0130843

Fonte: Autores (2017)

A Tabela 2 apresenta o Vetor dos Pesos (que representa o grau de importância ou ordenamento) de cada critério.

Tabela 2 – Pesos dos critérios do primeiro nível

Grupo de Indicadores	Vetor peso
Liquidez	0.1976825
Rentabilidade	0.2322234
Atividade	0.1747292
Endividamento	0.1976825
Valor de mercado	0.1976825

Fonte: Autores (2017)

O próximo passo é a avaliação dos critérios (indicadores específicos) em relação a cada subcritério (grupo de indicadores do segundo nível).

Da mesma forma, os procedimentos foram realizados com cada critério e subcritério utilizando-se dos índices obtidos dos relatórios financeiros de 2015 e das funções descritas anteriormente, gerando assim os vetores dos pesos de cada subcritério e de cada empresa correspondente a cada subcritério detalhado a seguir.

Para os indicadores específicos, os pesos foram atribuídos com igual importância, conforme apresentado na Tabela 3.

Tabela 3 – Peso dos grupos e indicadores específicos

Liquidez (0.1976825)		Rentabilidade (0.2322234)		Endividamento (0.1976825)		Atividade (0.1747292)		Valor de Mercado (0.1976825)	
Geral	0,3333	Margem Bruta	0,25	Grau End	0,25	GAT	0,25	Lucro por ação	0,50
Corrente	0,3333	Margem Líq	0,25	Comp End	0,25	PMRE	0,25		
Seca	0,3333	ROE	0,25	Imob PL	0,25	PMRV	0,25	Preço/Lucro	0,50
		ROA	0,25	Imob Rec	0,25	PMPC	0,25		

Fonte: Autores (2017)

O próximo passo, foi estabelecer a posição individual de cada empresa em cada indicador específico com o objetivo de hierarquizar as empresas.

A matriz de comparação dos 17 subcritérios do segundo nível apresentou consistência aceitável, conforme Tabela 4.

Tabela 4 – Verificação da consistência da matriz de Liquidez

Indicador específico	**lambda**	**IC**	**CR**
Liquidez Geral	12.10228	0.0093349	0.0063073
Liquidez Corrente	12.14874	0.0135219	0.0091364
Liquidez Seca	12.1582	0.014382	0.0097176
Margem Bruta	12.15979	0.0145266	0.0098152
Margem Líquida	12.31555	0.0286865	0.0193828
ROE (Retorno sobre o Capital Próprio)	12.55257	0.0502336	0.0339416
ROA (Retorno sobre o Ativo Total)	12.3614	0.0328549	0.0221993
GAT (Giro do Ativo Total)	12.15681	0.0142555	0.0096321
PMRE (Prazo Médio de Rotação de Estoques)	12.17724	0.016113	0.0108872
PMRV (Prazo Médio de Recebimento de Vendas)	12.16242	0.0147653	0.0099766
PMPC (Prazo Médio de Pagamento de Compras)	12.13777	0.0125248	0.0084627
Grau de Endividamento	12.0497	0.0043548	0.0029425
Composição de Endividamento	12.16443	0.0149481	0.0101001
Imobilização do Patrimônio Líquido	12.20053	0.0182296	0.0123173
Imobilização de Recursos não correntes	12.10085	0.0091684	0.0061949
Preço por lucro	12.13627	0.0123878	0.0083702
Lucro por ação	12.73402	0.066729	0.0450872

Fonte: Autores (2017)

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Assim, com a consistência aceitável, os pesos obtidos para cada empresa e para cada indicador de liquidez é apresentado na tabela 5.

Tabela 5 – Vetor dos pesos dos Indicadores de Liquidez para as empresas

Empresa	Liquidez Geral	Liquidez Corrente	Liquidez Seca
E1	0.0517845	0.0591332	0.0375035
E2	0.0574427	0.0625228	0.1008749
E3	0.2413525	0.1828740	0.1805039
E4	0.0657903	0.0774927	0.0736396
E5	0.0574427	0.0997098	0.1095241
E6	0.0574427	0.0823601	0.0977903
E7	0.1856613	0.1657859	0.1270122
E8	0.0574427	0.0625228	0.0500466
E9	0.0574427	0.0647972	0.0589736
E10	0.0615115	0.0689009	0.0977903
E11	0.0492432	0.0209342	0.0315187
E12	0.0574427	0.0529663	0.0348224

Fonte: Autores (2017)

Os resultados dos Indicadores de Liquidez demonstram que a empresa E3 foi a que possui a melhor liquidez no ano de 2015. Do mesmo modo destaca-se a empresa E7 com desempenho semelhante à empresa E3, porém ligeiramente menor. A E11 se destaca aqui como o pior desempenho em todos os indicadores de liquidez.

Já para os Indicadores de Rentabilidade que apresenta o maior peso dentre os 5 critérios do primeiro nível, as empresas E2 e E7 possuem melhor Margem Bruta. Contudo, E2 se diferencia como melhor Margem Líquida, ROA (retorno sobre Ativos) e ROE (retorno sobre Patrimônio Líquido) quase empatada com E1 que se saiu ligeiramente melhor neste quesito, conforme apresentado na Tabela 6.

Tabela 6 - Vetor dos pesos dos Indicadores de Rentabilidade para as empresas

Empresa	Margem Bruta	Margem Líquida.	ROE	ROA
E1	0.0126508	0.0507350	0.2679721	0.1671716
E2	0.1455669	0.4430104	0.2537501	0.2806478
E3	0.0883227	0.0507350	0.0194936	0.0150933
E4	0.1002939	0.0537537	0.0504620	0.0551657
E5	0.0707123	0.0537537	0.0499537	0.0501944
E6	0.0784680	0.0507350	0.0435135	0.0408389
E7	0.1455669	0.0743516	0.0850560	0.1514614
E8	0.0784680	0.0507350	0.0324064	0.0300319
E9	0.0825074	0.0537537	0.0662253	0.0594094
E10	0.0825074	0.0507350	0.0315271	0.0300319
E11	0.0364678	0.0139482	0.0128052	0.0280099
E12	0.0784680	0.0537537	0.0868350	0.0919437

Fonte: Autores (2017)

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Para os Indicadores de Atividade que apresenta o menor peso dentre os 5 critérios do primeiro nível, a E2 possui o melhor desempenho para o indicador específico GAT, para PMPC a empresa E1 se destaca. Para o PMRE a empresa E10 apresenta melhor desempenho e a empresa E11 se destaca em PMRV.

Tabela 7 - Vetor dos pesos dos Indicadores de Atividade para as empresas

Empresa	GAT	PMRE	PMRV	PMPC
E1	0.04648687	0.04884272	0.03005815	0.39594452
E2	0.19441951	0.05474835	0.16350554	0.06675085
E3	0.02588281	0.02539491	0.01773641	0.08185393
E4	0.08234419	0.07554206	0.07057551	0.07012272
E5	0.09457300	0.15156910	0.06438958	0.07089941
E6	0.09457300	0.12706979	0.06293128	0.03853970
E7	0.06205660	0.03737104	0.07475941	0.03853970
E8	0.08234419	0.07153108	0.07723080	0.03635584
E9	0.09457300	0.09040989	0.10177267	0.05301452
E10	0.07967875	0.15558008	0.04798496	0.07089941
E11	0.06072388	0.09040989	0.21036659	0.03853970
E12	0.08234419	0.07153108	0.07868910	0.03853970

Fonte: Autores (2017)

Os indicadores específicos do grupo de Endividamento, E3 e E7 apresentam melhor desempenho em Grau de Endividamento e Imobilização do Patrimônio Líquido. Porém, para Composição do Endividamento e Imobilização de Recursos não Correntes, as empresas E5 e E6 são mais eficientes, respectivamente. A E11 se destaca aqui como o pior desempenho em todos os indicadores de endividamento.

Tabela 8 - Vetor dos pesos dos Indicadores de Endividamento para as empresas

Empresa	Grau	Composição	Imob Pl	Imob Rec
E1	0.0286264	0.0817708	0.0536036	0.0877131
E2	0.0645601	0.0751773	0.0760420	0.0809076
E3	0.2215774	0.0929148	0.1332537	0.0940779
E4	0.0669362	0.0870487	0.0924803	0.0877131
E5	0.0614949	0.1284043	0.0849783	0.0877131
E6	0.0614949	0.1015158	0.0983825	0.1033012
E7	0.2215774	0.0817708	0.1108478	0.0877131
E8	0.0614949	0.0817708	0.0849783	0.0877131
E9	0.0614949	0.0817708	0.0760420	0.0877131
E10	0.0669362	0.0870487	0.0924803	0.0877131
E11	0.0223121	0.0294489	0.0249014	0.0307015
E12	0.0614949	0.0713582	0.0720101	0.0770203

Fonte: Autores (2017)

Finalizando com os Indicadores de Valor de Mercado, a E10 apresenta a melhor avaliação do mercado (Indicador Preço por Lucro) e a E2 se destaca por possuir maior Lucro por Ação.

Tabela 9- Vetor dos pesos dos Indicadores de Valor de Mercado para as empresas

Empresa	Preço por lucro	Lucro por ação
E1	0.0304103	0.0299439
E2	0.0817572	0.3426364
E3	0.0304103	0.0307093
E4	0.0961891	0.0810068
E5	0.0612219	0.0723119
E6	0.0857574	0.0463688
E7	0.0653796	0.0584817
E8	0.0504587	0.0440726
E9	0.0961891	0.0810068
E10	0.3402831	0.0440726
E11	0.0304103	0.0101392
E12	0.0315331	0.1592501

Fonte: Autores (2017)

Para obter o *ranking* final das empresas é necessário realizar a multiplicação do peso de cada grupo de indicador, por cada peso do indicador específico e obter a soma dos produtos. Realizado este processo obtém-se a ordenação das empresas. A Tabela 10 apresenta o *ranking* final das empresas de acordo com desempenho em relação aos indicadores em 2015.

Tabela 10 - Ranking das empresas em relação dos indicadores de 2015

Posição	Empresa	Peso	Posição	Empresa	Peso
1	E2	0.1572998	7	E4	0.0764420
2	E7	0.1043661	8	E9	0.0746748
3	E10	0.0963219	9	E6	0.0732514
4	E3	0.0893380	10	E12	0.0722558
5	E1	0.0799010	11	E8	0.0589704
6	E5	0.0783925	12	E11	0.0387863

Fonte: Autores (2017)

Verifica-se de acordo com a Tabela 10, que a empresa E2 apresenta uma melhor posição dentre as demais. As empresas E7 e E10 ocupam a segunda e terceira posição respectivamente.

Estes resultados corroboram com (ASSAF NETO, 2011), que esclarecem que a empresa pode possuir boa capacidade de liquidar suas obrigações, contudo não necessariamente apresentará o mesmo potencial de rentabilidade. Como também pode apresentar melhor desempenho no Indicador Preço por lucro (quanto que os investidores estão dispostos a pagar por cada R$ 1,00 de lucro que a empresa tiver) e ao mesmo tempo possuir alto nível de endividamento (empresa pode ter se endividado mais do que deveria).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Conclusão:

A avaliação de desempenho organizacional é importante para verificar se os objetivos estabelecidos pela empresa estão sendo alcançados, ajudando ainda na melhor aplicação dos recursos. Para mensuração do desempenho, procura-se medir as consequências financeiras e econômicas das decisões de gestões passadas.

Este estudo objetivou definir por meio dos indicadores econômico-financeiros, um *ranking* das melhores empresas do setor de construção civil listadas na BM&FBovespa, com a utilização do método AHP. Para atender ao objetivo, foi utilizada uma metodologia descritiva, realizada por meio de análise documental, com abordagem quantitativa utilizando a técnica AHP implementada no *software* R, que mostrou ser de grande versatilidade na implementação do método. Algumas funções foram utilizadas do pacote básico do R como é o caso da função eigen() para obter os autovalores e autovetores da matriz de comparação; outras funções foram implementadas na própria linguagem como é o caso da função fnotamag() e fconsistencia() criada pelos autores deste trabalho.

Com o *ranking* final, obteve-se a classificação das empresas Construtora Adolpho Lindenberg SA., Ez Tec Empreendimentos e Participações SA. e Rodobens Negócios Imobiliários SA. como as empresas com o melhor desempenho no setor da construção civil. Este resultado indica que são as empresas que possuem um conjunto de indicadores que demonstram um melhor desempenho econômico-financeiro em relação as demais empresas.

Os resultados apurados demonstram que de fato a Construtora Adolpho Lindenberg SA (E2) apresenta o melhor desempenho econômico financeiro no ano de 2015. A empresa possui mais de 60 anos de atuação, entregou cerca de 500 empreendimentos a mais de 7.000 clientes em todo o Brasil.

Apesar da crise no setor, a Construtora Adolpho Lindenberg SA, ao final do ano de 2015, totalizou 312,1 mil m² em obras, atingiu R$ 45,5 milhões de Receita Líquida (um aumento de 1,1% em relação ao período de 2014), totalizou R$ 14,6 milhões em Lucro Líquido (um aumento de 74,9% quando comparado a 2014), ainda obteve um aumento de 70,6% no Patrimônio Líquido totalizando R$ 26,9 milhões no encerramento do período.

O resultado também converge com o Ranking do ITC, que em 2015 as três estão entre as 7 maiores empresas, dentre as listadas na BM&FBovespa.

Com o resultado, pode-se comprovar o que especialistas defendem na teoria, verificou-se que a empresa pode possuir bom desempenho em um indicador, contudo não necessariamente apresentará o mesmo potencial em outro indicador.

Referências:

ASSAF NETO, Alexandre; LIMA, Fabiano Guasti. **Curso de Administração financeira**. 2ª. ed. São Paulo: Atlas, 2011.

ITC, **Ranking ITC: As 100 maiores construtoras**. URL http://rankingitc.com.br/, 2016.

IUDÍCIBUS, Sérgio de. **Análise de Balanços**. São Paulo: Atlas, 2010.

R CORE TEAM. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2016.

SAATY, T. L. **Método** de **análise hierárquica**. São Paulo: McGraw-Hill; Makron, 1991.

SHIMIZU, T. **Decisão nas organizações.** São Paulo: Atlas, 2006.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 26 - RISCO SISTÊMICO NA REDE BANCÁRIA BRASILEIRA: UMA ABORDAGEM COM VINE-CÓPULA

Andrea Ugolini [20]
Miguel A. Rivera-Castro [21]

Resumo

Neste estudo, examinamos o risco sistêmico do sistema bancário brasileiro, utilizando o valor em risco condicional como medida de risco. Modelamos a dependência condicional multivariada entre bancos brasileiros por meio do uso de uma estrutura hierárquica de dependência em árvore dada por um modelo Vine-cópula. Os resultados indicam que, para o período de janeiro de 2007 a janeiro de 2016, o risco sistêmico do sistema financeiro brasileiro aumentou dramaticamente durante a crise financeira global. Nossa evidência também indica que o principal impacto sistêmico do setor bancário sobre o conjunto do sistema financeiro brasileiro tem-se originado desde o Bradesco e o Itaú. Enquanto que, os outros bancos desempenharam um papel de transmissão de risco menor, em destaque o Banco do Brasil que fornece menos risco para o sistema. Os resultados têm implicações para a regulamentação de capital das instituições financeiras e para as decisões dos investidores de gestão de risco.

Palavras-Chave: Risco Sistémico, Sistema Bancário Brasileiro, CoVaR, Cópula, Vine-cópula

Abstract

In this study, the tail systemic risk of the Brazilian banking system is examined, using the conditional quantile as the risk measure. Multivariate conditional dependence between Brazilian banks is modelled with a Vine-copula hierarchical structure. The results demonstrate that, during the period from January 2007 to January 2016, Brazilian financial systemic risk increased drastically in the course of the global financial crisis period. Our empirical findings show that Bradesco and Itaú are the origin of the larger systemic shocks from the banking system to the financial system network. While the other banks played a minor risk transmission role, in particular Banco do Brasil which provides less risk to the system. The results have implications for the capital regulation of financial institutions and for risk managers' decisions.

Keywords: Systemic Risk, Brazilian Banking System, CoVaR, Copula, Vine-copula

Introdução

As crises financeiras recentes têm atraído a atenção de investidores e reguladores, devido à fragilidade do sistema financeiro e do potencial risco sistêmico decorrente de um não cumprimento dos compromissos de pagamentos dos bancos. Desde o colapso do Lehman Brothers, em meados de setembro de 2008, tornou-se crucial quantificar o risco sistêmico para os tomadores de decisões políticos, dado que

[20] Post graduate programme in management - PPGA, Unifacs, Rua Dr.José Peroba 251, 41770-235 , Salvador, Brazil. E-mail: andreaugolini@me.com.
[21] Post graduate programme in management - PPGA, Unifacs, Rua Dr.José Peroba 251, 41770-235 , Salvador, Brazil. E-mail: miguel.castro@pro.unifacs.br.

o regulamento macro-prudencial requer uma avaliação de como uma frágil posição de uma instituição financeira pode comprometer o desempenho de outras instituições financeiras. É também crucial para determinar quanto capital regulamentar uma instituição financeira precisa acumular, a fim de cobrir este tipo de risco (Das e Uppal (2004); Rosemberg e Schuermann, 2006). O Conselho de Estabilidade Financeira (2010) afirmou que as instituições financeiras sistemicamente importantes devem ser obrigadas a assumir uma maior capacidade de absorção de perdas, uma vez que contribuem mais para o risco global do sistema financeiro. Neste artigo, vamos abordar a questão do risco sistêmico, quantificando a contribuição de risco de cada banco para os outros bancos. Nós descrevemos um modelo hierárquico de dependência com cópula para os bancos que levam em consideração a interconexão e dependência condicional e também as características específicas de dependência caudal entre os bancos. A partir deste modelo, podemos avaliar como o perigo de uma instituição em particular aumenta o valor em risco (*VaR*) de uma outra instituição, levando-se em conta os efeitos diretos e indiretos.

A medida de risco mais amplamente utilizada é o *VaR,* que quantifica a máxima perda de uma instituição financeira para um dado nível de confiança e horizonte temporal. Esta medida é, no entanto, centrada no risco individual de uma instituição e deixa de considerar os potenciais efeitos colaterais que um não cumprimento pode ter sobre outras instituições. Portanto, a literatura acadêmica sobre as políticas macro-prudenciais tem centrado a própria atenção sobre a contabilização da contribuição de cada banco ao risco de outras instituições e/ou para todo o sistema financeiro, desenvolvendo uma gama de medidas contra riscos sistêmicos (Bisias et al. 2012; Bernal et al. 2014). Huang et al. (2009) desenvolveram um indicador de risco sistêmico para a dificuldade financeira sistêmica dado pelo preço de *credit default swaps* (CDS). Usando dados CDS, Segoviano e Goodhart (2009) construíram um índice de estabilidade bancária para avaliarem a dependência interbancária de eventos extremos. Rodríguez-Moreno e Peña (2013) forneceram evidências sobre a adequação do uso de dados CDS para estimarem o risco sistemático. Acharya et al. (2010) introduziram a esperada diminuição (Expected Shortfall) sistêmica e a esperada diminuição marginal como indicadores para quantificarem o risco de a situação se agravar e, ainda, as contribuições das instituições financeiras ao risco. Brownlees e Engle (2012) desenvolveram uma medida de risco sistêmico chamado SRISK, representando-se o montante de capital necessário para restaurar uma exigência de capital mínimo. Allen et al. (2012) propuseram uma medida do risco sistêmico chamado CATFIN, que pode predizer o declínio das atividades de empréstimos agregados no banco com 6 meses de antecedência. Billio et al. (2012) apresentaram cinco medidas de risco sistêmico que capturam o contágio e o efeito da exposição nas relações entre instituições financeiras. Engle e Manganelli (2004) desenvolveram um modelo de valor em risco autoregressivo condicional (*CaViaR*) que usa regressão quantílica para capturar o comportamento dos retornos nas caudas. Recentemente, Adrian e Brunnermeier (2011) e Girardi e Ergün (2013) propuseram o *VaR* condicional (*CoVaR*) como uma nova medida de risco sistêmico. Esta medida considera a perda máxima esperada de uma instituição financeira para um nível de confiança e condicionada ao facto de que outra instituição está numa situação de instabilidade medida polo seu *VaR*.

Objetivo

A partir desse panorama, o nosso trabalho contribui para a literatura atual sobre a medição do risco sistêmico no sistema financeiro. Em primeiro lugar, modelamos a

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

estrutura de dependência entre os bancos usando uma estrutura de árvore hierárquica multivariada, chamada de Vine-cópula (Joe, 1996), a partir da qual podemos calcular o *CoVaR*. Esta estrutura de dependência hierárquica permite que o risco que representa um banco para outros bancos seja avaliado, levando-se em conta as relações diretas e indiretas entre os bancos. Além disso, dado que a estrutura hierárquica pode ser decomposta em um conjunto de cópulas bivariadas (chamadas par-cópulas) que capturam a dependência entre duas variáveis nas quais se usam diferentes especificações de cópula, podemos levar em conta as diferentes características de dependência, como as dependências média, caudal simétrica ou assimétrica. Nesse sentido, a Vine-cópula permite que as distribuições marginais e a estrutura de dependência multivariada possam ser modeladas separadamente, independente de na estrutura de dependência ser possível capturar a dinâmica e as assimetrias de volatilidade específicas nas séries univariadas dos retornos do banco. Estas duas últimas características são cruciais para a obtenção de uma estimativa precisa do *CoVaR*.

Com base nesta descrição teórica, este trabalho tem como objetivo medir o impacto sistêmico de crise financeira em um banco brasileiro cotado com outros bancos cotados no sistema financeiro brasileiro, usando dados diários para o período de 01 de janeiro de 2007 a 18 janeiro de 2016, para o índice financeiro brasileiro e para os sete bancos brasileiros cotados, ou seja, Banco ABC, Banco do Brasil, Banco Bradesco, Banco PanAmericano, Banco do Estado do Rio Grande do Sul, Banco Itaú e Paraná Banco.

Material e Métodos:

Várias medidas de risco sistêmico têm sido propostas na literatura, para se quantificar o impacto de uma instituição financeira potencialmente geradora de risco sobre o sistema financeiro em seu conjunto ou sobre outras instituições financeiras (Adrian e Brunnemeier, 2011, Girardi e Ergün, 2013). Para a nossa pesquisa, optamos por utilizar o *CoVaR* para quantificarmos o risco sistêmico como o efeito da situação de risco de determinada instituição financeira sobre o *VaR* do sistema financeiro ou de determinada instituição financeira.

Definição do *CoVaR*

O *CoVaR* de uma instituição financeira é o quantil condicionado pelo fato de que outra instituição financeira está em uma situação de crise, ou em um quantil extremamente baixo. Sendo X_t^1 o retorno do banco 1, e X_t^2 o retorno do banco 2. O *CoVaR* do banco 1 ao tempo *t* para um dado α-quantil, e β-quantil do banco 2 pode ser definido formalmente como $\Pr(X_t^1 \le \mathrm{CoVaR}_{\alpha,\beta,t}^{1|2} \mid X_t^2 \le \mathrm{VaR}_{\beta,t}^2) = \alpha$, e pode ser calculado por:

$$\mathrm{CoVaR}_{\alpha,\beta,t}^{1|2} = F_{X_t^1 | X_t^2 \le \mathrm{VaR}_{\beta,t}^2}^{-1}(\alpha). \qquad (1)$$

de modo que $F_{X_t^1 | X_t^2 \le \mathrm{VaR}_{\beta,t}^2}^{-1}(\alpha)$ é a inversa da função de distribuição de X_t^1 condicionada sobre o feito que $\Pr(X_t^2 \le VaR_{b,t}^2) = b$. Isso é um quantil incondicional, que em finanças é denominado de Valor em Risco. Com esta nossa metodologia, substancialmente, determinamos o quantil de uma distribuição condicional que requer informações sobre a dependência conjunta bivariada entre X_t^1 e X_t^2.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Um passo adicional na metodologia é exigir informações relativas à dependência multivariada entre todas as instituições financeiras no sistema financeiro. Em outras palavras, calcular os efeitos secundários sobre X_t^1 que podem surgir como resultado da dependência com o banco 2 e outros bancos. Nesse sentido, considerando a estrutura de dependência multivariada, assim como *n* bancos no sistema financeiro, podemos redefinir o quantil condicional como $\Pr(X_t^1 \le \mathrm{CoVaR}_{\alpha,\beta,t}^{1|2} \mid X_t^2 \le \mathrm{VaR}_{\beta,t}^2, X_t^3, ..., X_t^n) = \alpha$,que pode ser calculado por:

$$\mathrm{CoVaR}_{\alpha,\beta,t}^{1|2} = F^{-1}_{X_t^1 | X_t^2 \le \mathrm{VaR}_{\beta,t}^2, X_t^3, \cdots, X_t^4}(\alpha)\,. \qquad (2)$$

Assim, computando-se o *CoVaR* na Equação (2), consiste em se determinar o quantil da distribuição condicional, no entanto, levam-se em conta as circunstâncias financeiras e das dependências dos outros bancos com bancos 1 (Reboredo e Ugolini, 2015).

Estimação do VaR e do CoVaR com cópula e Vine-cópula.

A partir da informação sobre a média e a variância de X_t , podemos calcular o *VaR* da distribuição do retorno como:

$$\mathrm{VaR}_{\alpha,t}^2 = \mu_t + F_\nu^{-1}(\beta)\sigma^t\,, \qquad (3)$$

sendo que $F_\nu^{-1}(\beta)$ denota o β-quantil incondicional de uma distribuição t-Student. Para calcularmos o *CoVaR* da distribuição dos retornos, usamos funções cópula (Joe, 1997 e Nelsen, 2006). Observe-se que $\Pr(X_t^1 \le \mathrm{CoVaR}_{\alpha,\beta,t}^{1|2} \mid X_t^2 \le \mathrm{VaR}_{\beta,t}^2) = \alpha$ e $\Pr(X_t^1 \le \mathrm{CoVaR}_{\alpha,\beta,t}^{1|2} \mid X_t^2 \le \mathrm{VaR}_{\beta,t}^2, X_t^3, \ldots, X_t^n) = \alpha$ pode ser escrita, respectivamente, como:

$$\frac{F_{X_t^1, X_t^2}(\mathrm{CoVaR}_{\alpha,\beta,t}^{1|2}, \mathrm{VaR}_{\beta,t}^2)}{F_{X_t^2}(\mathrm{VaR}_{\beta,t}^2)} = \alpha\,, \qquad (4)$$

$$\frac{F_{X_t^1, X_t^2 | X_t^3, \ldots, X_t^n}(\mathrm{CoVaR}_{\alpha,\beta,t}^{1|2}, \mathrm{VaR}_{\beta,t}^2)}{F_{X_t^2 | X_t^3, \ldots, X_t^n}(\mathrm{VaR}_{\beta,t}^2)} = \alpha\,. \qquad (5)$$

Assim, os quantiles condicionais para a distribuição dos retornos exigem informações sobre a função de distribuição conjunta de X_t^1 e X_t^2, $F_{X_t^1, X_t^2}(\cdot)$. Leve-se em conta que o teorema de Sklar (1959) nos permite expressar a função distribuição conjunta em termos de uma função cópula C, $C(F_X(x), F_Y(y)) = F_{XY}(x,y)$ onde, as Equações (4) e (5) podem ser escritas como:

$$C_{X_t^1, X_t^2}\left(F_{X_t^1}\left(CoVaR_{a,b,t}^{1|2}\right), F_{X_t^2}\left(VaR_{b,t}^2\right)\right) = a\,b\,, \qquad (6)$$

$$C_{X_t^1, X_t^2 | X_t^3 \cdots X_t^n}\left(F_{X_t^1, X_t^2 | X_t^3 \cdots X_t^n}\left(CoVaR_{a,b,t}^{1|2}\right), F_{X_t^2 | X_t^3 \cdots X_t^n}\left(VaR_{b,t}^2\right)\right) = a\,b\,. \qquad (7)$$

Assim, podemos caracterizar os valores do *CoVaR* em termos de uma função cópula bivariada ou multivariada. Com a primeira caracterização, Equação (6),

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

podemos calcular o $F_{X_t^1}\left(CoVaR_{a,b,t}^{1|2}\right)$ invertendo-se a função cópula, dados os valores de a e de $F_{X_t^2}\left(VaR_{b,t}^{2}\right)=b$, que denotamos como $\hat{F}_{X_t^1}\left(CoVaR_{a,b,t}^{1|2}\right)$.

Subsequentemente, invertendo-se a função de distribuição marginal de X_t^1 obteremos o valor do *CoVaR* como:

$$CoVaR_{a,b,t}^{1|2}=F_{X_t^1}^{-1}\left(\hat{F}_{X_t^1}\left(CoVaR_{a,b,t}^{1|2}\right)\right). \quad (8)$$

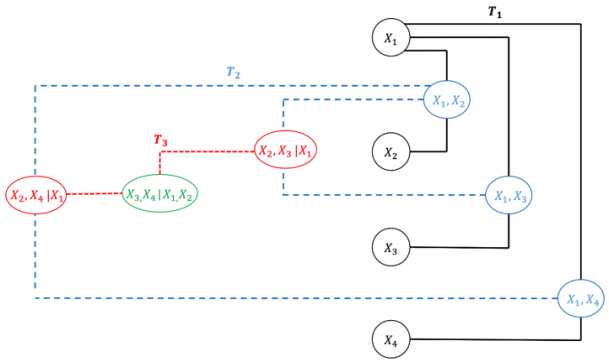
Da mesma forma, usando-se funções cópula em uma configuração multivariada a partir da Equação (7), é mais atraente, pois se permite a transmissão direta da dificuldade financeira do banco de 2 ao banco 1, mas também se explicam efeitos indiretos através do impacto que uma dificuldade de banco 2 pode ter sobre outros bancos. No entanto, exige-se a dependência de modelagem multivariada em *n* dimensões. Para esse fim, foram consideradas as Vine-cópulas que permitem decompor a densidade multivariada no produto das densidades marginais e uma cópula multivariada que é gerada através de uma construção hierárquica que é decomposta em uma cascata de cópulas bivariadas chamadas de Pair-cópula. Foi considerado um tipo de Vine-cópula a C-Vine que tem densidade multivariada:

$$f\left(x_1,x_2,\ldots,x_n\right)=\prod_{k=1}^{n}f_k\left(x_k\right)\prod_{h=2}^{n}c_{1,h}\left(F_1\left(x_1\right),F_h\left(x_h\right)\right)$$
$$\prod_{j=2}^{n-1}\prod_{i=1}^{n-j}c_{j,j+1|1,\ldots,j-1}\left(F\left(x_j\mid x_1,\ldots,x_{j-1}\right),F\left(x_{j+1}\mid x_1,\ldots,x_{j-1}\right)\right), \quad (9)$$

onde $c_{j,j+1|1,\ldots,j-1}$ é a cópula condicional, e em que a função de distribuição condicional da variável x_i, dada a variável x_j, é dada pela (Joe (1997)):

$$F_{i|j}(x_i\mid x_j)=\frac{\partial C_{ij}\left(F_i(x_i),F_j(x_j)\right)}{\partial F_j(x_j)}. \quad (10)$$

Figura 1: Estrutura C-Vine.



A Figura 1 representa a cópula C-Vine por meio de uma estrutura hierárquica, de tal modo que, em um primeiro nível da árvore, *n* nódulos são conectados por arestas que representam a dependência entre duas variáveis, e, em níveis de árvores sucessivas, os nós são obtidos a partir do conjunto de arestas do nível da árvore anterior. Cada árvore (T) tem uma estrutura de estrelas, e apenas uma variável desempenha um papel fundamental. A dependência é medida entre a variável central

com as restantes variáveis na primeira árvore, utilizando-se cópulas bivariadas, como indicado pelo segundo termo da Equação (9), ou usando-se cópulas bivariadas condicionais nas árvores restantes, como indicado pelo terceiro termo na Equação (9). Uma vez que a dependência em cada árvore é modelada, a árvore é sucessivamente expandida, de modo que os nós das respectivas árvores são configurados pelas bordas das árvores anteriores, tal como representado na Figura 1. Em cada árvore, a variável central que regula a dependência é identificada como aquela que maximiza a soma de dependências emparelhadas medidas por tau de Kendall.

Computando-se o CoVaR por meio das funções cópula, temos várias vantagens. Em primeiro lugar, as cópulas oferecem flexibilidade e permitem a modelagem separada das marginais e estruturas de dependência. Isto é crucialmente importante quando a dependência quantílica difere, e quando a função de distribuição conjunta não é elíptica, ou quando os dados têm características especiais (tais como heterocedasticidade condicional). Em segundo lugar, a obtenção dos quantis condicionais das cópulas é computacionalmente fácil, já que só se precisa de informações sobre a cópula, sobre a distribuição marginal dos retornos das ações de um banco e sobre a probabilidade cumulativa de quantiles de outro banco.

Em nosso estudo empírico, foram utilizadas diferentes especificações de cópula estáticas, a fim de se capturarem diferentes características de dependência: nenhuma dependência caudal (Gaussian, Plackett e Frank), dependência caudal simétrica (Student-t) e dependência caudal assimétrica (Gumbel, Rotated Gumbel BB1 e BB7).

Resultados e Discussão:

Neste trabalho, examinou-se empiricamente o impacto sistêmico do não cumprimento dos compromissos de cada banco listado na bolsa de Valores de São Paulo para com outros bancos e para o conjunto do sistema financeiro bancário brasileiro (BFIndex), sendo usados dados diários para o período de 01 de janeiro de 2007 a 18 de janeiro de 2016. O conjunto de bancos incluiu sete bancos: Banco ABC (ABC), Banco do Brasil (BB), Banco Bradesco (Bradesco), Banco Pan-Americano (Pan), Banco do Estado do Rio Grande do Sul (Banrisul), Banco Itaú (Itaú) e Paraná Banco (Paraná), para os quais temos informações dos preços para toda a amostra. Também tomamos informações sobre o índice financeiro do Brasil, que captura o comportamento de todo o sistema financeiro. Ao analisarmos o impacto sistêmico de um específico banco brasileiro sobre o sistema financeiro, excluímos este banco do índice financeiro, e assim fomos re-calcular o índice; desse modo, excluíram-se os efeitos diretos das flutuações dos preços deste banco no índice brasileiro (ver, por exemplo, López-Espinosa et al. (2012)). Os dados foram obtidos a partir de Bloomberg, e os retornos foram calculados numa base de composição contínua.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figura 2: Gráfico das séries temporais dos preços das ações dos bancos brasileiros e o índice financeiro do Brasil (BFIndex) da BM&FBOVESPA.

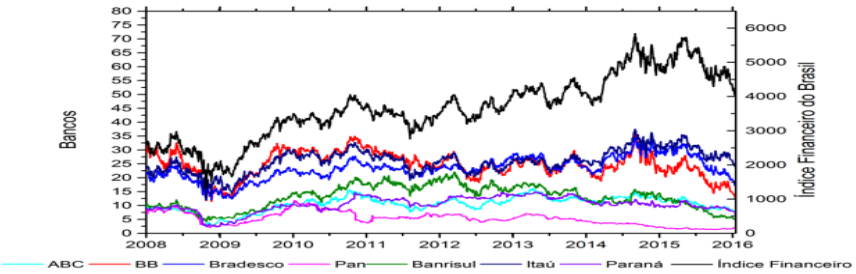


Tabela 1. Estatísticas descritivas para os bancos e o índice financeiro do Brasil.

	ABC	BB	Bradesco	Pan	Banrisul	Itaú	Paraná	Sistema
Mean	0.000	0.000	0.000	-0.001	0.000	0.000	0.000	0.000
Std. Dev	0.026	0.028	0.023	0.030	0.028	0.024	0.024	0.021
Minimum	-0.188	-0.189	-0.122	-0.369	-0.142	-0.129	-0.187	-0.128
Minimum	0.182	0.188	0.200	0.235	0.160	0.210	0.152	0.190
Skewness	-0.029	0.018	0.411	-1.177	0.073	0.500	-0.365	0.461
Kurtosis	8.471	7.676	8.983	29.085	5.777	9.650	11.600	11.018
JB^{i}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LB^{IV}	0.000	0.032	0.000	0.301	0.041	0.000	0.000	0.000
$ARCH\text{-}LM^{IV}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notas: Dados diários para o período do 01 de Janeiro de 2007 a 18 de Janeiro de 2016. A tabela informa as estatísticas básicas para as séries de retornos, incluindo média (mean), desvio padrão (Std. Dev.), assimetria e curtose. JB refere-se às estatísticas empíricas para o teste de Jarque-Bera de normalidade baseado em assimetria e excesso de curtose. LB refere-se às estatísticas empíricas para o teste de Ljung-Box para a correlação serial na série retorno calculado com 20 defasagens. ARCH se refere às estatísticas empíricas do teste estatístico de autorregressivo condicionalmente heterocedástico de ordem 10.

A Figura 2 mostra um gráfico de séries temporais dos preços das ações para os bancos brasileiros e o índice financeiro do Brasil. Na primeira parte pode-se observar uma queda moderada, que é o início da crise financeira global, e também no final da mostra vemos outra queda, devido à crise política no Brasil. A Tabela 1 mostra que os retornos têm características semelhantes; eles não apresentaram qualquer tendência significativa, como também os desvios-padrão foram maiores do que os retornos médios. Todos os bancos apresentaram volatilidade semelhante em termos de desvio padrão, enquanto os retornos do índice financeiro do Brasil têm menos volatilidade. Observamos caudas pesadas que evidenciam o fato de que o coeficiente de Kurtosis fica acima de 3. Entretanto, o teste de Jarque-Bera rejeitou a hipótese nula de normalidade. O estatístico O estatístico autorregressivo condicionalmente heterocedástico-multiplicadores de Lagrange (ARCH-LM) e as estatísticas de Ljung-Box para retornos ao quadrado indicam que todas as séries apresentam efeitos ARCH.

Calculamos os valores VaR e CoVaR no nível de confiança de 95% (a = 0.05, b = 0.05), utilizando as funções de distribuição marginal univariadas e os melhores

ajustes do par-copula estimada da estrutura hierárquica C-Vine e das melhores cópulas bivariadas. A Figura 5 mostra evidência gráfica do mapa da rede das relações e do tamanho do risco sistêmico normalizado, ou seja Vine $CoVaR/VaR$, que representa a medida do risco sistêmico que a instituição individual leva ou recebe desde todos os outros sete bancos ao longo do período de amostragem. Cada uma das arestas representa o risco sistêmico normalizado recebido e cedido. Consideramos somente aresta o qual valor do risco sistêmico normalizado é > 2, ou seja, o dobro do *VaR* instituição analisada. A cor é o tamanho do risco que pode ser muito elevado (aresta continua vermelha, valor do risco sistêmico normalizado > 4), elevado (aresta tracejada azul, 3 < valor do risco < 4) e mediano (aresta tracejada longa verde, 2 < valor do risco < 4). O tamanho da bola que representa cada uma das instituições nos indica o total do risco sistêmico gerado para uma instituição, versus os diferentes bancos. As estatísticas descritivas são apresentadas na Tabela 2 - 3

Tabela 2: Estatísticas descritivas para o *VaR* e o Vine *CoVaR*.

	Estrutura Vine	*VaR*	Vine *CoVaR*
Bradesco		-0.034	
		(-0.01)	
Bradesco|BB|Brasil	(2,3)		-0.103
			(-0.05)
Bradesco|ABC	(1,3)		-0.126
			(-0.06)
Bradesco|Pan	(4,3)		-0.097
			(-0.05)
Bradesco|Itaú	(6,3)		-0.141
			(-0.07)
Bradesco|Banrisul	(5,3)		-0.133
			(-0.06)
Bradesco|Paraná	(7,3)		-0.116
			(-0.05)
BB		-0.042	
		(-0.02)	
BB|Bradesco	(2,3)		-0.103
			(-0.04)
BB|ABC	(1,2|3)		-0.114
			(-0.05)
BB|Pan	(2,4|3)		-0.132
			(-0.06)
BB|Itaú	(2,6|3)		-0.146
			(-0.07)
BB|Banrisul	(2,5|3)		-0.098
			(-0.05)
BB|Paraná	(2,7|3)		-0.137
			(-0.06)
ABC		-0.04	
		(-0.01)	
ABC|Bradesco	(1,3)		-0.121
			(-0.05)
ABC|BB	(1,2|3)		-0.135
			(-0.06)
ABC|Pan	(1,4|2,3)		-0.041
			(-0.02)
ABC|Itaú	(1,6|2,3)		-0.039
			(-0.02)
ABC|Banrisul	(1,5|2,3)		-0.044
			(-0.02)
ABC|Paraná	(1,7|2,3)		-0.128
			(-0.06)
Pan		-0.047	
		(-0.02)	
Pan|Bradesco	(4,3)		-0.094
			(-0.04)
Pan|BB	(2,4|3)		-0.142
			(-0.06)
Pan|ABC	(1,4|2,3)		-0.036
			(-0.02)
Pan|Itaú	(4,6|1,2,3)		-0.062
			(-0.03)
Pan|Banrisul	(4,5|1,2,3)		-0.032
			(-0.02)
Pan|Paraná	(4,7|1,2,3)		-0.036
			(-0.02)

Notas. A tabela mostra com o primeiro o valor é a média a ló largo de todo o período. Valores do desvio padrão (entre parênteses).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Tabela 3: Estatísticas descritivas para o *VaR* e o Vine *CoVaR*.

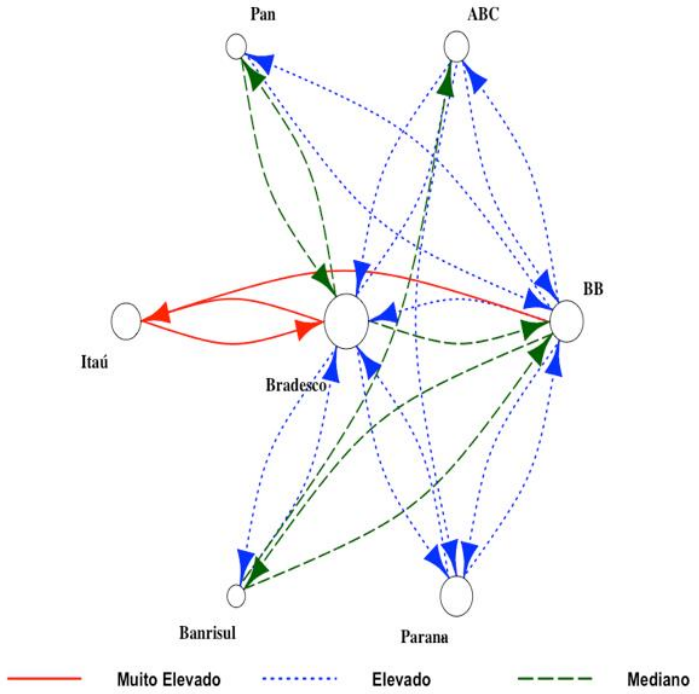
	Estrutura Vine	*VaR*	Vine *CoVaR*
Itaú		-0.036	
		(-0.01)	
Itaú|Bradesco	(6,3)		-0.132
			(-0.06)
Itaú|BB	(2,6|3)		-0.045
			(-0.04)
Itaú|ABC	(1,6|2,3)		-0.04
			(-0.03)
Itaú|Pan	(4,6|1,2,3)		-0.072
			(-0.04)
Itaú|Banrisul	(5,6|1,2,3,4)		-0.04
			(-0.03)
Itaú|Paraná	(6,7|1,2,3,4)		-0.074
			(-0.06)
Banrisul		-0.045	
		(-0.01)	
Banrisul|Bradesco	(5,3)		-0.128
			(-0.05)
Banrisul|BB	(2,5|3)		-0.087
			(-0.04)
Banrisul|ABC	(1,5|2,3)		-0.091
			(-0.04)
Banrisul|Pan	(4,5|1,2,3)		-0.062
			(-0.03)
Banrisul|Itaú	(5,6|1,2,3,4)		-0.067
			(-0.03)
Banrisul|Paraná	(5,7|1,2,3,4,6)		-0.07
			(-0.04)
Paraná		-0.038	
		(-0.02)	
Paraná|Bradesco	(7,3)		-0.112
			(-0.05)
Paraná|BB	(2,7|3)		-0.126
			(-0.05)
Paraná|ABC	(1,7|2,3)		-0.043
			(-0.02)
Paraná|Pan	(4,7|1,2,3)		-0.023
			(-0.01)
Paraná|Itaú	(6,7|1,2,3,4)		-0.024
			(-0.01)
Paraná|Banrisul	(5,7|1,2,3,4,6)		-0.025
			(-0.02)

Notas. Ver notas na Tabela 8.

Observe-se que na Figura 2 é revelado que o banco Bradesco transmite mais risco sistêmico para os outros bancos, especialmente para o banco Itaú. Além disso, o Bradesco transmite risco em maneira elevada a todos os outros bancos, com exceção do BB e Pan, que têm uma transmissão mediana. Outra instituição que gera muito risco sistêmico é o BB, que gera principalmente muito risco para o banco Itaú, e, em medida elevada, a outros bancos, excetos para o Banrisul. Quanto ao Itaú, gera bastante risco com respeito aos demais, no entanto, transmite o risco sistêmico de forma muito elevada, exclusivamente para os grandes bancos (Bradesco e BB), apresentando conexão com outro banco muito débil. Os bancos menores em nossa amostra, Banrisul Pan e Abc, tiveram um impacto sistêmico limitado sobre os outros bancos.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figura 2: Mapa da rede do risco sistêmico dos bancos brasileiros.



Em resumo: nossos resultados sobre os efeitos de risco sistêmico indicam que, apesar do fato de que a dependência média entre os bancos foi relativamente alta, a dependência de cauda - crucial para a avaliação dos efeitos de risco sistêmico - não era necessariamente simétrica: encontramos o risco sistêmico do banco 1 para o banco 2, mas não o inverso. Nossos resultados apontam para o papel predominante do Bradesco em termos de risco sistêmico, seja em recepção, seja em transmissão do risco para o restante dos bancos. O BB, apesar de seu tamanho ser maior, desempenhou um papel menor em termos de risco sistêmico. Este resultado é consistente com a ideia de que o BB é um banco com uma forte influência do governo brasileiro. Também o Itaú, apesar de seu tamanho ser maior, desempenhou um papel menor em termos de risco sistêmico, ainda que esteja muito conectado com os outros dois grandes bancos. Nossos resultados apontam para o fato de que os menores bancos, Banrisul, Pan e ABC não desempenham um papel fundamental em termos de risco sistêmico; na verdade, eles nem sequer transmitem riscos entre eles.

Nossos resultados têm três implicações principais. Em primeiro lugar, as estimativas de risco sistêmico devem levar em conta a dependência de cauda e o fato de que isso pode mudar durante o tempo de dificuldades financeiras. A nossa estrutura de dependência multivariada revela que a ausência de dependência condicional na cauda inferior tem um impacto significativo sobre o tamanho dos efeitos sistêmicos e do risco. Em segundo lugar, os nossos resultados têm implicações para os reguladores, como o capital necessário, devido ao risco sistêmico que varia de acordo com as instituições financeiras e através do tempo, mas é especialmente importante durante os episódios de perigo financeiro, quando o risco sistêmico aumenta dramaticamente e levanta dúvidas quanto à capacidade de um sistema financeiro para resistir à dificuldade de uma magnitude, como a da recente crise financeira. Isso abre espaço para se debater a questão da implementação de um sistema dinâmico de requisitos de capital para se responder ao risco sistêmico. Em terceiro lugar, nossos resultados também têm implicações para os investidores em

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

termos de projeto de portfólio e gestão de riscos. Embora a evidência da dependência média positiva indique que não há oportunidades de hedge para os investidores que utilizam ativos bancários, a nossa evidência de risco sistêmico indica que os ganhos em termos de risco em uma carteira podem ser obtidos, observando-se que alguns bancos não têm efeitos de risco sistêmico sobre os outros bancos.

Conclusão:

As recentes crises financeiras têm levantado preocupações públicas e regulamentares relativas ao impacto do risco sistêmico de instituições financeiras falidas ou em desagregação. A avaliação precisa do risco sistêmico, e é crucial para uma regulação eficaz dos riscos e melhoria do impacto da crise financeira sobre o desempenho dos sistemas financeiros.

Nós medimos o impacto sistêmico decorrente de um não cumprimento dos compromissos de um banco e a sua influência para outros bancos usando *CoVaR* como uma medida de risco sistêmico. Para modelar a estrutura de dependência multivariada entre os bancos, foi utilizado um modelo hierárquico de dependência chamado Vine cópula que é capaz de explicar as interconexões, a dependência condicional e características específicas da dependência caudal entre os bancos. Nós também utilizamos um modelo de cópula bivariada de dependência para cada banco e o índice financeiro brasileiro para medirmos a contribuição de risco de cada banco para o conjunto do sistema financeiro brasileiro.

Nossos resultados empíricos - com base em dados relativos ao período de 01 janeiro de 2007 a 18 janeiro de 2016 - indicam que a dependência multivariada entre os bancos é dada por uma estrutura hierárquica C-Vine, no qual predomina Bradesco na determinação da estrutura de dependência condicional. Todos os bancos covariam, em média, ao longo do período da amostra, mostrando evidências de dependência caudal diferentes. Além disso, a contribuição do risco sistêmico de cada banco para outros bancos manteve-se semelhante ao longo do período analisado. Bradesco desempenhou um papel preponderante na medida em que transmite e recebe o risco sistêmico para/de os outros bancos. Itaú desempenhou um papel menor.

Referências:

Acharya, V. V., Pedersen, L. H., Philippon, T., & Richardson, M. P. (2010). Measuring systemic risk.

Adrian, T., & Brunnermeier, M. K. (2011). *CoVaR* (No. w17454). National Bureau of Economic Research.

Allen, L., Bali, T. G., & Tang, Y. (2012). Does systemic risk in the financial sector predict future economic downturns?. *Review of Financial Studies*, *25*(10), 3000-3036.

Bernal, O., Gnabo, J. Y., & Guilmin, G. (2014). Assessing the contribution of banks, insurance and other financial services to systemic risk. *Journal of Banking & Finance*, *47*, 270-287.

Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, *104*(3), 535-559.

Bisias, D., Flood, M. D., Lo, A. W., & Valavanis, S. (2012). A survey of systemic risk analytics. *US Department of Treasury, Office of Financial Research*, (0001).

Brownlees, C. T., & Engle, R. (2012). Volatility, correlation and tails for systemic risk measurement. working paper.

Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367-381.

Girardi, G., & Ergün, A. T. (2013). Systemic risk measurement: Multivariate GARCH estimation of CoVaR. *Journal of Banking & Finance*, *37*(8), 3169-3180.

Huang, X., Zhou, H., & Zhu, H. (2009). A framework for assessing the systemic risk of major financial institutions. *Journal of Banking & Finance*, *33*(11), 2036-2049.

Joe, H. (1996). Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, 120-141.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC Press.

Nelsen, Roger B. "An introduction to copulas, 2nd. *New York: SpringerScience Business Media* (2006).

Reboredo, J.C & Ugolini, A. (2015). Systemic risk in European sovereign debt markets: A CoVaR-copula approach. *Journal of International Money and Finance, 51*, 214–244.

Rodríguez-Moreno, M., & Peña, J. I. (2013). Systemic risk measures: The simpler the better?. *Journal of Banking & Finance*, *37*(6), 1817-1831.

Rosenberg, J. V., & Schuermann, T. (2006). A general approach to integrated risk management with skewed, fat-tailed risks. *Journal of Financial economics*, *79*(3), 569-614.

Das, S. R., & Uppal, R. (2004). Systemic risk and international portfolio choice. *The Journal of Finance*, *59*(6), 2809-2834.

Segoviano, M. A., & Goodhart, C. (2009). Banking stability measures.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Université Paris 8*.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 28 - ESQUEMA OPERACIONAL DE BAIXO CUSTO PARA VERIFICAÇÃO ESTATÍSTICA DE MODELOS NUMÉRICOS DE PREVISÃO DO TEMPO

Nilza Barros da Silva [22]
Natália Santos Lopes[23]

Resumo

Apesar da importância da validação estatística, apenas alguns poucos serviços nacionais de previsão do tempo contam com um sistema implementado e automatizado. Uma das prováveis razões para esta dificuldade é o tempo gasto no preparo dos dados, no que se refere ao controle de qualidade e todo o processo necessário para combinar os pares de observações com as previsões. Além disso, muitos centros de previsão ou de pesquisa vivem uma realidade de constantes restrições orçamentárias. O objetivo deste trabalho é mostrar que estes centros podem implementar seu próprio sistema de verificação a partir do uso de softwares livres ou abertos. O esquema de validação estatística que será apresentado esta implementado no Centro de Hidrografia da Marinha desde 2009. Esta solução foi aplicada, pois os dados (observações e previsões) encontravam-se espalhadas em vários servidores e não havia um processo para avaliar o desempenho dos modelos de previsão. O aplicativo R é utilizado para consultar e alimentar o banco dados, para construir gráficos, tabelas e estatísticas. Os produtos finais são disponibilizados em servidores e na Internet. Em resumo, o processo apresentado pode ser utilizado por centros de previsão ou de pesquisa que não têm um banco de dados organizado e por organizações com poucos recursos para investir em programas proprietários.

Palavras-Chave: agendamento, verificação, Shell, RmySQL, scripts

Abstract

Regardless of the acknowledge importance of verification system there only a few National Weather Service (NWS) with an operational verification system. One of the reasons is the time spent on data management issues such as performing quality control and pairing forecast with observation. Nevertheless, there are many NWS or small research facilities that have to deal with low budget and lack of resources. This work presents a low-cost operational forecast verification system that use only free and open-source software. The goal is to show that small or low budget NWS can implement their own operational verification system. The verification process is implemented at Navy Hydrography Center since 2009. That solution was applied because their data (observations and forecast) were scattered across multiple servers and there were no structured activities to show how their models were improving. R scripts are used to consult and feed the database. All products such as graphics, tables and statistics are produced by R scripts. All process is executed periodically at a set time on Linux. The end products are available to the forecaster in a server or on the Internet. In conclusion, that process can be used by those NWS that do not have an organized database or by those organizations that work with low budgets.

Keywords: scheduling, verification, Shell, RmySQL, scripts

[22] Centro de Hidrografia da Marinha (CHM) – nilza.barros@marinha.mil.br
[23] Centro de Mísseis e Armas Submarinas da Marinha (CMASM) – natalia.lopes@marinha.mil.br

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

A Marinha do Brasil, por meio do Centro de Hidrografia da Marinha (CHM), tem por obrigação produzir e divulgar análises e previsões meteorológicas para a área marítima de responsabilidade do Brasil, conhecida internacionalmente como METAREA V. Em virtude disso, o CHM é a organização militar responsável por gerar diariamente previsões oriundas de modelos atmosféricos e oceanográficos a fim de contribuir para a melhoria da qualidade das informações ambientais elaboradas e disseminadas pelo Serviço Meteorológico Marinho. Contudo, a qualidade destas previsões deve ser auferida e medida por meio de técnicas estatísticas, conhecida como verificação.

A verificação estatística é o processo de avaliação da qualidade das previsões e envolve comparações entre previsões e observações. Além disso, ela também apresenta outros objetivos, tais como: medir o desempenho dos modelos quando suas características forem alteradas; comparação com outros modelos a fim de identificar qual tem melhor desempenho para cada tipo de parâmetro; monitorar a distribuição espacial e temporal dos erros de modo a permitir que o previsor do tempo conheça as limitações do modelo, entre outros.

Entretanto, até que se obtenha o conjunto de dados a serem validados, faz-se necessário um pré-processamento dos mesmos. Durante a construção do sistema de verificação, a maior parte do tempo é utilizado na extração dos dados, no controle de qualidade e na combinação de pares de observação e previsão (CASATI *et al*.,2008). Isto ocorre, pois muitas vezes o formato dos dados de um determinado modelo numérico difere das observações e vice-versa. Esta etapa, normalmente, exigirá do analista o conhecimento de vários programas para decodificar os dados originais e deixá-los no formato final, onde será possível iniciar toda a análise estatística. Além do mais, as variáveis provenientes de modelos numéricos apresentam uma larga variação espacial e temporal, o que exige uma grande capacidade de síntese e consolidação das informações, de forma a permitir uma análise estatística que seja útil e confiável para os usuários finais.

Desta forma, faz-se necessário a organização e o agrupamento de dados que usualmente encontram-se distribuídos em diferentes locais, tais como: servidores, sites na Internet, sistemas de distribuição de dados, etc.

Objetivo

O objetivo deste trabalho é apresentar um sistema de baixo custo, que permitirá a criação de uma estrutura de pré-processamento de informações, a partir do uso de softwares livres.

Embora seja apresentada uma solução aplicada na validação estatística de modelos de previsão do tempo, nada impede que o método ou processo seja adaptado para qualquer tipo de projeto que apresente características como: dados oriundos de fontes diversas, necessidade de organização de um grande volume de dados, grande variação espacial e temporal de dados e apresentação de resultados consolidados em estatísticas, tabelas e gráficos.

O software R, além da geração das estatísticas e dos gráficos, é a ferramenta que faz a ligação entre os diversos aplicativos utilizados neste trabalho, como por exemplo, o MySQL cuja a interface é acessada pelo pacote RMySQL.

Material e Métodos:

Esta solução foi desenvolvida a partir das necessidades listadas abaixo:

Criação de rotinas de validações estatísticas dos modelos de previsão do tempo operados pelo CHM;

Organização e agrupamento de dados dispersos; e

Resultados objetivos que demonstrem se os modelos numéricos, cuja operacionalização exige um grande aporte de recursos humanos e financeiros, apresentam o desempenho esperado.

Todo o sistema foi instalado no sistema operacional Linux. Para que sejam possíveis as funcionalidades aqui citadas há necessidade dos programas listados abaixo:

MySQL: sistema de gerenciamento de banco de dados (BD) que utiliza a linguagem SQL (*Structured Query Language)*;

phpMyAdmin (opcional): permite a administração do MySQL pelo *browser.* A interface amigável possibilita, para aqueles que não tenham conhecimento da linguagem SQL, o acesso e a gerência do BD com facilidade;

Apache: servidor Web, dependência exigida para uso do phpMyAdmin;

R (R Development Core Team, 2016): software livre para computação estatística e construção de gráficos;

RStudio (opcional): interface gráfica para uso do R;

RMySQL (OOMS et al.,2016): interface para uso e acesso do MySQL dentro da interface do R;

Rscript: é um *front-end* que permite que o Linux interprete os comandos do R pelo *bash*, o que possibilita que os scripts sejam escritos e executados em R diretamente sem necessidade do uso da interface do R.

Com exceção dos pacotes do R, que devem ser instalados dentro deste programa, os outros podem facilmente ser instalados com os gerenciadores de pacotes disponíveis para Linux como o APT.

Segue abaixo, os conhecimentos mínimos necessários para operacionalização do sistema:

Sistema operacional Linux (instalação de programas e comandos básicos);

Básico de Shell Script;

Utilização Crontab do Linux;

Básico de SQL; e

Programação em R.

Resultados e Discussão:

Resumo Esquema de Validação

O esquema apresentado na figura 1 ilustra o funcionamento do sistema que consiste de:

Agendamento de tarefas com utilização do crontab do Linux que permite a alimentação do banco de dados denominado Verificação com dados de diversos modelos de previsão tempo. O mesmo banco de dados é alimentado com dados de observações que podem estar disponíveis tanto em servidores próprios como na Internet;

Acesso ao BD, com uso do pacote RMySQL, dentro dos próprios scripts responsáveis pela geração das figuras e pelos cálculos das estatísticas. Isso é possível graças ao uso do *front-end* Rscript que permite que os comandos do R sejam reconhecidos pelo Shell do Linux sem necessidade de acesso à interface do R.

Utilização do phpMyAdmin para acesso ao BD, criação de novo bancos ou tabelas, etc.

Todas as etapas acima foram automatizadas e permitem que diariamente sejam armazenados os dados das previsões dos modelos e as observações. Nesta automatização são gerados produtos diários, mensais ou trimestrais. Para estas

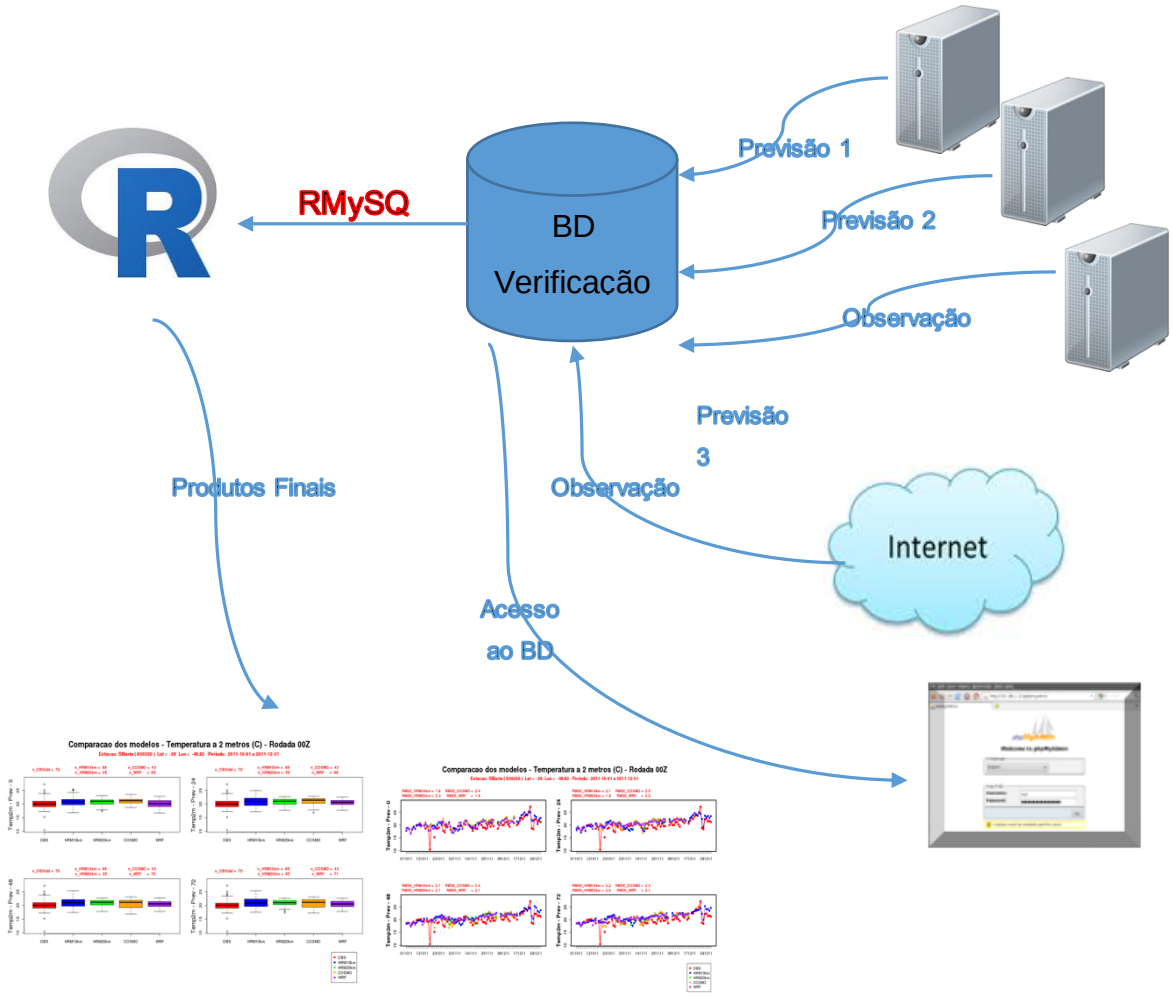
customizações são utilizados o agendador de tarefas do Linux (Crontab) e todas as funcionalidades dos scripts em R utilizando o Rscript.

Uso do Crontab

O uso desta funcionalidade possibilita a automatização dos processos de alimentação diária do banco de dados Verificação e, em período definidos, da geração dos produtos em forma de tabelas e gráficos. Como pode ser observado na figura 2 o Crontab admite a definição do dia, da data e da hora de rodada de determinado script.

A figura 2 exibe o agendamento diário para diversas tarefas como, por exemplo, a extração das previsões dos modelos e a alimentação de diversas tabelas no banco de dados. Os scripts estão escritos em Shell ou em R.

Figura 1- Esquema de validação estatística



```
# Crontab do usuario gempak
#
#####################################################################
#                       Scripts Ativos
#####################################################################
#
# Script de extracao SYNOP do IDD para BD  ORMVERIF
#
30 * * * * /home/gempak/ormverif/scripts/extrai_syn.sh 00
55 * * * * /home/gempak/ormverif/scripts/extrai_syn.sh 12
#--------------------------------------------------------------------
# Script  extrai  GRIB (COSMO 10KM )
# Alimenta BD verificacao - tabela FCT_COSMO e FCT_COSMO_near
#
40 02 * * * /home/gempak/cosmo/scripts/geradados.sh 00
#--------------------------------------------------------------------
# Script  extrai  GRIB (WRF )
# Alimenta BD verificacao - tabela FCT_WRF e FCT_WRF_near
#
30 08 * * * /home/gempak/wrf/scripts/geradados.sh 00
#--------------------------------------------------------------------
# Scripts para alimentar BD Verificacao - tabela OBS_idd 00Z
#
30 08 * * * /home/gempak/ormverif/scripts/LeSflistFile 00
#--------------------------------------------------------------------
# Script  gera graficos para verificacao e envia para DPNT01
#
00 09 * * * /home/gempak/ormverif/scripts/verif_fig_modelos dia
30 09 * * * /home/gempak/ormverif/scripts/GrafHorarioModelos.R dia NULL NULL COSMO
#--------------------------------------------------------------------
```

Figura 2 - Agendamentos do Crontab

Uso do Banco de Dados e do phpMyAdmin

O objetivo principal para uso do MySQL foi tirar proveito das funcionalidades de um BD no que se refere à organização, ao armazenamento e, principalmente, às

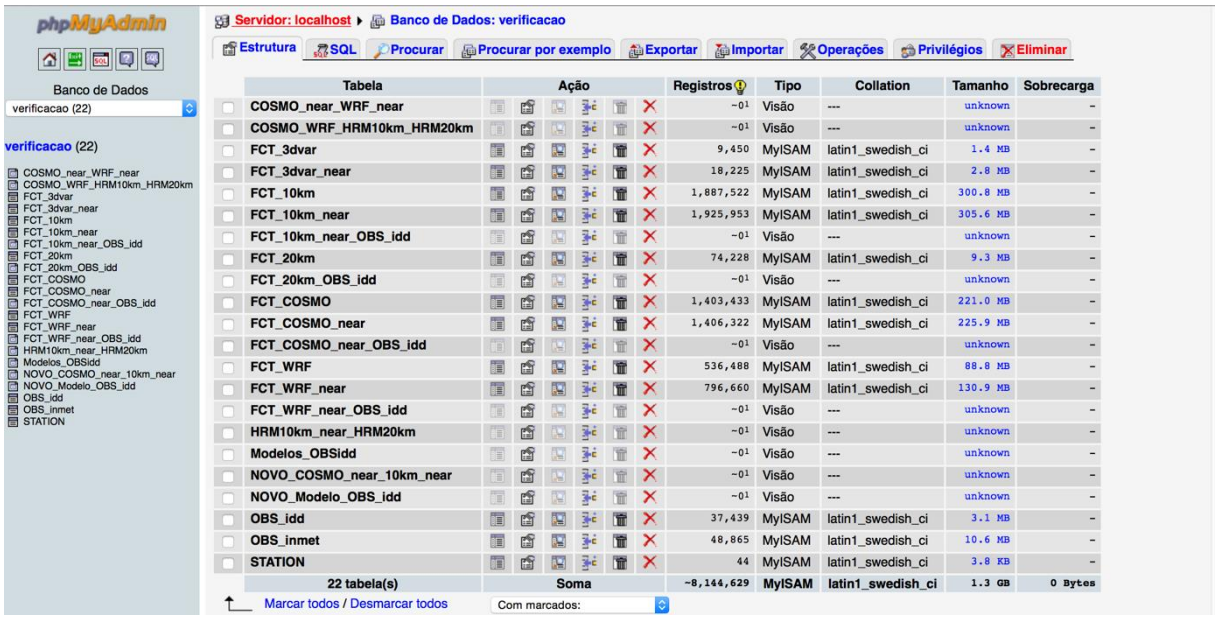


Figura 3 - Banco de Dados Verificação

possibilidades de se fazer o cruzamento das informações e dos registros de forma mais ágil e com menos chances de erro. Uma vez que, neste tipo de validação, faz se necessário a combinação das observações geradas em determinado dia com as previsões de até 5 dias antes.

A figura 3 exibe o Banco de Dados verificação a partir da interface do phpMyAdmin**.**

Scripts operacionais desenvolvidos com o R

Alimentação do BD

As figuras 4 e 5 apresentam os comandos do RMySQL utilizados para o acesso ao banco de dados verificação. Observa-se que a forma é bastante simples, mesmo para usuários pouco habituados ao uso de comandos SQL.

```
#! /usr/bin/Rscript --vanilla
rm (list=ls()) # remove todas as variaveis do ambiente R
args <- commandArgs(TRUE)

###############################################################################################
#### Leitura de Dados do COSMO                                                             ####
####-------------------------------------------------------------------------------------####
####                                                                                       ####
#### Objetivo: ler dados do COSMO, ajustar estacoes e  alimentar o BD verificacao   ####
###############################################################################################
#Diretorio de Trabalho
wrk.dir <- "/home/gempak/ormverif/scripts/"
dat.dir <- "/home/gempak/cosmo/saidas/"
date.dir <- "/home/gempak/datas/"
stn.dir <- "/home/gempak/cosmo/scripts/"
setwd(wrk.dir)
library(RMySQL)
drv=dbDriver("MySQL")
con=dbConnect(drv,dbname='verificacao', user='xxxx',password='xxxx')
```

Figura 4 - Acesso ao BD MySQL pelo comando dbDriver e dbConnect do RMySQL

```
208 ##############################################################################################
209 #         RMYSQL - ALIMENTAR O BANCO
210 ##############################################################################################
211 if( npt.near=="N"){
212 if(dbExistsTable (con,paste("FCT_COSMO", sep=""))) {
213 feeddb<-dbWriteTable(con,paste("FCT_COSMO", sep=""),fct.cosmo,append=T,row.names=FALSE)
214 print(" Resultado alimentacao BD")
215 print(feeddb)
216 }else {
217 feeddb<-dbWriteTable(con,paste("FCT_COSMO", sep=""),fct.cosmo, sep="",row.names=FALSE)
218 }
219 dbDisconnect(con)
220
221 }else{
222 if(dbExistsTable (con,paste("FCT_COSMO_near", sep=""))) {
223 feeddb<-dbWriteTable(con,paste("FCT_COSMO_near", sep=""),fct.cosmo,append=T,row.names=FALSE)
224 print(" Resultado alimentacao BD")
225 print(feeddb)
226 }else {
227 feeddb<-dbWriteTable(con,paste("FCT_COSMO_near", sep=""),fct.cosmo, sep="",row.names=FALSE)
228 }
229 dbDisconnect(con)
230 }
```

Figura 5 - Comando dbWriteTable para alimentar o BD

A figura 6 exibe um exemplo de uso de comandos SQL para extração de dados de uma determinada tabela do banco de Dados Verificação. Como pode ser observado os comandos são executados dentro do próprio script do R. Nesta etapa há necessidade do conhecimento de alguns comandos em SQL.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

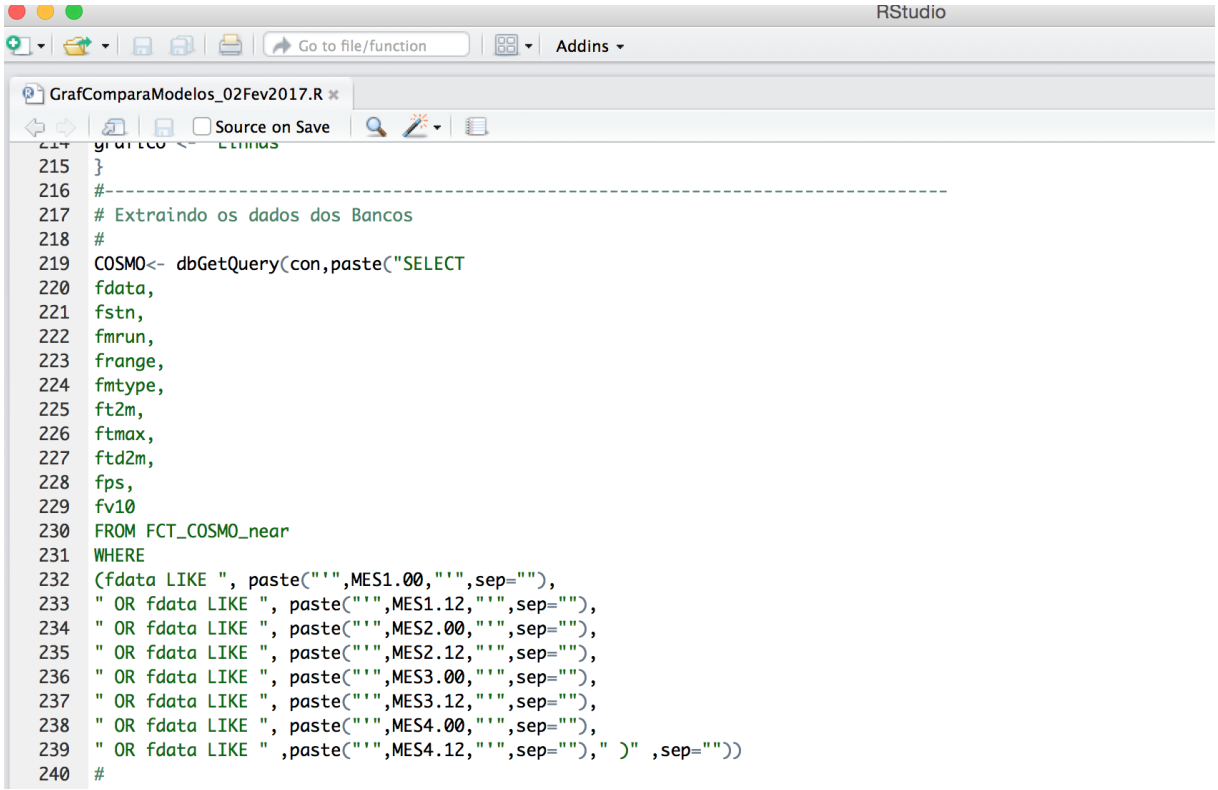


Figura 6 - Uso do comando dbGetQuery

Exemplos de produtos gerados

As figuras 7, 8 e 9 exibem exemplos de produtos que são gerados para os previsores do tempo, onde se comparam os dados observados em estações meteorológicas com as previsões geradas pelos modelos. No mesmo gráfico são exibidas estatísticas recomendadas por WILKS (2006) que permitem identificar qual modelo teve o melhor desempenho no período estudado.

Estes gráficos e estatísticas permitem que o previsor avalie o desempenho dos modelos numéricos, qual deles apresenta melhor resultado, se um dado modelo superestimar ou subestimar determinada previsão. Este tipo de análise auxilia o previsor no momento de gerar suas próprias previsões, pois o conhecimento, por exemplo, que um dado modelo tende a superestimar a previsão de temperatura permitirá que o profissional faça o ajuste necessário quando da confecção dos boletins de previsão do tempo que serão divulgados para o usuário final.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

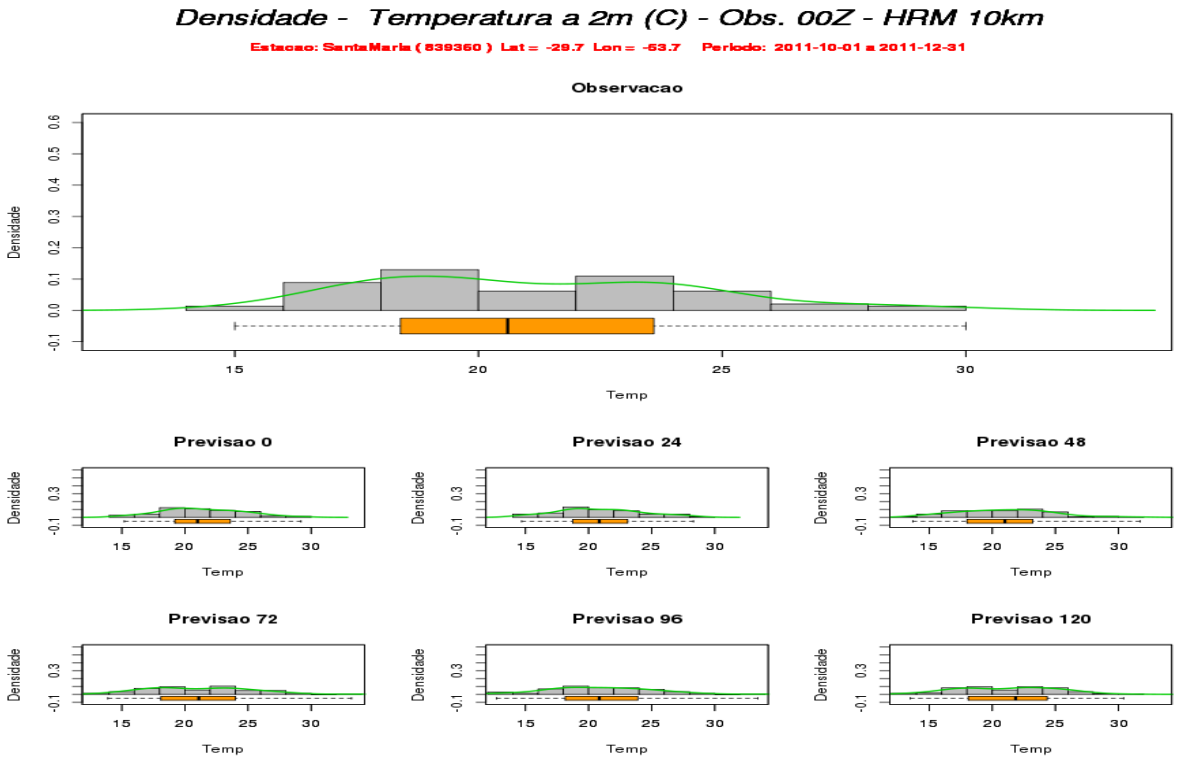


Figura 7 – Grafico densidade e box-plot para medir desempenho de modelo de atmosféricos para as previsões de 0 a 120 horas.

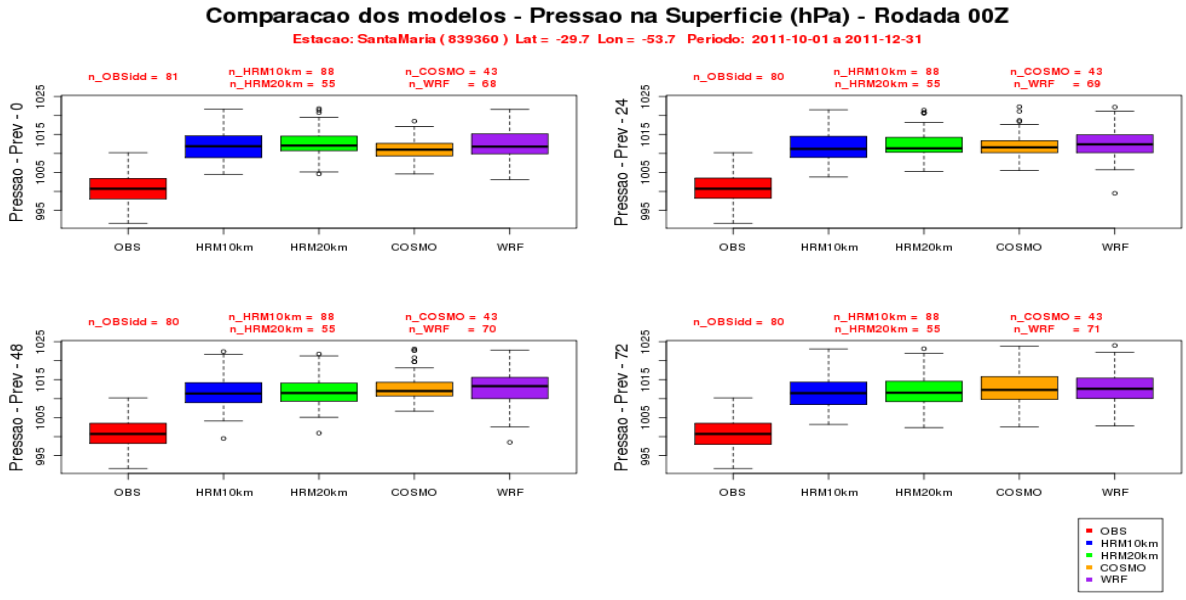


Figura 8 – Box-plot para comparação entre a observação e vários modelos atmosféricos para as previsões de 0 a 72 horas.

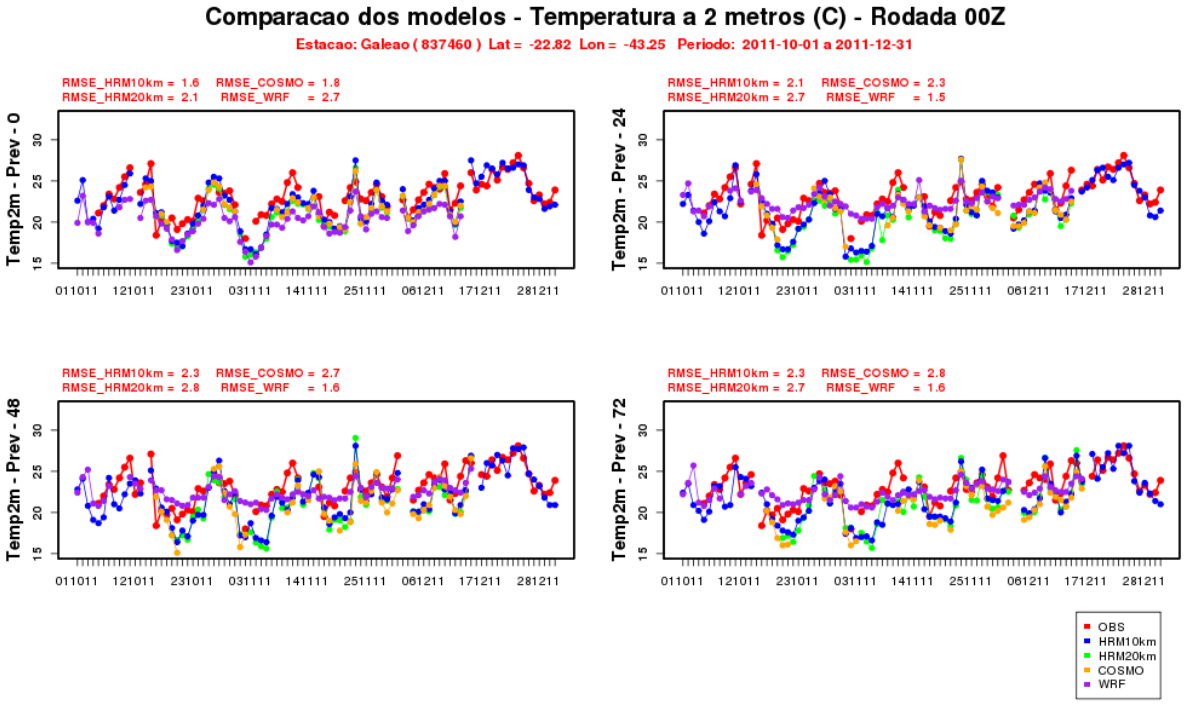


Figura 9 – Gráficos de linhas e estatística para comparação entre diversos modelos atmosféricos para as previsões de 0 a 72 horas.

Conclusão:

Destacam-se a seguir os ganhos obtidos com a operacionalização do esquema apresentado:

permite a consolidação das informações de interesse em um banco de dados que pode ser facilmente acessado com uso de pacotes disponíveis no próprio R;

O uso das funcionalidades de um agendador de tarefas como o crontab do Linux permite a automatização de todas as etapas do processo;

o Rscript é um pacote muito importante para aqueles que querem utilizar o R em todas as etapas do esquema, pois evita o uso de programas intermediários entre o R e o banco de dados e entre o R e a geração de estatísticas, gráficos e tabelas;

o pré-processamento das informações demandará menos tempo ao analista, que poderá dedicar-se às análises estatísticas e à geração de novos produtos que atendam aos usuário finais, neste caso, aos previsores do tempo; e

Uso de softwares livres que implicam em menores custos.

Embora o esquema apresentado tenha sido operacionalizado em Linux, com alguns ajustes poderá ser implementado em qualquer outro sistema operacional. O

intuito deste trabalho é mostrar um processo de baixo custo que pode ser utilizado em qualquer organização que necessite analisar uma grande quantidade de dados de forma contínua e sistematizada.

Referências:

CASATI, B., WILSON, L.J., STEPHENSON, D. B., NURMI, P., GHELLI, A., POCERNICH, M., DAMRATH, U., EBERT, E.E., BROWN, B. G., MASON, S., *Forecast verification: current status and future directions*. Meteorological Applications, v.15, pp. 3-18, Mar. 2008.

OOMS ,JEROEN, JAMES,DAVID, DEBROY, SAIKAT, WICKHAM, HADLEY e HORNER,JEFFREY (2016). RMySQL: Database Interface and MySQL Driver for R. R package version 0.10.9. https://CRAN.R-project.org/package=RMySQL

R Core Team (2016). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

WILKS, D.S., 1995, Statistical methods in the atmospheric sciences. 2nd ed. San Diego: Academic Press, 2006, 627 p.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 30 - MODELO DE REGRESSÃO LOG-SIMÉTRICA: UMA APLICAÇÃO COM DADOS DE CINEMA

Marcelo dos Santos Ventura[24]
Helton Saulo[25]

Resumo

O presente trabalho é destinado a aplicar o modelo Log-simétrico a dados de cinema, explorar a modelagem de arrecadação dos filmes, e ressaltar as possibilidades criadas com uma base de dados para filmes. Com o modelo de regressão log-simétrico, foram analisados os dados de 155 filmes contidos no site IMDB(Internet Movie Database). As variáveis incluídas no modelo são as notas dos usuários, o número de votantes, o custo dos filmes e sua receita. A variável resposta, a arrecadação dos filmes, é modelada em um submodelo para a mediana e outro para a assimetria. O primeiro possui como variáveis explicativas os custos dos filmes e o número de votos. A variável notas médias foi escolhida para modelar a assimetria. Os resultados sugerem que o modelo ajusta os dados gerando estatísticas significantes para a modelagem da mediana e assimetria considerando a variável resposta como arrecadação. A família log-exponencial é escolhida utilizando critérios de seleção de modelo.

Palavras-Chave: (de 3 a 5 palavras) IMDb, Log-simétrica,Cinema.

Abstract

This work is destined to apply the Log-symmetric model to movie data, explore movie gross modelling, and highlight possibilities created with a movie database. By the regression Log-symmetric model, there were analyzed data of 155 movies on IMDb's (Internet Movie Database) webpage. The ranking, numbers of votes, budgets, and grosses of each movie were included in the model. As response variable, movie's grosses is modeled in a sub model for the median, and another for skewness. The first one has independent variables like movie budgets and numbers of votes. The variable ranking was choosed to model skewness. Results suggests that model fits data generating significant statistics for median and skewness modelling considering gross as the dependent variable. Log-exponential family is assumed by model selection criteria.

Keywords: IMDb, log-symmetric, movies

Introdução

O software estatístico R permite aos usuários uma atmosfera de constante colaboração e evolução. Embora seja um software estatístico, ele também cria aplicações em áreas além da Estatística como Economia, Psicologia, Administração e Biologia. Uma das possibilidades a serem exploradas é a criação de bases de dados com extração de informações de sites e suas aplicações a modelos estatísticos.

[24] UFG (Universidade Federal de Goiás), marcelos.ventura@gmail.com
[25] UFG (Universidade Federal de Goiás), heltonsaulo@gmail.com

O modelo de regressão Log-simétrico possui interessantes propriedades estatísticas, entre elas a modelagem da mediana e assimetria (dispersão relativa) da variável resposta. A primeira, é encontrada com uma função não linear, e a segunda, uma semi-paramétrica. Como (VANEGAS; PAULA,2016) cita em seu artigo, a variável resposta assimétrica, contínua, e estritamente positiva, possui ampla utilidade em diversos campos da ciência. Além disso, a distribuição dos erros multiplicativos independentes e aleatórios é pertencente a classe log-simétrica, que contém as distribuições log-Normal, log-Student-t, log-Slash, Birnbaum-Saunders, entre outras.

O site IMDb (Internet Movie Database) contém dados sobre filmes, jogos, e séries audiovisuais. A base de dados teve seu início em 1989 com o britânico Col Needham, quando trabalhava na empresa Usenet. Um ano depois a base já detinha informações sobre atores, atrizes e diretores em mais de 10000 séries televisivas e filmes. Como nessa época ainda não existiam navegadores de internet, as informações eram divulgadas em scripts de shell(linguagem de script utilizada em sistemas operacionais) para o sistema UNIX, e a base era conhecida como "rec.arts.movies movie database". No entanto em 1993, com o desenvolvimento da World Wide Web a base de dados então ficou disponível e em 1996 a Internet Movie Database ltd. se tornou um empreendimento comercial. Atualmente o site é propriedade da Amazon.com.

Neste trabalho a variável explicada é a arrecadação de cada filme (em dólares), e as variáveis explicativas são as notas, o número de avaliadores, e o orçamento dos filmes. Uma vantagem em utilizar a regressão citada para esses dados é o melhor ajuste do modelo com a flexibilização da modelagem ao permitir diversas famílias de distribuição para os dados. Ou seja, aplicando os dados ao modelo de regressão log-simétrico, é possível escolher uma família presente na classe Log-Simétrica que se adeque melhor aos dados. Essa decisão é tomada utilizando critérios de ajuste, além de estatísticas como AIC e BIC, critérios de informação de Akaike e Bayesiano respectivamente.

Objetivo

O presente trabalho é destinado a aplicar o modelo Log-simétrico a dados de cinema, explorar a modelagem de arrecadação dos filmes com a escolha das famílias da classe log-simétrica, e ressaltar as possibilidades criadas com uma base de dados para filmes. Com o modelo de regressão log-simétrico, foram analisados os dados de

155 filmes contidos no site IMDB(Internet Movie Database). As variáveis incluídas no modelo são as notas dos usuários, o número de votantes, o custo dos filmes e sua receita. As variáveis explicativas para a mediana são o custo dos filmes e o número de votantes, enquanto a assimetria é modelada com as notas dos filmes.

Material e Métodos:

Como (VANEGAS; PAULA et al.,2016) explora em seu artigo, existem muitas distribuições que suportam o intervalo (0, ∞). No entanto, a distribuição log-normal obteve sucesso em uma considerável quantidade de aplicações. A classe de distribuição log-simétrica é uma flexibilização da distribuição log-normal, de forma a considerar distribuições bimodais, e as que possuam caudas maiores ou menores que a distribuição log-normal. Além de considerar outras distribuições além da distribuição normal, o modelo é análogo ao MLG (modelo linear generalizado), porém com ferramentas estatísticas mais aprimoradas. Como será visto nos resultados, o modelo permite que uma família da classe Log-simétrica seja escolhida para os dados, de forma a obter o melhor ajuste possível.

Além disso, a classe log-simétrica possui dois parâmetros, que são a mediana e a assimetria. Para medidas de posição, dispersão, assimetria, e curtose normalmente a estatística sugere as suas estimações utilizando os momentos da variável. No entanto, essa abordagem pode ser problemática para distribuições assimétricas, dada a possibilidade de um momento ser infinito ou seu cálculo complexo. Por isso como (VANEGAS; PAULA et al.,2016) deriva em seu artigo, temos as medidas referidas para a classe log-simétrica, mesmo quando a distribuição em questão não possuir momentos. As possibilidades criadas pelo modelo tornam ele mais flexível estatisticamente em relação a métodos de estimação como Mínimos quadrados ordinários, e modelos que considerem apenas uma família de distribuição, ou que adotem assimetria constante para todas as observações.

Com relação aos dados utilizados, a média dos anos dos filmes presentes na base é de 1992, e o terceiro quartil é igual a 2007. Isso mostra que mesmo com constantes filmes sendo produzidos, e a coleta de dados ter sido no dia 31 de janeiro de 2017, mês que antecedeu a cerimônia do Oscar, os usuários em sua maioria não se limitam a avaliar somente filmes recentes, como observado na Figura 1 abaixo:

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

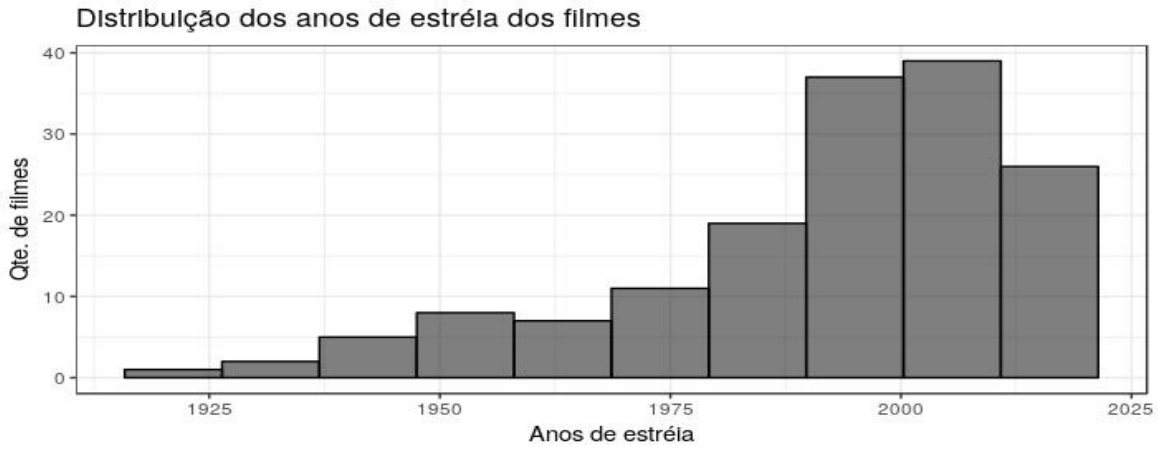


Figura 1 – Distribuição dos anos de estréias dos filmes

A variável explicada do modelo é a arrecadação dos filmes. Há uma grande concentração de filmes abaixo da arrecadação de 200 milhões de dólares, mais precisamente 120 filmes. A média dessa variável é de U$$127.400.000 enquanto a mediana é U$$60.990.000. A variável resposta possui as características necessárias para a modelagem via regressão log-simétrica: assimetria, dados positivos, e continuidade. Todas as arrecadações foram medidas apenas nos Estados Unidos, incluindo filmes estrangeiros. A Figura 2 é mostrada a seguir:

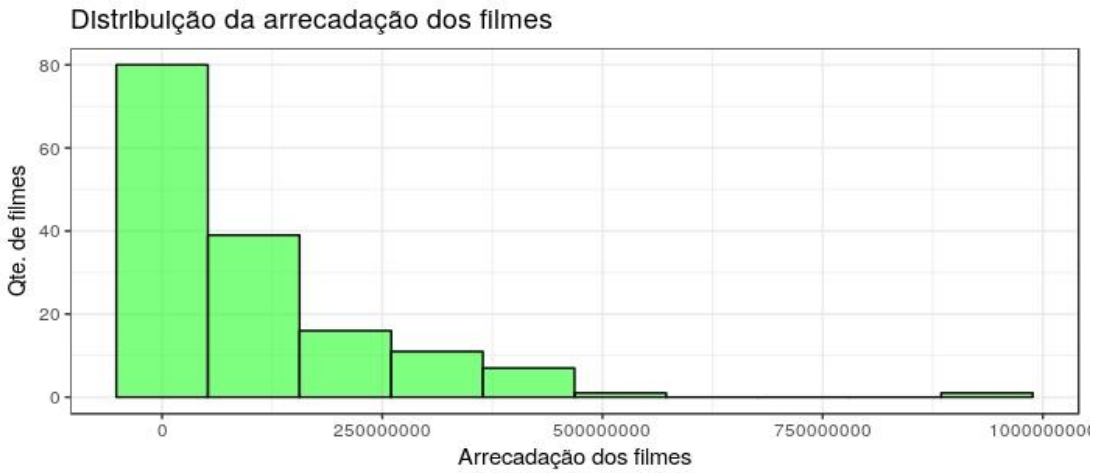


Figura 2 – Distribuição das receitas por filme

A primeira variável explicativa a se considerar serão as notas. Observando as descrições a seguir e a Figura 3, é possível considerar uma distribuição assimétrica à direita. Dos 155 filmes da subamostra, 25 apenas se encontram após a nota 8.5. Nessa variável, a média é de 8.325 e a mediana é de 8.3. Existem 17 filmes com nota mínima de 8, entre eles Monstros S.A. (2001), Zootopia (2016), Tubarão (1975), Exterminador do Futuro (1984), e Os Suspeitos (2013). Com nota máxima de 9.2 estão Um sonho de liberdade (1994), e O poderoso chefão (1972), com respectivas receitas de U$$28.341.469 e U$$37.874.302.

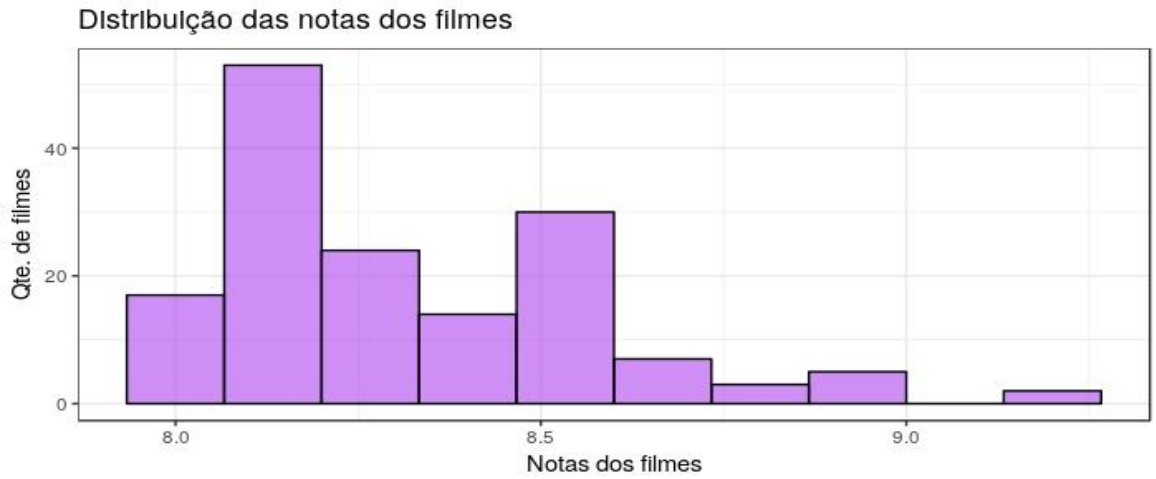


Figura 3 – Distribuição das notas por filme

A segunda variável explicativa, o orçamento, foi ajustada a preços de 2016 da mesma forma que a variável arrecadação. O custo médio dos filmes contidos na base é de U$$ 50.180.000, e a mediana U$$ 28.250.000. O filme com menor orçamento, de U$$ 254.289, é Filhos do paraíso, lançado nos EUA em 2009 e produzido no Irã. Já o filme com maior orçamento a preços de 2016 é O cavaleiro das trevas ressurge (2012), com custo de U$$269.942.197. Observando a Figura 4, é notável o mesmo tipo de assimetria da variável arrecadação:

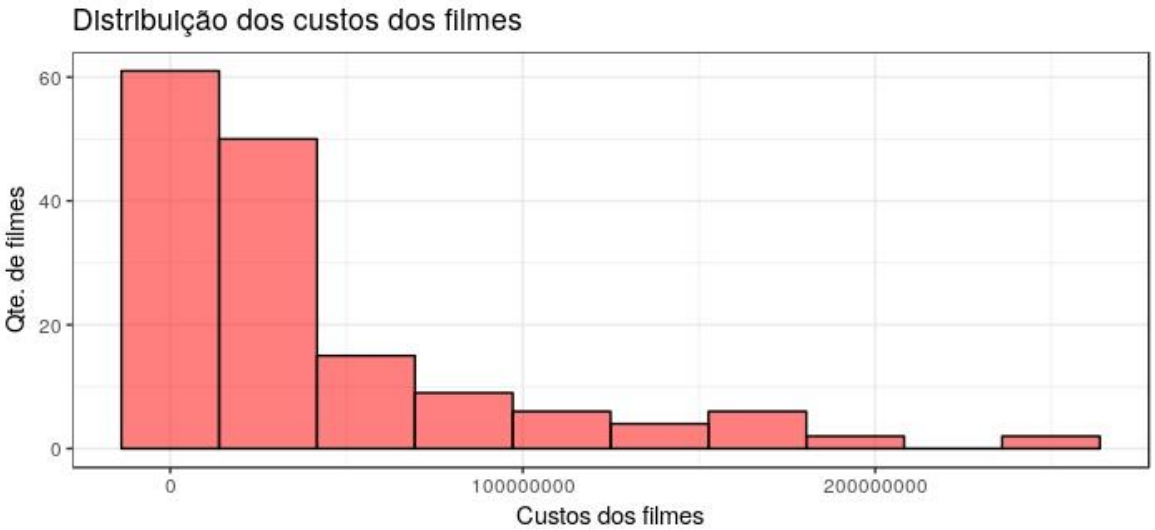


Figura 4 – Distribuição dos custos por filme

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Por fim, como pode ser visto na Figura 5 abaixo, o número de votos é uma variável assimétrica à direita, ainda que em menor grau que as demais variáveis. A média dessa variável é 550.300, e a mediana é 514.100. O filme da amostra com menos votos é Filhos do Paraíso, com 32.185 votos. O mais votado é Um sonho de liberdade, o filme número 1 da lista de 250 filmes.

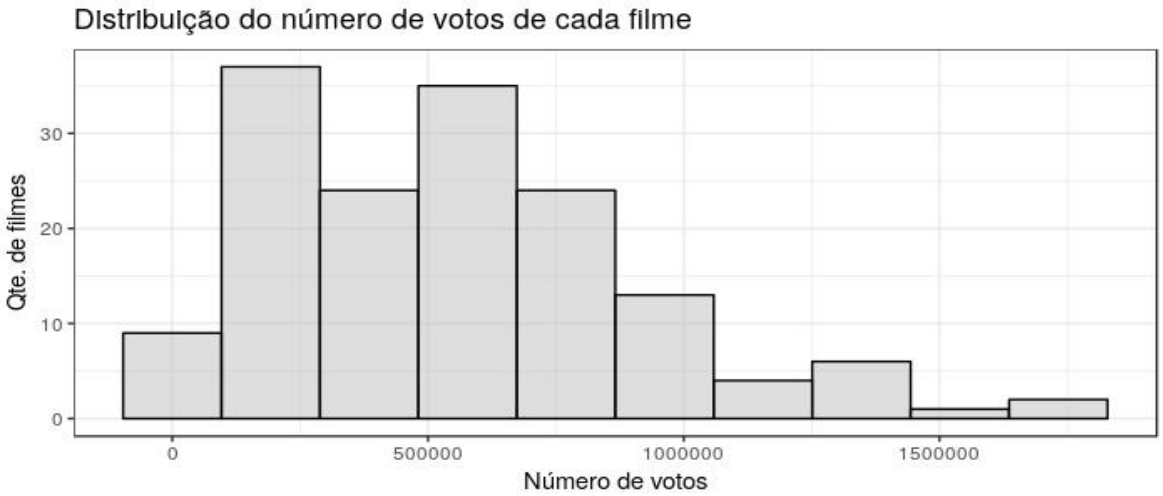


Figura 5 – Distribuição dos votos por filme

Resultados e Discussão:

Os resultados são descritos a seguir na Tabela 1. Primeiro foi gerada uma tabela com as estatísticas de ajuste, AIC e BIC. Com a família escolhida, os coeficientes são revelados.

Tabela 1 – Resultados para famílias da classe log-simétrica

Famílias	ζ	Υ	-2log-veross.	AIC	BIC
Log-Student	2.55	0.098426	508.952	518.952	534.17
Log-Power exponencial	**0.48**	**0.096057**	**497.884**	**514.967**	**540.962**
Log-Hiperbólica	1.3	0.097673	498.955	515.69	541.154
Log-Slash	1.6	0.098303	501.386	517.108	541.032
Log-Normal-contaminada	0.355 e 0.09	0.114677	513.02	524.141	541.063

Observando a Tabela 1, as famílas consideradas são descritas na primeira coluna. O parâmetro extra ζ foi escolhido a partir da minimização da estatística Y. Ela é útil visto que números menores caracterizam maiores ajustes, de tal forma os melhores parâmetros podem ser escolhidos. Comparando as estatísticas da terceira coluna a família Log-Power- exponencial possui melhor ajuste. O mesmo ocorre nas duas próximas colunas. Por fim, a estatística BIC sugere a família Log-Student, ainda

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

que ela fique em quarto lugar nas estatísticas de ajuste. Provavelmente isso ocorre pelo tamanho da amostra.

A seguir, são descritos os resultados para a modelagem da mediana gerados pelo software estatístico R, considerando a família log-power-exponencial.

```
******************** Median/Location submodel ********************
link: identity
 ******** Parametric component

                   Estimate  Std.Err  z-value              Pr(>|z|)
 (Intercept)  16.0968760903   0.1647  97.7616 <  0.00000000000000022 ***
 votes         0.0000017768   0.0000   5.6957      0.000000012289876 ***
 orcmil        0.0000127309   0.0000   7.2234      0.000000000000507 ***
```

Com os resultados acima, podemos observar que todos os coeficientes são significativos a aproximadamente 99% de confiança para o submodelo da mediana. Para o submodelo da assimetria, com aproximadamente 87% de confiança, temos os resultados a seguir:

```
**********************Skewness/Dispersion  submodel******************
link: logarithmic
******** Parametric component

Estimate Std.Err z-value Pr(>|z|) (Intercept)
             0.31711  0.3656   0.8674
0.3857


******** Nonparametric component

 Smooth.param Basis.dimen                d.f. Statistic p-value
psp(rating)       0.1454      8.000 4.541     12.51   0.13
```

Por fim os ajustes do modelo são comparados com ajustes em um MQO, de forma a buscar comparações com um modelo mais tradicional que modela somente a média, e não considera assimetria. A comparação é feita a seguir:

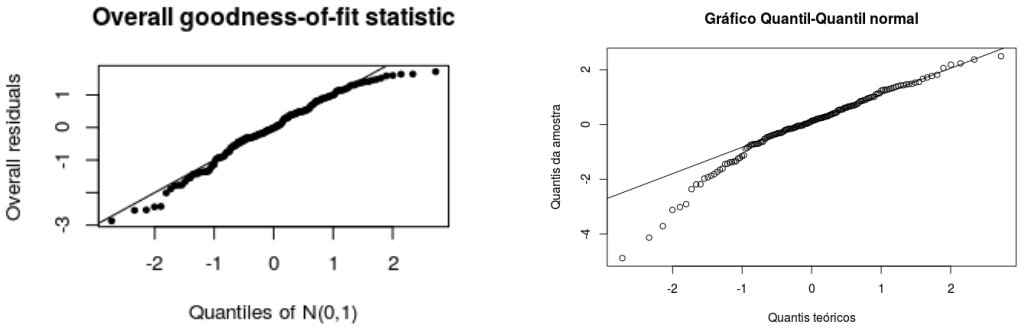


Figura 6: Comparação de ajustes: regressão log simétrica vs. mínimos quadrados ordinários

Conclusão:

Dados os modelos apresentados e a comparação com os ajustes de Mínimos quadrados ordinários evidenciou-se a importância da modelagem da assimetria, de forma a obter um modelo de regressão mais completo e robusto. Os resultados sugerem que o modelo ajusta os dados gerando estatísticas significantes para a modelagem da mediana e assimetria considerando a variável resposta como arrecadação.

Além disso, com erros mais ajustados, flexibilidade estatística na escolha das famílias dentro de um só modelo, e possibilidade de trabalhar uma distribuição que não seja necessariamente normal, o modelo se torna mais amplo de forma a atender demandas de diferentes áreas de estudo, bem como diferentes famílias de distribuições. Ainda que a indústria de cinema possua estrutura financeira mais complexa que a apresentada, o site IMDb.com possui completa e relevante informação sobre produção de filmes, e bilheterias.

Referências:

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: *Breakthroughs in statistics*. [S.l.]: Springer, 1992. p. 610–624.

EPSTEIN, E. J. *The Hollywood Economist 2.0: The Hidden Financial Reality Behind the Movies*. Reprint. Melville House, 2012. ISBN 1612190502,9781612190501. Disponível em:<http://gen.lib.rus.ec/book/index.php?md5=a6fb4aa49e4d0e3101401624eb56878c>.

SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. .

VANEGAS, L. H.; PAULA, G. A. A semiparametric approach for joint modeling of median and skewness. *Test*, Springer, v. 24, n. 1, p. 110–135, 2015.

VANEGAS, L. H.; PAULA, G. A. An extension of log-symmetric regression models: R codes and applications. *Journal of Statistical Computation and Simulation*, v. 86, n. 9, p. 1709–1735, 2016. Disponível em:<http://dx.doi.org/10.1080/00949655.2015.1081689>.

VANEGAS, L. H.; PAULA, G. A. et al. Log-symmetric distributions: statistical properties and parameter estimation. *Brazilian Journal of Probability and Statistics*, Brazilian Statistical Association, v. 30, n. 2, p. 196–220, 2016.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 32 - ANÁLISE DE DADOS PLUVIOMÉTRICOS NO MUNICÍPIO DE JOINVILLE COM USO DOS PACOTES HYFO E HYDROTSM

Natassia Cardoso Bilésimo[26]
Elisa Henning[27]
Edgar Odebrecht[28]
Andrea Cristina Konrath[29]

Resumo

O tratamento adequado dos dados pluviométricos é importante durante a identificação dos fenômenos meteorológicos típicos em uma região, pois torna possível o reconhecimento de padrões que podem auxiliar na sua gestão de recursos hídricos. Neste sentido, esse trabalho tem como objetivo realizar o preenchimento de dados faltantes e a análise exploratória dos dados pluviométricos de Joinville, do período entre 2012 e 2016, com base registros diários de dez estações meteorológicas da Defesa Civil do município. Para isso, foi utilizado uma função existente no pacote Hyfo, pertencente ao software R. Esta função se baseia no cálculo do coeficiente de correlação de cada par de estações. As estatísticas descritivas foram calculadas através de funções contidas no pacote HydroTSM. Por meio dos resultados parciais dessas estatísticas sobre a série de dados, por exemplo o valor mínimo, a mediana, a média, e o desvio padrão, foi possível identificar alguns problemas relacionados à coleta de dados e suas prováveis causas.

Palavras-Chave: Dados pluviométricos, preenchimento de dados faltantes, análise exploratória, pacotes estatísticos.

Abstract

The proper treatment of data on rainfall is important during identification of meteorological phenomenons commom to a region, since it becomes possible to recognize patterns that could assist on water resource management. On such account, this work has as an objective to supply data where absent and have an exploratory analysis of rainfall data in Joinville, between the period of 2012 and 2016, with basis in daily records of ten meteorological stations of the city´s Civil Defense. For that, an existing function of the Hyfo package was implemented, which belongs to the R software. This function is based on a calculation of the correlation coefficient of each pair of station. The descriptive statistics were calculated by the functions presented within the HydroTSM package, and through the partial results of these statistics from the series of data, for example minimum value, medium, average, maximum value and standard deviation, it became possible to identify problems related to the gathering of data and probable causes.

Keywords: Rainfall data, missing values, exploratory analysis, statistical packets.

[26] Universidade do Estado de Santa Catarina natassiabilesimo@gmaill.com
[27] Universidade do Estado de Santa Catarina elisa.henning@udesc.br
[28] Universidade do Estado de Santa Catarina edgar.odebrecht@udesc.br
[29] Universidade Federal de Santa Catarina andrea.ck@ufsc.br

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

Estudos sobre o comportamento pluviométrico de uma região podem auxiliar na definição dos níveis de observação, alerta, emergência e calamidade pública. Logo, também podem embasar um plano a ser utilizado pela Defesa Civil ou outro órgão gestor, com o objetivo de minimizar os impactos, como deslizamentos, causados pelos desastres naturais. Neste sentido é importante analisar a variabilidade espacial e temporal de atributos naturais, tendo destaque a precipitação pluviométrica (SILVA; GUIMARÃES, TAVARES, 2003).

A relação entre precipitação pluviométrica e ocorrência de deslizamentos tem sido estudada por diversos autores (2015). Especificamente para o município de Joinville, Bilesimo (2016) verificou a relação existente entre as chuvas e os deslizamentos em Joinville para o período de janeiro de 2012 a agosto de 2014. Nesse estudo, foi percebida a influência de fenômenos meteorológicos como a Zona de Convergência do Atlântico Sul e os ciclones, na formação das áreas de instabilidades na região. Assim sendo, torna-se importante identificar os fenômenos meteorológicos típicos e a altura pluviométrica característica e com isso identificar possíveis padrões que possam auxiliar na gestão de recursos hídricos da região de Joinville. Considerando a importância de estudar esses fenômenos, é fundamental tratar adequadamente os dados pluviométricos.

A obtenção de informações dos dados resultantes de uma pesquisa não é uma tarefa fácil. Desta forma, qualquer análise estatística inicia com uma análise exploratória de dados, no sentido em que o pesquisador possa se familiarizar com as observações. Esta análise gera as primeiras informações sobre os dados sem levar em consideração as suposições de algum modelo probabilístico. Geralmente a analise inicia com a aplicação de técnicas de Estatística Descritiva, já possibilitando ao pesquisador, muitas vezes, a obtenção de algumas respostas.

Um problema comum em qualquer pesquisa científica é a ocorrência de dados faltantes. Para resolver esse problema surgiram as técnicas de imputação de dados faltantes (NUNES; KLÜCK; FACHEL, 2009). Estas técnicas têm por objetivo completar os dados faltantes e permitir a análise com todos os indivíduos, observações ou períodos de um estudo. Uma das técnicas de imputação mais comumente utilizada é a regressão linear simples. Esta técnica permite preencher os dados faltantes por meio do que se chama de imputação única, ou seja, o dado ausente é preenchido

uma única vez e então se utiliza o banco de dados completo nas análises. (NUNES; KLÜCK; FACHEL, 2009).

Objetivo

Este trabalho tem como objetivo realizar o preenchimento de dados faltantes e análise exploratória dos dados pluviométricos de Joinville com o software R (R CORE TEAM, 2016).

Material e Métodos

A cidade de Joinville está localizada na região norte do Estado de Santa Catarina, com área de 1125,70 m² e altitude de 4,5 metros do nível do mar. Com o clima de tipo úmido a superúmido, a temperatura média anual é de 22,63°C (IPPUJ, 2014).

Para a análise exploratória dos registros pluviométricos foram utilizados os registros diários das estações meteorológicas da Defesa Civil de Joinville no período de 2012 a 2016. Foram utilizados dados de dez estações: Águas de Joinville, Cubatão, Estrada Sul, FlotFlux, Guanabara, Iate Club, Itaum, Jativoca Paraíso e Unidade de Obras, sendo que suas respectivas localizações são mostradas na Figura 1.

Foi realizado o preenchimento dos dados faltantes a partir da função "*fillgap*", existente para tal finalidade, do pacote Hyfo (XU, 2016). Esta função se baseia no cálculo do coeficiente de correlação de cada par de estações. A partir da estação com maior coeficiente de correlação, o preenchimento é realizado por meio de regressão linear simples (HIRSCH *et al*., 1993).

As estatísticas descritivas foram calculadas com auxílio do pacote HydroTSM (ZAMBRANO-BIGIARINI, 2014).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figura 1 – Localização das estações meteorológicas.

Fonte: Os autores

Resultados e Discussão

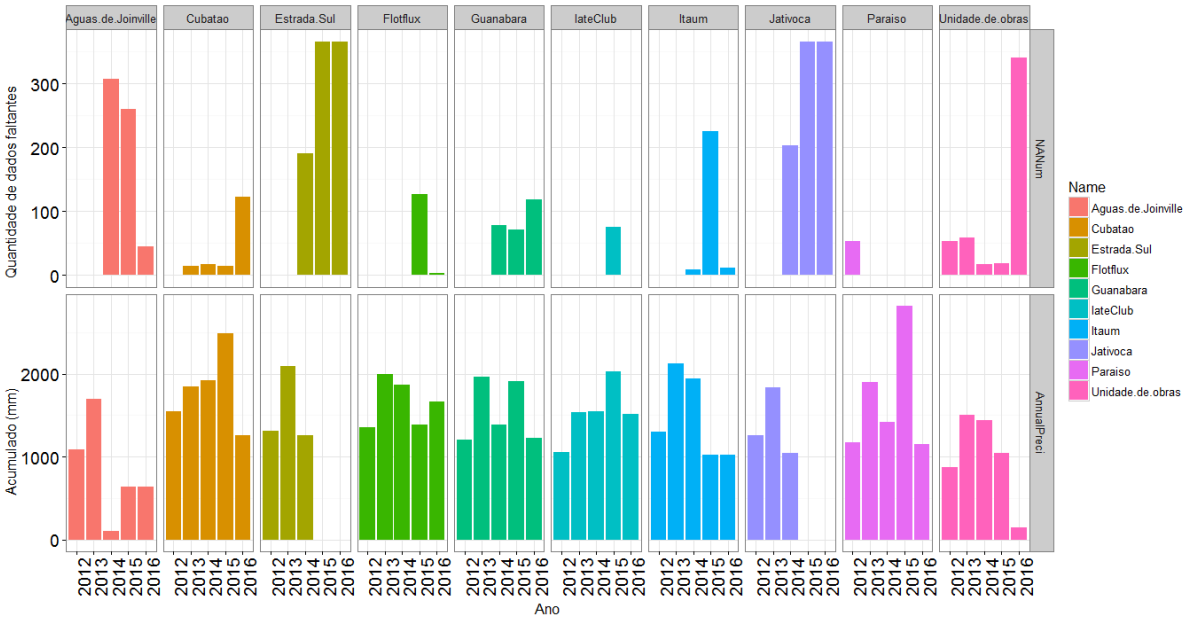
Na Figura 2 é possível visualizar os gráficos obtidos para cada estação meteorológica utilizando a função "getAnnual" presente no pacote Hyfo. Na parte superior são mostrados gráficos que indicam a quantidade de dados faltantes, e na parte inferior os valores para altura pluviométrica acumulada em cada uma das estações.

Pode-se verificar que as estações com mais dados faltantes são: Jativoca, Águas de Joinville e Estrada Sul. Por sua vez, a estação Paraíso possui dados faltantes apenas no seu primeiro ano de funcionamento. Já, a estação IateClub também apresentou poucos dados faltantes, sendo que eles ocorreram apenas em 2015.

Com relação ao acumulado anual de chuva, viu-se que a estação Paraíso e Cubatão apresentaram os maiores índices, uma vez que essas mesmas estações possuem poucos dados faltantes. No entanto, muitos dados faltantes foram verificados nas estações Jativoca e Estrada Sul, comprometendo a obtenção de valores para os acumulados anuais dessas duas estações, o que explica a inexistência de barras nos gráficos dos anos de 2015 e 2016.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

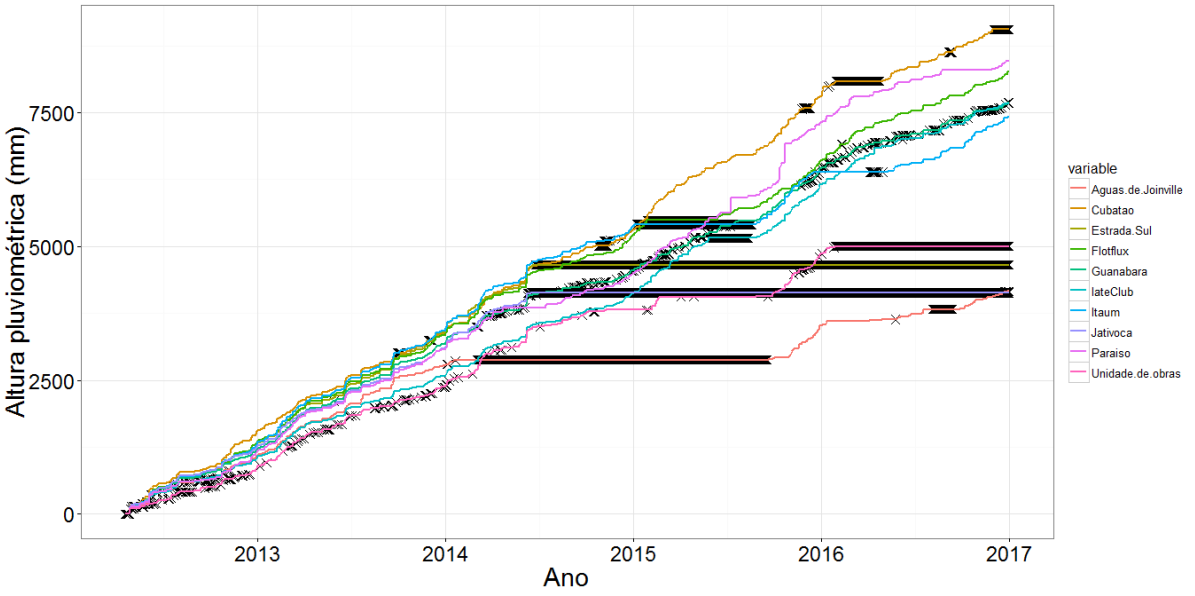
Figura 2 – Gráficos indicando a quantidade de dados faltantes e acumulados de chuva por ano em cada estação.



Fonte: Os autores

O acumulado de chuvas durante o período em estudo pode ser visualizado no gráfico da Figura 3. Os sinais “X” indicam um dado faltante, e se forem recorrentes por longos períodos são sinalizados pela linha em preto.

Figura 3 – Alturas pluviométricas acumuladas por cada estação ao longo do período.



Fonte: Os autores

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

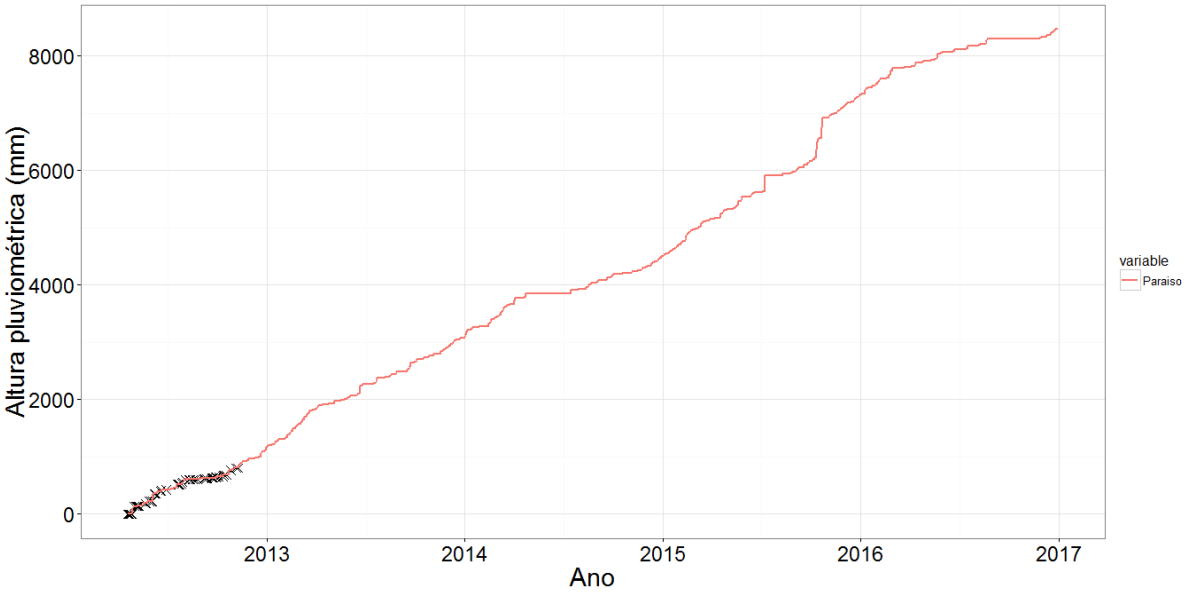
Com esse gráfico, pode-se perceber que o comportamento das linhas que representam os acumulados de chuva foram semelhantes em todas as estações até o ano de 2014. A partir de 2014 percebe-se um aumento da falta de registro de dados.

Buscando identificar as razões pelas quais as estações ficaram sem registros por longos períodos, entrevistas foram realizadas com o corpo técnico da Defesa Civil, por meio das quais pode-se perceber que o principal motivo dessas interrupções é falta de um programa de manutenção preventiva em toda a rede de monitoramento. Conforme foi relatado pela equipe, esta rede estaria completando cinco anos de funcionamento, e as peças para reposição em estoque estão em falta. A reposição de peças está vinculada a processos de licitação, que pelas suas características podem ser demorados. Outras causas, também levantadas foram o vandalismo, que fez com que algumas estações parassem de efetuar os registros por algum tempo, e os problemas com a estação repetidora, que por duas vezes foi derrubada devido a temporais ocorridos na região.

Para melhorar a visualização dos acumulados de chuvas, optou-se por separar os acumulados de cada estação. Dessa forma, foi possível identificar quando ocorreram chuvas intensas que fizeram com que o gráfico apresentasse um “salto nos valores” do eixo das ordenadas, ou seja, um aumento no valor da altura pluviométrica acumulada num curto espaço de tempo. O gráfico (Figura 4) refere-se a estação Paraíso. Percebe-se no gráfico a falta de dados no início do funcionamento dessa estação meteorológica.

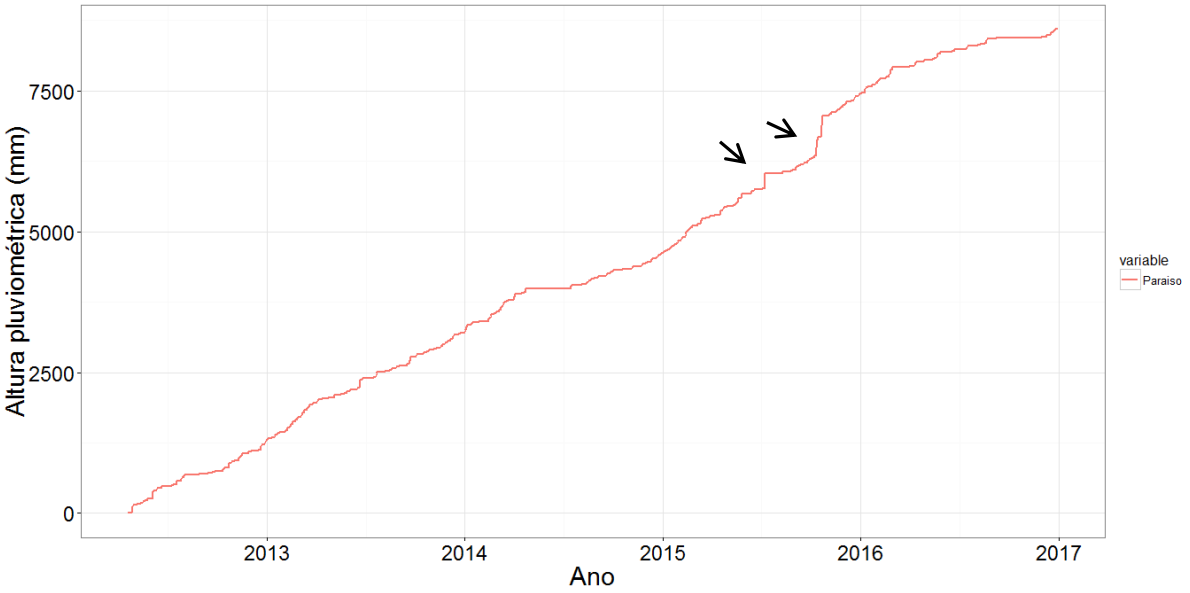
Os gráficos das alturas pluviométricas acumuladas durante o período analisado foram refeitos após o preenchimento. Na Figura 5 apresenta-se o gráfico da altura pluviométrica registrada pela estação Paraíso, após o preenchimento de dados faltantes, realizado com o auxílio do pacote Hyfo. Podem ser visualizados dois desníveis (saltos na direção vertical) próximos ao ano de 2016. Estudar estes desníveis e sua relação com deslizamentos (MENDES, 2015) é um dos objetivos na continuidade da pesquisa.

Figura 4 – Altura pluviométrica acumulada registrada pela estação Paraíso antes do preenchimento.



Fonte: Os autores.

Figura 5 – Altura pluviométrica acumulada registrada pela estação Paraíso após preenchimento.

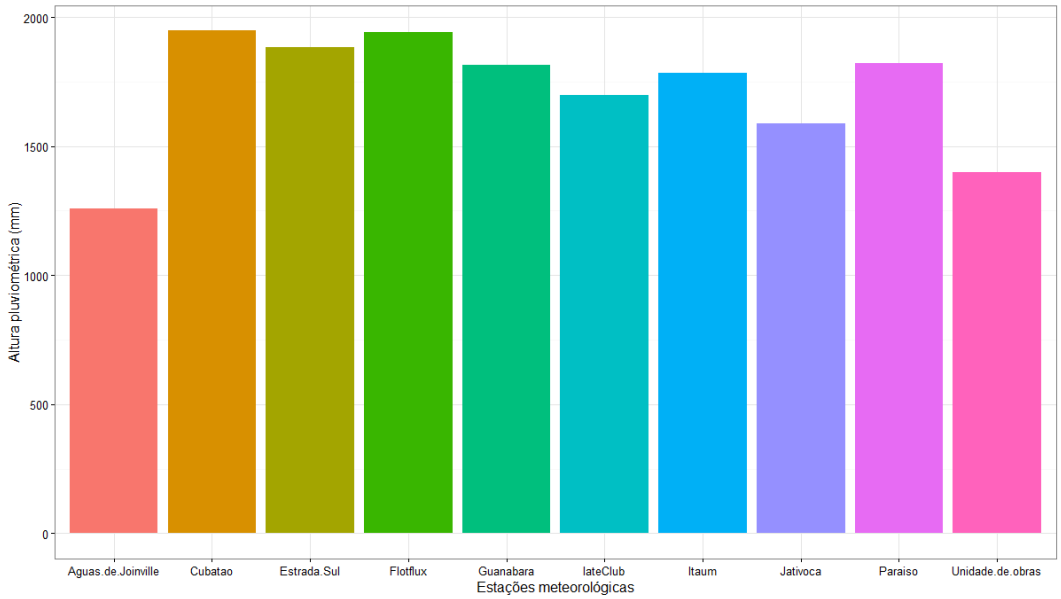


Fonte: Os autores

Na Figura 6 está um gráfico de barras gerado para a média anual de precipitação para cada estação após o preenchimento de dados. Na Tabela 1 estão as medidas descritivas obtidas com auxílio do pacote HydroTSM. A média diária de precipitação na maioria das estações durante o período analisado foi de

aproximadamente 5 mm, sendo que a estação Águas de Joinville (AJ) apresenta o menor valor numérico, 3,6mm. As estações Estrada Sul e Unidade de Obras foram abreviadas na tabela para ES e UO, respectivamente. As maiores médias obtidas foram nas estações Paraíso, Itaum e Estrada Sul, com valores superiores a 2000 mm, e a menor média, na estação IateClub.

Figura 6 – Média anual de precipitação para o período após o preenchimento.



Fonte: Os autores

Tabela 1 – Resumo numérico dos dados pluviométricos após o preenchimento

	Mínimo	1ºquartil	Mediana	Média	3ºquartil	Máximo	Desvio Padrão
Paraíso	0	0	0,23	5,25	4,35	274,8	14,35
Cubatão	0	0	0,46	5,52	5,41	129,7	11,92
Jativoca	0	0	0,23	4,42	3,66	126,8	10,26
Guanabara	0	0	0,21	4,99	4,34	120,6	11,33
Itaum	0	0	0,00	4,88	3,97	123,0	11,46
Flotflux	0	0	0,23	5,40	4,80	125,2	11,72
IateClub	0	0	0,23	4,64	4,34	97,2	10,19
AJ	0	0	0,15	3,60	2,87	103,3	8,59
ES	0	0	0,23	5,24	5,09	120,3	13,56
UO	0	0	0,00	3,94	2,51	119,4	9,74

Fonte: Os autores.

Com esses resultados, pode-se verificar que o valor máximo (274,8) de altura pluviométrica observado na estação Paraíso é numericamente superior aos valores das demais estações. Ressalta-se que esta estação é que teve menor número de

dados preenchidos, ou seja, mais observações verdadeiras. Por outro lado, a estação Unidade de Obras, que teve muitos valores preenchidos apresenta valores numericamente mais baixos para todas as medidas descritivas calculadas.

Com relação à forma da distribuição, verifica-se que indicam assimetria positiva, pois todas as estações possuem valores de média bem superiores à mediana. A variabilidade também é alta, como pode ser observado pelos valores dos desvios padrão.

Os histogramas e boxplots, que confirmam a assimetria positiva, por razões de espaço, não são mostrados neste documento. A continuidade dos trabalhos prevê a aplicação do Método da Curva de Dupla Massa para verificar a consistência dos dados após o preenchimento (GOMES; MONTENEGRO; VALENÇA, 2010).

O Método da Curva de Dupla Massa pode ser utilizado em séries mensais e anuais, e consiste em acumular os valores de precipitação mensal ou anual da estação em estudo e de uma outra estação, localizada na mesma região e que possua informações confiáveis. Esses valores acumulados são plotados em um gráfico cartesiano, sendo que no eixo das ordenadas estão os valores do posto em estudo, e no eixo das abcissas, os valores do posto que serve como base para a comparação (AGÊNCIA NACIONAL DE ÁGUAS, 2012).

Conclusão

Este trabalho teve como objetivo principal o preenchimento dos dados faltantes e análise dos dados pluviométricos das estações meteorológicas da Defesa Civil de Joinville, no período de 19 de abril de 2012 a 11 de setembro de 2016. Foram aplicados os pacotes Hyfo e HydroTSM. Os pacotes Hyfo e HydroTSM do software R se mostraram alternativas viáveis para a análise de dados pluviométricos. A grande quantidade de dados faltantes sinaliza a necessidade de ações que precisam ser tomadas pela gestão pública no sentido de melhorar o funcionamento da rede de monitoramento existente. A falta de registros pode dificultar a realização de análises mais precisas, os quais são ferramentas indispensáveis para a gestão dos recursos hídricos disponíveis na região.

Referências

AGÊNCIA NACIONAL DE ÁGUAS. Orientações para consistência de dados pluviométricos. Brasília, 2012. Disponível em:

http://arquivos.ana.gov.br/infohidrologicas/cadastro/OrientacoesParaConsistenciaDadosPluviometricos-VersaoJul12.pdf . Acesso em: 18 abr. 2017.

BILESIMO, N. C.. Análise da relação entre chuvas e os deslizamentos em Joinville, de janeiro de 2012 a agosto de 2014. 2015. 132f. Monografia (Graduação em Engenharia Civil). Universidade do Estado de Santa Catarina. Joinville, 2015.

FUNDAÇÃO INSTITUTO DE PESQUISA E PLANEJAMENTO PARA O DESENVOLVIMENTO SUSTENTÁVEL DE JOINVILLE (JOINVILLE). Joinville Cidade em Dados 2014. Joinville: Prefeitura Municipal, 2014.

GOMES, L. F. C.; MONTENEGRO, S. M. G. L.; VALENÇA, M. J. S. Modelo baseado na técnica de redes neurais para previsão de vazões na bacia do Rio São Francisco. **Revista Brasileira de Recursos Hídricos**, v. 15, n. 1, p. 05-15, 2010.

HIRSCH, R. M., *et al*. Statistical analysis of hydrologic data. **In**: MAIDMENT, D.R. (Ed.), **Handbook of Hydrology**. New York: McGraw-Hill. p. 17.1–17.55, 1993.

MENDES, R. M. *et al*. Estudo de limiares críticos de chuva deflagradores de deslizamentos no município de São José dos Campos/SP (Brasil). **Territorium**, n. 22, p. 119-129, 2015.

NUNES, L. N. *et al*.. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos . **Caderno de Saúde Pública**, Rio de Janeiro, 25(2):268-278, fev, 2009.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Disponível em: https://www.R-project.org/

SILVA, J. W.; GUIMARÃES, E. C.; TAVARES, M. Variabilidade temporal da precipitação mensal e anual na estação climatológica de Uberaba-MG. **Ciênc. agrotec**., v. 27, p. 665-674, 2003.

XU, Y. (2016). hyfo: Hydrology and Climate Forecasting. Rpackage version 1.3.6. Disponível em: https://CRAN.R-project.org/package=hyfo. Acesso em 09/01/2017.

ZAMBRANO-BIGIARINI, M. 2014. hydroTSM: Time series management,analysis and interpolation for hydrological modelling. R package version 0.4-2-1. Disponível em: https://CRAN.R-project.org/package=hydroTSM. Acesso em 09/01/2017.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 33 - USO DO R PARA COMPARAÇÃO DE ARQUIVOS CLIMÁTICOS: UMA ANÁLISE DA APLICAÇÃO DO ARQUIVO CLIMÁTICO DE ITAPOÁ NA CIDADE DE JOINVILLE

Rodrigo Jensen Cechinel[30]
Elisa Henning[31]
Ana Mirthes Hackenberg[32]

Resumo

As simulações termo-energéticas em edificações preveem o uso de dados climáticos da cidade onde o projeto está inserido. Porém, na ausência destes, recomenda utilizar o arquivo climático de uma cidade próxima, desde que pertencente à mesma zona bioclimática e com latitudes semelhantes. O objetivo principal deste trabalho é comparar os dados de temperaturas da cidade de Joinville com os da cidade de Itapoá. Espera-se que a partir dos resultados parciais seja possível contribuir para avaliar de forma efetiva a confiabilidade da aplicação dos dados climáticos de Itapoá para simulações termo-energéticas relativas à cidade de Joinville. Foram comparados os dados climáticos do arquivo EPW de Itapoá que é oriundo de dados do INMET (Instituto Nacional de Meteorologia) registrado por RORIZ (2010) aos dados fornecidos pela estação meteorológica da EPAGRI em Joinville (ISOPPO, 2016). A variável investigada neste trabalho é a temperatura média de bulbo seco. Nesta pesquisa serão aplicadas duas abordagens distintas. Inicialmente será realizada uma análise comparativa a partir da análise visual de gráficos e medidas descritivas. Em seguida deseja-se verificar se as duas são geradas pelo mesmo processo estocástico. Os resultados parciais apontam para semelhanças entre as duas séries.

Palavras-Chave: Arquivo climático; simulação termo-energética de edificações; R; séries temporais.

Abstract

The thermo-energetic simulations in buildings predict the use of climatic data of the city where the project is inserted. However, in the absence of these, it recommends to use the climatic archive of a nearby city, since it belongs to the same bioclimatic zone and with similar latitudes. The main objective of this work is to compare the temperature data of the city of Joinville with those of the city of Itapoá. It is expected that from the partial results it will be possible to contribute to an effective evaluation of the reliability of the application of the climate data from Itapoá to thermo-energetic simulations related to the city of Joinville. The climatic data of the EPW file from Itapoá, derived from INMET (National Meteorological Institute) data recorded by RORIZ (2010), were compared to the data provided by the EPAGRI meteorological station in Joinville (ISOPPO, 2016). The variable investigated in this work is the average dry bulb temperature. In this research, two different approaches will be applied. Initially, a comparative analysis will be performed based on visual analysis of graphs and descriptive measures. Then we want to verify if the two are **generated** by the same stochastic process. The partial results point to similarities between the two series.

[30]Universidade do Estado de Santa Catarina /macrorodrigo@gmail.com
[31]Universidade do Estado de Santa Catarina / elisa.henning@udesc.br
[32] Universidade do Estado de Santa Catarina / ana.hackenberg@udesc.br

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Keywords: Climatic file; Thermo-energy simulation of buildings; R; Time series.

Introdução

Há uma tendência crescente em se adequar os projetos das edificações da melhor forma possível, de modo a atender às necessidades de seus ocupantes aos fatores ambientais e externos do local considerando as características climáticas. Sob esta ótica, o conforto térmico e a eficiência energética se configuram como fatores importantes nas edificações.

A crise energética, em especial a do petróleo na década de 70, forçou o mundo a buscar fontes alternativas de energia e a otimização do seu uso. Foram criados os primeiros programas para avaliação de desempenho de térmico em edificações buscando eficiência energética em seu uso (MENDES *et al*., 2005). Estes programas de simulação de edificações auxiliam a identificar as melhores práticas construtivas, modelando e prevendo o comportamento de cada tomada de decisão.

O Procel Edifica é a principal certificação em edificações, é de caráter opcional no Brasil e apresenta uma lista de simuladores para os mais diversos fins, incluindo o programa Energyplus (PROCEL EDIFICA, 2006). Dentre os programas existentes, o Energyplus, é um programa para simulação energética de edificações que permite engenheiros, arquitetos e pesquisadores desenvolverem estudos em consumo de energia, aquecimento, refrigeração, ventilação, iluminação e uso de água (DOE, 2016). A precisão destes resultados está intimamente ligada ao tipo, precisão e confiabilidade do arquivo climático utilizado no processo.

No que tange aos aspectos legais, a NBR 15.575 (ABNT, 2013) destaca o uso do programa Energyplus que, aliado a dados climáticos da cidade onde o projeto está inserido, considera dias típicos de verão e inverno para definir as características do meio onde a edificação está incluída. Estes arquivos climáticos devem estar acreditados por instituição de reconhecida capacidade técnica e disponíveis para consulta, assim como suas respectivas fontes de dados.

Arquivos climáticos são dados obtidos em estações climáticas que devem ser tratados de modo a fornecerem um resultado adequado para uso em programa de simulação. Eles podem fornecer informações mensais, medias diárias ou leituras horárias. Os melhores resultados de simulação consideram arquivos climáticos que possuam 8760 horas de informações completas num ano típico de 365 dias. Com a

instrumentação de monitoramento adequada, apresentam informações para as seguintes condicionantes climatológicas: temperatura de bulbo seco, temperatura de bulbo úmido, velocidade de vento, direção do vento, radiação solar direta, radiação solar direta e difusa, cobertura de nuvens, pressão atmosférica e temperatura do solo.

A elaboração de arquivos climáticos, no entanto, pode ser um grande desafio no Brasil. Apesar da disseminação de estações meteorológicas de superfície nas últimas décadas, a grande maioria das cidades brasileiras ainda não é contemplada com dados climáticos em um formato compatível aos programas de simulação (GRIGOLETTI; FLORES; SANTOS, 2016). Os poucos arquivos existentes são uma lacuna para a simulação de projetos de edificações. Sua base de dados é obtida em estações meteorológicas que nem sempre possuem dados completos ou uma série histórica adequada. A prática, apoiada pela legislação vigente, tem consolidado o uso de arquivos climáticos de outras cidades para o caso da sua inexistência.

A NBR 15.575 (ABNT, 2013) é a norma de desempenho para edificações com fins habitacionais e indica que em simulações computacionais devam ser usados dados climáticos da cidade onde o projeto está inserido. Porém, na ausência destes, recomenda utilizar o arquivo climático de uma cidade próxima, desde que pertencente à mesma zona bioclimática e com latitudes semelhantes.

A cidade catarinense de Joinville não possui arquivo climático para uso no programa Energyplus. Apesar de haver estações meteorológicas na região operadas pela Empresa Brasileira de Infraestrutura Aeroportuária (INFRAERO), pela Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI) e pela Universidade de Joinville (UNIVILLE), não há uma que possua uma série histórica longa (recomendável superior a dez anos), com dados horários completos e que estejam acessíveis à comunidade. Destacam-se desta maneira as limitações para o desenvolvimento de um arquivo climático local, assim como da maioria das cidades brasileiras. Desta forma, para trabalhos de simulação em Joinville, deve ser adotado o arquivo climático de maior proximidade, no caso o da cidade de Itapoá (HENNING et al., 2016; CECHINEL; HACKENBERG; TONDO, 2016; TONDO; CECHINEL; HACKENBERG, 2016). Assim, uma dúvida motivou a proposição deste trabalho: os dois arquivos climáticos, de Joinville e Itapoá, apresentam realmente características similares?

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Objetivo

O objetivo principal deste trabalho é comparar os dados de temperaturas da cidade de Joinville com os da cidade de Itapoá. É uma primeira proposta de comparação dos dois arquivos. Espera-se que a partir dos resultados parciais seja possível contribuir para avaliar de forma efetiva a confiabilidade da aplicação dos dados climáticos de Itapoá para simulações termo-energéticas relativas à cidade de Joinville.

Material e Métodos

Nesta pesquisa foram comparados os dados climáticos do arquivo EPW de Itapoá que é oriundo de dados do INMET (Instituto Nacional de Meteorologia) registrado por RORIZ (2010) aos dados fornecidos pela estação meteorológica da EPAGRI em Joinville (ISOPPO, 2016). A variável investigada neste trabalho é a temperatura média de bulbo seco.

Nesta pesquisa serão aplicadas duas abordagens distintas. Inicialmente será realizada uma análise comparativa a partir da análise visual de gráficos e medidas descritivas. Este enfoque foi aplicado na comparação de diferentes tipos de arquivos climáticos, por exemplo, os formatos TRY e SWERA (SHELLER *et al*, 2015; GUIMARÃES; CARLO, 2015). Também foi utilizado na comparação de variáveis climáticas com objetivo de aperfeiçoar os Zoneamentos Bioclimáticos específicos para diferentes tipologias construtivas (RORIZ, 2013). INVIADATA *et al*. (2016) ainda aplicaram medidas gráficas para analisar o agrupamento de 411 cidades brasileiras com base no desempenho térmico de edificações residenciais.

Na segunda metodologia deseja-se verificar se as duas são geradas pelo mesmo processo estocástico. Silva, Ferreira e Sáfadi (2000) propuseram uma alternativa de comparação de séries temporais que consiste em avaliar a diferença entre as duas séries analisadas. De acordo com os autores, para comprovar que as duas séries são originárias de um mesmo processo estocástico é necessário verificar a inexistência de tendência, sazonalidade e confirmar que os resíduos do modelo ajustado são um ruído branco. Para avaliar a tendência foi aplicado o teste Cox-Stuart (COSTA; SÁFADI, 2010); para a sazonalidade o teste Canova-Hansen (HYNDMAN, 2016) e por meio do teste Box-Pierce (COSTA, SÁFADI, 2010) verificou-se se os resíduos se comportavam como um ruído branco. Todas as análises foram realizadas

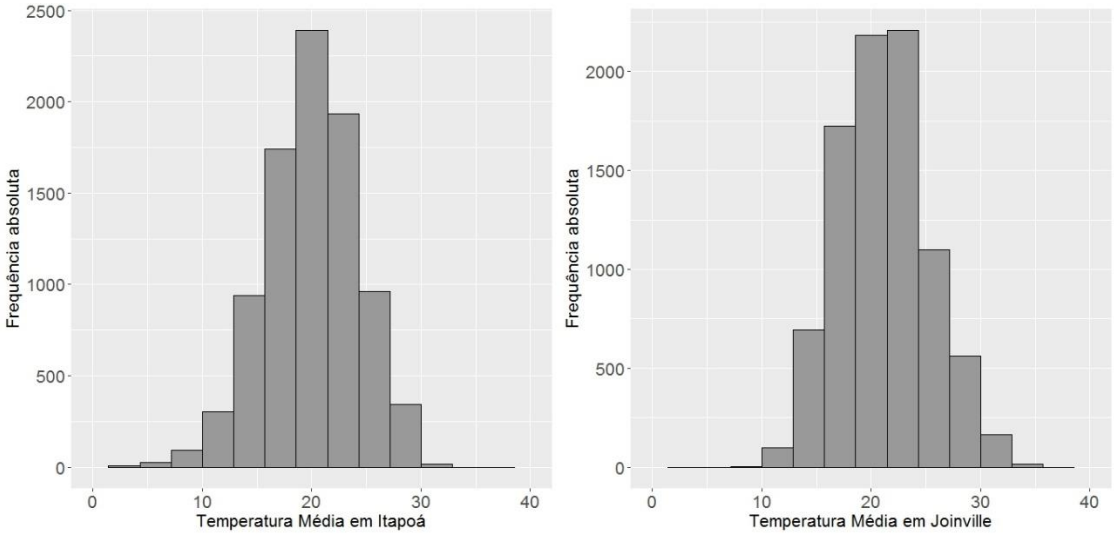
Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

no ambiente R (R CORE TEAM, 2016), com o auxilio dos pacotes ggplot2 (WICKHAM, 2009), randtests (CAEIRO; MATEUS, 2014) e forecast (HYNDMAN, 2016).

Resultados e Discussão

Resultados parciais apontam para a similaridade entre as duas séries de dados de temperatura de bulbo seco. A partir da análise visual dos histogramas (Figura 1), pode-se verificar que existem possivelmente padrões similares na forma da distribuição dos dados.

Figura1 – Histogramas das temperaturas médias de Itapoá e Joinville



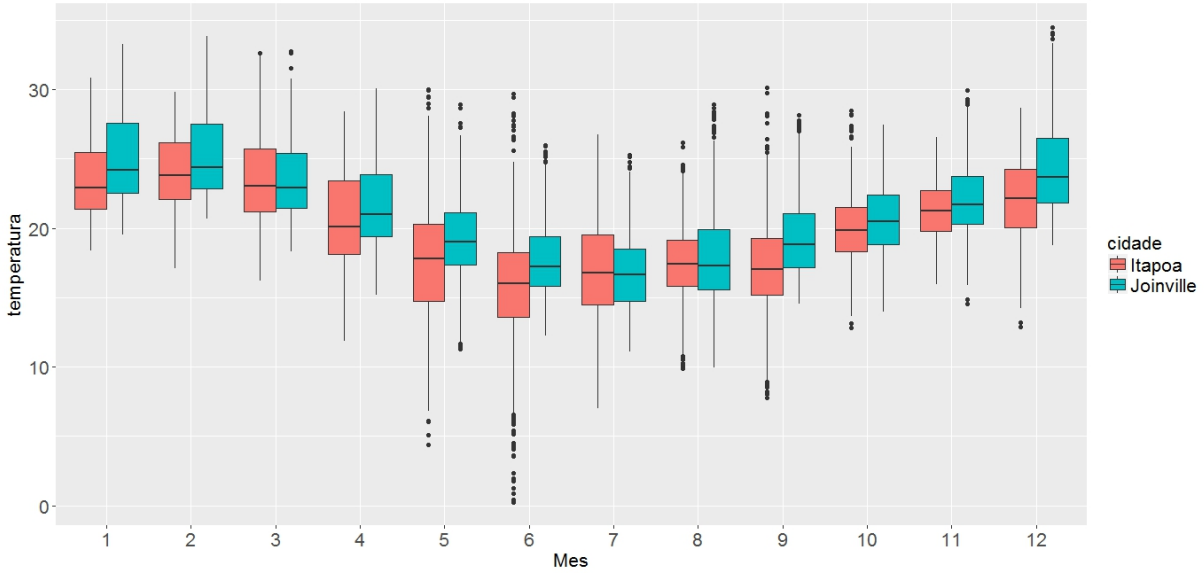
Fonte: os autores

As temperaturas médias das duas cidades apresentam distribuição aparentemente simétrica com indicação de leve assimetria negativa para a cidade de Itapoá, possivelmente apontando para temperaturas menores em alguns dias nesta cidade. As temperaturas médias das duas cidades ficam próximas de 20°C, com média 19,9°C (desvio padrão 4,2°C) em Itapoá e média 21,1°C (desvio padrão 4,1°C) em Joinville.

Foram construídos boxplots das temperaturas, por mês, para avaliar o padrão de comportamento mensal (Figura 2). Verifica-se um padrão semelhante ao longo do período analisado, todavia com indicação de temperaturas um pouco mais baixas em Itapoá. Pode-se atribuir como uma das razões o fato de Itapoá ser um município litorâneo com praia e assim têm maior incidência de ventos.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figuras 2 – Boxplots das temperaturas médias de Itapoá e Joinville



Fonte: Os autores

Do mesmo modo, por meio da segunda abordagem, de acordo com os resultados, as duas séries são originárias de um mesmo processo estocástico. A série composta pela diferença não apresenta tendência (p-valor = 0,12226) e nem sazonalidade. Os resíduos também se comportam como um ruído branco (p-valor = 0,9418).

Estes resultados são preliminares e ressalta-se que é essencial avaliar as demais variáveis climáticas, temperaturas máximas e mínimas, direção do vento e umidade relativa, entre outras, necessárias para as simulações.

A proposta de metodologia, com as duas abordagens, se mostra uma alternativa viável à aplicação desejada. Para complementar e dar mais confiabilidade aos resultados, a aplicação de medidas de similaridade e dissimilaridade são opções para a continuidade dos trabalhos.

Conclusão

Neste trabalho foram comparadas as séries de temperaturas médias das cidades de Itapoá e Joinville, com objetivo de avaliar se estas apresentam comportamentos similares. Esta análise tem como motivação a aplicação do arquivo climático de Itapoá em simulações computacionais de edificações para o município de Joinville, que não possui arquivo climático. Duas abordagens foram aplicadas e

resultados parciais apontam para semelhanças entre as duas séries. Ressalta-se são resultados preliminares e é fundamental avaliar as outras variáveis climáticas necessárias. A aplicação de outras técnicas de comparação pode contribuir para maior embasamento e efetividade dos resultados.

Referências

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). NBR 15.220-3: Desempenho térmico de edificações. Rio de Janeiro, 2005.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). NBR 15.575-1: Edificações habitacionais - Desempenho. Rio de Janeiro, 2013.

CAEIRO, F.; MATEUS, A. 2014. randtests: Testingrandomness in R. R Package 1.0. Disponível em: <https://CRAN.R-project.org/package=randtests>. Acesso em 06 jan. 2017.

CECHINEL, Rodrigo J.; HACKENBERG, Ana M.; TONDO, Gabriela H. Desempenho térmico em habitações de interesse social inseridas na cidade de Joinville e recomendações para melhoria dos parâmetros mínimos construtivos. **In:** Encontro Nacional de Tecnologia do Ambiente Construído (ENTAC), 16., 2016, São Paulo. Anais... Porto Alegre: ANTAC, 2016. Disponível em: <http://www.infohab.org.br/entac/2016/ENTAC2016_paper_122.pdf>. Acesso em: 17 jan. 2017.

COSTA, F. M.; SÁFADI, T. Comparação estatística de duas séries de material particulado (Mp10) na cidade de São Paulo. **Revista Brasileira de Biometria**, v. 28, n. 3, p. 23-38, 2010.

GRIGOLETTI, G. C; FLORES, M. G.; SANTOS, J. C. P. Tratamento de dados climáticos de Santa Maria, RS, para análise de desempenho térmico de edificações. **Ambiente Construído**, v. 16, n. 1, p. 123-141, 2016.

GUIMARÃES, I. B. B.; CARLO, J. C. Comparação estatística entre arquivos climáticos desenvolvido com métodos diferentes. **In**: Proceedings of EURO ELECS 2015. 2015, Guimarães. Proceedings...Guimarães: Universidade do Minho, 2015. p. 2303 – 2312.

HENNING, E.; CECHINEL, R. J.; TONDO, G. H.; HACKENBERG, A. M.; OLIVEIRA, T.A.C. Aplicação do DOE no conforto térmico em edificações. **Revista TMQ Techniques, Methodologies and Quality**. Vol. 7. Edição Especial: Técnicas Avançadas da Qualidade. 2016. Disponível em: <http://publicacoes.apq.pt/aplicacao-do-doe-conforto-termico/>. Acesso em 17 jan. 2017.

HYNDMAN, R. J. 2016. forecast: Forecasting functions for time series and linear models. R package version 7.3. Disponível em: <http://github.com/robjhyndman/forecast>. Acesso em 05 jan. 2017.

INVIDIATA, Andrea; MELO, Ana Paula; VERSAGE, Rogerio; SOUSA, Raquel Fernandes de; LAMBERTS, Roberto. Análise de agrupamento de 411 cidades brasileiras baseado em indicadores de desempenho de edificações residenciais naturalmente ventiladas. **In**: Encontro Nacional de Tecnologia do Ambiente Construído (ENTAC), 16., 2016, São Paulo. Anais... Porto Alegre: ANTAC, 2016. Disponível em: <http://www.infohab.org.br/entac/2016/ENTAC2016_paper_14.pdf>. Acesso em: 17 jan. 2017.

ISOPPO, G. Dados climatológicos da estação da EPAGRI em Joinville entre 2013 à 2015. [mensagem pessoal]. Mensagem recebida por macrorodrigo@gmail.com em 23/05/2016.

MENDES, N.; WESTPHAL, F. S.; LAMBERTS, R.; NETO, J. A. B. C. Uso de Instrumentos Computacionais Para Análise do Desempenho Térmico e Energético de Edificações no Brasil. **Ambiente Construído**. Porto Alegre, v.5, n.4, p.47-68, out./dez. 2005.

PROCEL EDIFICA, Simuladores. Florianópolis, SC, 2006. Disponível em: <http://www.procelinfo.com.br/main.asp?TeamID ={796B68CB-2559-401F-A481-DC3D145F572E>. Acesso em: 05 jan. 2017.

R CORE TEAM. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>. Acesso em 05 jan. 2017.

RORIZ, M. Arquivos climático em formato EPW. Florianópolis, SC, 2010. Disponível em: https://labeee.ufsc.br/downloads/arquivos-climaticos/formato-epw . Acesso em:11 jan. 2017.

RORIZ, M. Classificação de Climas do Brasil – Versão 2. **Associação Nacional de Tecnologia do Ambiente Construído (ANTAC).** São Carlos, SP. Novembro de 2013. Disponível em: < http://publicacoes.apq.pt/aplicacao-do-doe-conforto-termico/>. Acesso em: 17 jan. 2017.

SCHELLER, C. et al. Análise de Arquivos Climáticos Para a Simulação de Desempenho Energético de Edificações. Florianópolis: UFSC/Centro Brasileiro de Eficiência Energética em Edificações, 2015. Disponível em: <http://www.labeee.ufsc.br/node/635>. Acesso em 05 jan. 2017.

SILVA, R. B. V.; FERREIRA, D. F.; SÁFADI, T. Modelos de séries temporais aplicados á série dos índices de preços ao consumidor na região de Lavras, MG, no período de 1992 a 1999. Organizações Rurais & Agroindustriais, v. 2, n. 2, 2000.

TONDO, Gabriela Hanna; HACKENBERG, Ana Mirthes; CECHINEL, Rodrigo Jensen .Analysis of Thermal Comfort in Schools: Comparison between the prescriptive method and simulation. **International Conferenceon Passive and Low Energy Architecture (PLEA)**, 32., 2016, Los Angeles. Anais. Disponível em: <http://www.plea2016.org/download/PLEA%202016%20Volume%201.pdf>. Acesso em 17 jan. 2017.

WICKHAM,H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag. New York, 2009.

DOE. UNITED STATES DEPARTMENT OF ENERGY'S (DOE). Energyplus. Disponível em: <https://energyplus.net/>. Acesso em: 1 jun. 2016.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 35 - DEVELOPMENT OF A VIRTUAL ANALYSER THROUGH PARTIAL LEAST SQUARE REGRESSION AND VARIANCE INFLUENCE PROJECTION TO ESTIMATE THE CONTENT OF MAPD CONTAMINANTS IN A TRICKLE-BED REACTOR USING R

Ana Rosa Massa[33]
Vicente Braga Barbosa[34]
Karla Patrícia Silva de Oliveira Esquerre[35]
Adonias Magdiel Silva Ferreira[36]

Resumo

Analisadores em linha fornecem uma resposta rápida de composição em comparação às análises laboratoriais. Porém, esses estão sujeitos a frequentes interferências e contaminações, que comprometem o funcionamento do equipamento. Durantes estas, haverá perda de informações vitais, a não ser que haja uma técnica que permita estimar tais informações de maneira confiável. O presente trabalho tem por objetivo desenvolver um analisador virtual para estimar a concentração dos contaminantes metilacetileno e propadieno (MAPD) em um reator trickle-bed em uma indústria de propileno. Uma análise multivariada de dados de uma campanha catalítica, coletados por cromatográfos a gás e termopares, é utilizada no desenvolvimento, através do R, de um modelo de Regressão por Mínimos Quadrados Parciais (PLSR) para dois leitos em série. Os modelos são desenvolvidos a partir de 13 variáveis de entrada e 11445/13540 observações para cada leito. Os modelos de PPLSR demonstraram ótima capacidade de predição e excelente performance, com R^2 de 0.87 e 0.95. Uma técnica de seleção de variáveis conhecida como Projeção de Influência da Variância (VIP) é aplicada, e as variáveis mais importantes são selecionadas para consequente redução de dimensionalidade dos modelos. Após tal redução, ambos modelos preservaram suas capacidades de predição e notáveis performances.

Palavras-Chave: PLS, VIP, Analisadores Virtuais, teor de MAPD

Abstract

Online analysers grant a faster answer on the composition of products when compared with laboratory analysis. However, the former often requires frequent maintenance. During those, the loss of vital information could lead to a halt in production, unless another device allows for such information to be carefully estimated. This paper aims at developing a Virtual Analyser that can estimate the concentration of methylacetylene and propadiene (MAPD) contaminants in a trickle-bed reactor at a propene plant. A multivariate analysis of data from a catalytic campaign, collected by gas chromatographers and temperatures probes, is used to develop a Partial Least Square Regression (PLSR) model for two in series beds using R. The two models are developed with 13 predicting variables and 11445/13540 observations for each bed. The PLSR models have shown a great prediction capacity and a remarkable performance, with R^2 of 0.87 and 0.95. A variable selection technic called Variance Influence Projection (VIP) is applied and the most important variables are selected to

[33] Universidade Federal da Bahia (UFBA), anarosa.massa@hotmail.com
[34] Universidade Federal da Bahia (UFBA), vicentebragab@gmail.com
[35] Universidade Federal da Bahia (UFBA), karla.esquerre@gmail.com
[36] Universidade Federal da Bahia (UFBA), adoniasmagdiel@ufba.br

diminish the dimensionality of the models, which are still able to keep remarkable prediction capacities.

Keywords: PLS, VIP, Online Analyser, MAPD content

1. Introduction

The petrochemical industry is constantly innovating its methods and searching for processes that are safer and more efficient. A strong competition among different companies calls for complex technology that can guarantee both the product specification and a solid yield on the invested capital. In this context, online analysers grant a faster answer on the composition of products when compared with laboratory analysis. However, the former is often affected by the existing substances in the streamline, which harm and compromise its normal working state, requiring frequent maintenance. During those, the loss of vital information could lead to a halt in production, unless another device allows for such information to be carefully estimated.

During the production of propene in a petrochemical plant, the removal of contaminants happens in catalytic hydrogenation reactors, which usually contain two online analysers (gas chromatographers) responsible for determining the content of the inflow and outflow streams, in order to evaluate catalytic conversion and selectivity (Cohn, 2006).Other devices, such as temperature probes and flow meters, gather, together with the chromatographers, an immense amount of data about the process. One way to evaluate this data is through Partial Least Square Regression (PSLR), which has been widely used in the monitoring of industrial processes, since it usually provides empirical models with reasonable prediction capacity (Morellato, 2010).

As such, PLSR models can be used to estimate the content of contaminants in case the online analyser fails to provide such information, preventing a halt in production and a subsequent loss in capital. However, as mentioned by Galindo-Prieto*et al.* (2017), highly computerized analytical instrumentation can produce data sets with a large amount of predicting variables. Reducing the dimensionality of the model can lead to easier interpretations and more robust predictions. One method to achieve such reduction is the Variable Influence Projection (VIP), which has been widely used for selecting the important X-variables in a PLSR.

2. Objectives

This paper aims at developing a virtual analyser in order to estimate the concentration of methylacetylene and propadiene (MAPD) in a trickle-bed reactor at a propene plant through the multivariate analysis of data collected by an online analyser (gas chromatographer) and by temperature probes connected to the reactor. Furthermore, it also aims at using VIP in order to identify the most relevant process variables to the developed multivariate models, and analyse the performance of such models using only the selected variables.

3. Materials and Methods

3.1 Trickle-bed Reactor

Data was collected through a catalytic campaign for two in-series bed reactors, called bed A and bed B. Thirteen variables were monitored with a periodicity of 10 minutes. A delay of 3 to 7 minutes between inflow variables and outflow/analyser variables exist due to the reactor residence time and analyser cycle. Data was organized into two **X** matrices of 13 columns/predicting variables (fresh feed flow, recycle flow, fed hydrogen flow, combined feed temperature, pressure, temperature across six points in the reactor, outflow temperature and MAPD content in feed) and 11445/13540 rows/observations for bed A and B respectively. MAPD content in out stream for each bed were organized into **y** vectors of lengths 11445 and 13540.

3.2 Removal of Outliers through Principal Components Analysis (PCA)

Initial exploratory analysis of X-variables was done through PCA by selecting the components that together explain a minimum of 90% of captured variance. Q residuals (sum of squared residuals) and Hotelling's T^2 (weighted sum of squared scores) were used as a criterion to trace outliers. Observations with simultaneously high numeric values for both parameters are potential outliers (Ferreira *et al.*, 2015) and were hence removed from the dataset.

3.3 PLSR model

After outlier removal, data were divided into training (80%) and testing (20%) subsets. Rows of each subset were chosen randomly, and 10,000 different pairs of subsets were used to estimate the coefficients of PLSR models. The optimal value of latent variables for each model was determined through Venetian Blinds cross-

validation method (Souza, 2014). The model with highest Pearson Correlation Coefficient (R^2) for the testing subset was chosen as the PLSR for the bed.

3.4 Reduction of monitored process variables through VIP

VIP calculation provides the influence of each X-variable j in the data matrix X_{IJ} on the PLS model of a given response vector $\boldsymbol{y}_I$ as follows: (Tran*et al.*, 2015)

$$VIP_j = \sqrt{d \sum_{k=1}^{h} \boldsymbol{v}_k (\boldsymbol{w}_{kj})^2 / \sum_{k=1}^{h} \boldsymbol{v}_k} \quad (1)$$

in which, $\boldsymbol{v}_k = \boldsymbol{c}_k^2 \boldsymbol{t}_k' \boldsymbol{t}_k$ and $c_k = \frac{\boldsymbol{t}_k' \boldsymbol{y}(\boldsymbol{k})}{\boldsymbol{t}_k' \boldsymbol{t}_k}$; (2) and (3)

where d is the number of variables and h the number of latent variables in the PLS model. The proportion of the fraction of the explained variance of X_{IJ}, for each variable j over all latent variables is expressed by $\boldsymbol{v}_k$, which is weighted by the term $\boldsymbol{w}_{kj}$ that represents the covariance between X and y. The term c_k is calculated for each column of the PLS scores matrix T and for the predicted response $\boldsymbol{y}_I$.

As suggested by Tran*et al.* (2015), a threshold value for VIP equal to 1wasadopted as the criterion to define an important variable. After processes variables were selected, the same training and testing subsets chosen for best model in section 3.3 was used to develop a new PLSR with only the selected variables.

3.5 R Packages

All methodology described in previous sections were applied using software R. The packages used were *pls*[1], *plsdepot*[2],*mixOmics*[3] and *randtests*[4].

4. Results and Discussion

4.1 Exploratory Analysis and outlier detection in the dataset

[1] Bjørn-Helge Mevik, Ron Wehrens and Kristian HovdeLiland (2016). pls: Partial Least Squares and Principal Component Regression. R package version 2.6-0. https://CRAN.R-project.org/package=pls
[2] Gaston Sanchez (2012). plsdepot: Partial Least Squares (PLS) Data Analysis Methods. R package version 0.1.17. https://CRAN.R-project.org/package=plsdepot
[3] Kim-Anh Le Cao, Florian Rohart, Ignacio Gonzalez, Sebastien Dejean with key contributorsBenoit G autier, Francois Bartolo, (2016). mixOmics: Omics Data Integration Project. R package version 6.1.1. https://CRAN.R-project.org/package=mixOmics
[4] FredericoCaeiro and Ayana Mateus (2014). randtests: Testing randomness in R. R package version 1.0. https://CRAN.R-project.org/package=randtests

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Figure 01 shows the Influence Graphs (Q vs T^2) of the data sets before and after outlier removal. Around 5% of observations were defined as abnormal in bed A and 16% in bed B. Those were excluded in subsequent analysis. The relatively large proportion of outliers in bed B might be explained by a glitch in the chromatographer, which recorded the same MAPD content for several days. This paper, however, does not aim at analysing other probable reasons for the remaining proportion of outliers.

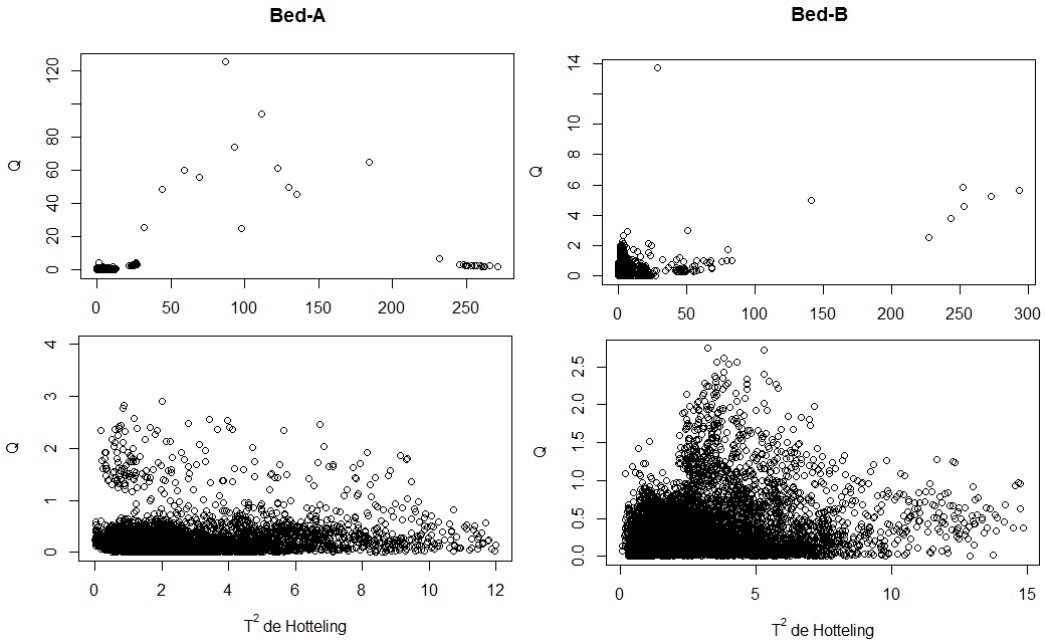


Figure 01: Influence Graphs before (top) and after (bottom) outlier removal

4.2 Predicting MAPD content in the out stream through PLSR

Figure 02 shows the R^2 for the predicted MAPD content in each of the 10,000 runs for bed A. The R^2 of the best PLSR model was 0.887 and it is highlighted. The Root Mean Square Error of Prediction (RMSEP) for the best model was 1131 ppm and its BIAS was -2.02×10^{-9} ppm. Such small value for the latter is expected given the large amount of observations. According the central limit theorem, as *n* tends to infinity, the residuals of a model tend to zero. Yet, such parameter is important to confirm the lack of a tendency in the model to over/underestimate MAPD content.

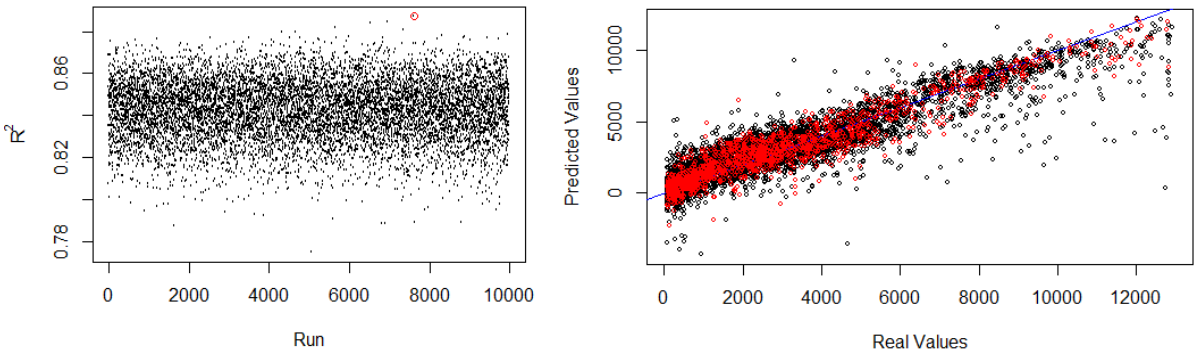


Figure 02: R^2 for PLSR models and Predicated *vs*. Real MAPD content for bed A

Figure 02 also displays the Real *vs*. Predicted values for MAPD content. Black dots are training points whether red dots are testing points. Most points fall around the blue line, which represents a perfect model.

Figure 03 shows the same information as Figure 02, but this time for bed B. The model performance for this bedwas much higher, with a R^2 of 0.950. RSMEP and BIAS were respectively 1339ppm and -8.16×10^{-10} ppm. The low value of the latter indicates once more the lack of significant systematic error.

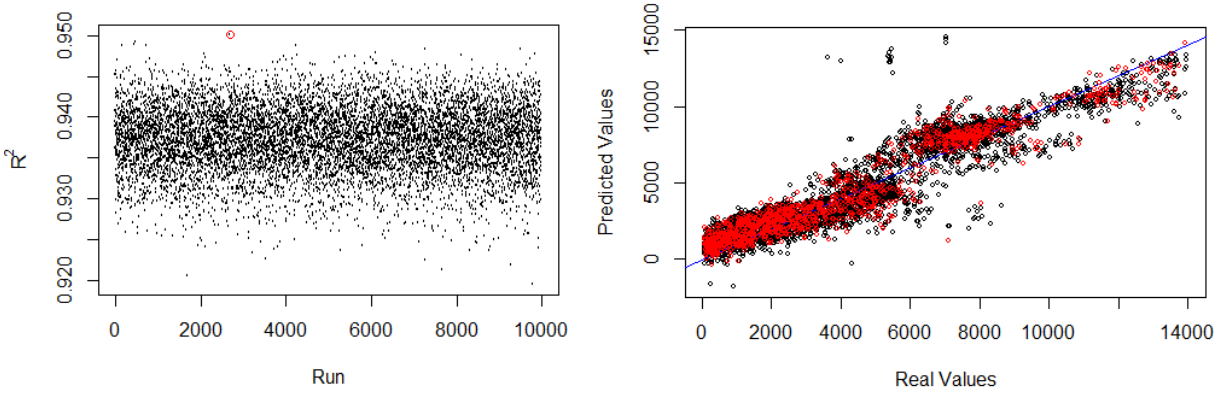


Figure 03: R^2 for PLSR models and Predicated *vs*. Real MAPD content for bed B

Both PLSR models for bed A and bed B adjusted remarkably well to the observable data, particularly the latter. A Bartel's rank test of randomness for the residuals of the PLSR model for bed A gives a p-value of 0.321 and for bed B a p-value of 0.062. For a significance level (α) of 0.05, there is no evidence to refuse the null hypothesis that the residuals are random.

4.3 Predicting MAPD content in the out stream through PLSR and VIP

The VIP scores for both beds are displayed in Figure 04, where the predicting variables are:

1: Fresh Feed Flow

2: Recycle Flow

3: Fed Hydrogen Flow

4: Combined feed temperature

5: Bed Pressure

6 – 11: Temperature across 6 different points in the reactor

12: Outflow temperature

13: MAPD content in feed

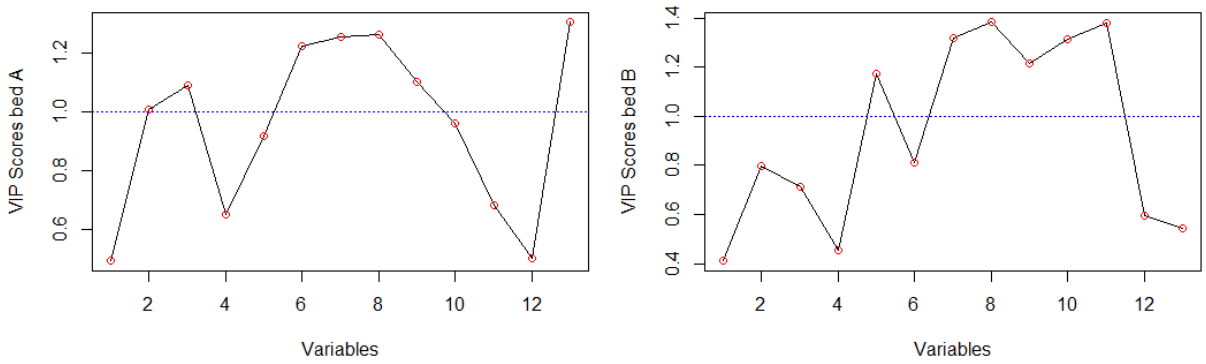


Figure 04: VIP scores for PLSR models for bed A and bed B

According to Figure 04, the selected variables for bed A are recycle flow, fed hydrogen flow, temperatures across four points in the reactor and feed MAPD content. For bed B, the selected variables were pressure and temperatures across five points in the reactor. It is important to notice that temperatures across the reactor scored VIP values greater than 1 for both beds, highlighting the importance of controlling such variables. The MAPD content in the feed is quite important for bed A, but not as much for bed B – its VIP score in the former (variable 13) is actually lower than 1.

Figure 05 displays Real *vs.* Predicted MAPD content for the PSLR models using only the selected variables. Visually, there is no great observable change for bed A, with points falling around the ideal blue line, although more spread, particularly at low values of MAPD. On the other hand, there is a visual drop in the performance of the model for bed B. In fact, the calculated residuals of the model for this bed show a p-value of 3×10^{-15}, whereas for model A such value is 0.611. Hence, with $\alpha = 0.05$, there is no evidence to support the null hypothesis that residuals from bed B PLSR model are random.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

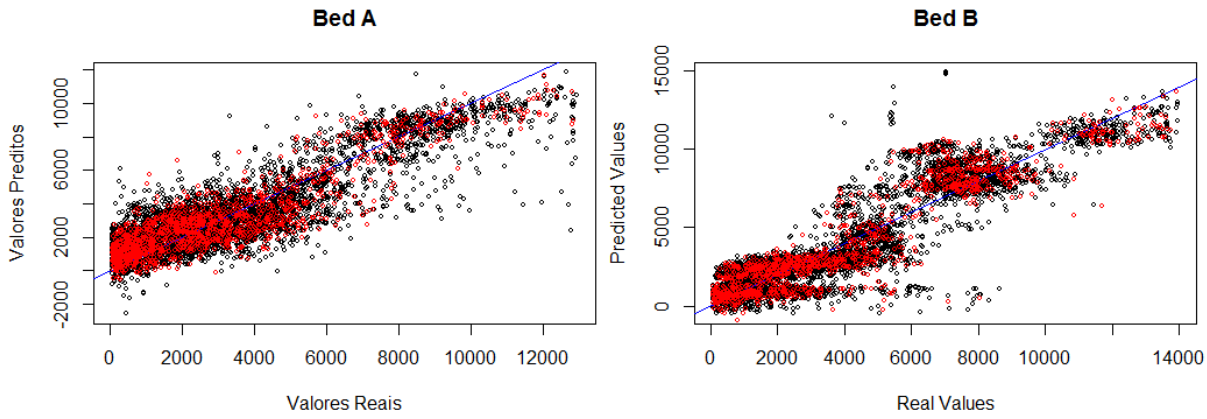


Figure 05: Predicted *vs.* Real MAPD content for PLSR models after variable selection

Table 01 summarises some parameters of models before and after variable selection. The number of process variables (PVs) decreased greatly for A and B. There was a decrease in the number of latent variables (LVs) for the PLSR model as well, leading to an easier interpretation of them. An increase of 16% and 19% in the RMSEP of bed A and B, respectively, can be noticed. Although the Pearson Correlation Coefficient slightly dropped for both beds, it still remained above 0.80, a remarkable performance given the industrial background of the data.

Table 01: Evaluation parameters for PLSR models before and after VIP

	Before variable selection				After variable selection			
	PVs	**LVs**	**R^2**	**RMSEP**	**PVs**	**LVs**	**R^2**	**RMSEP**
Bed A	13	5	0.887	1131 ppm	8	4	0.831	1312 ppm
Bed B	13	4	0.950	1065 ppm	6	3	0.921	1270 ppm

Although the PLSR model for bed B still has a remarkable R2 after variable selection (above 0.90), the assumption that residuals should be random could not be verified, as said before. This is a clear illustration of how deceivable R2 can be. Although the parameter indicates a suitable model, its failure to prove randomness of residuals nullifies its suitability. If such model is used, it could display extremely different performances for different MAPD values. As an example, the model can perform well for high values of MAPD but poorly predict low MAPD values. As such, the model should be used with caution.

5. Conclusions

The developed PLSR models in this paper show that it is possible to achieve a satisfactory virtual analyser that predicts the MAPD content in a trickle-bed reactor based on collected data from the process variables. PLSR models for two beds displayed an acceptable Pearson Correlation Coefficient of 0.887 and 0.950 and, according to Barlet's rank test, random residuals. In fact, the established models are relatively simple to be applied in case of need and have shown remarkable predictive capacity of MAPD content. Hence, there is a great potential for it to be implemented in the process control on the production of propene.

Besides that, the use of the VIP technique permitted a reduction in the model's dimensionality while still keeping suitable prediction capacity. However, only the PLSR model with variable selection for bed A showed evidence for residual randomness. Such was not found in the model for bed B, and such model should be used with caution as it could lead to wrong conclusions. In any case, the successful variable selection for the PLSR model in bed A shows the potential of the VIP technique.

6. References

Cohn, Pedro Estefano. *Analisadores Industriais: No Processo, Na Área De Utilidades, Na Supervisão Da Emissão De Poluentes E Na Segurança*. 1st ed. Rio de Janeiro: Interciencia, 2006. Print.

Ferreira, Daniela Souza, Ronei Jesus Poppi, and Juliana Azevedo Lima Pallone. "Evaluation of dietary fiber of Brazilian soybean (Glycine max) using near-infrared spectroscopy and chemometrics." *Journal of Cereal Science* 64 (May 2015): 43-47

Galindo-Prieto, Beatriz, Johan Trygg, and Paul Geladi. "A new approach for variable influence on projection (VIP) in O2PLS models." *Chemometrics and Intelligent Laboratory Systems* 160 (2017): 110-124

Morellato, Saulo Almeida. "Modelos de Regressão PLS com erros Heteroscedásticos"Master thesis, Universidade Federal de São Carlos, 2010. Retreived from https://repositorio.ufscar.br/bitstream/handle/ufscar/4541/2781.pdf?sequence= 1&isAllowed=y

Souza, Letícia Maria de. "Uso de espectroscopia no infravermelho médio, calibração multivariada e seleção de variáveis na quantificação de adulterantes em biodieseis e

suas misturas com diesel" Master thesis, Universidade Federal de Uberlândia, 2014.Retreived from https://repositorio.ufu.br/bitstream/123456789/17404/1/UsoEspectroscopiaInfravermelho.pdf

Tran, Thanh N., Nelson Lee Afanador, Lutgarde M.C. Buydens, Lionel Blanchet. "Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC)". *Chemometrics and Intelligent Laboratory Systems* 146 (2015) 297–304.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 38 - DONALD TRUMP E O TWEETER

Luiz Sá Lucas [37]
Felipe Souza [38]

Resumo

O trabalho tem como objetivo analisar tweets da segunda semana de março de 2017 referentes a Donald Trump através de mineração de texto (***text mining)***, abordando as forças e fraquezas da fonte de dados: os próprios tweets, devido ao seu pequeno tamanho – 140 caracteres. Foram obtidos, utilizando o package ***tweeteR***, do software R, na Web, cerca de cinco mil tweets sobre Trump (#TRUMP). A seguir os tweets foram analisados com o package ***tm*** e outros de análise multivariada, como ***cluster***, de nuvens de palavras (***wordcloud***) e de redes sociais (***igraph***) em R. Foi também efetuada uma análise de tópicos utilizando o package ***topicmodels***, também do software R.

Palavras-Chave: analytics, data mining, text mining, social data, tweeter

Abstract

The work aims to analyze tweets from the second week of March 2017, regarding Donald Trump, through text mining, addressing the strengths and weaknesses of the data source: the tweets themselves, due to their small size - 140 characters. We obtained, using the software ***tweeteR***, from the R software, on the Web, near five thousand tweets about Trump (#TRUMP). Next, the tweets were analyzed with the package ***tm*** and other multivariate analysis tools, such as ***cluster***, together with ***wordcloud*** and social networks analysis (***igraph***) in R. A topic analysis was also performed using the package ***topicmodels***, also from the R software.

Keywords: analytics, data mining, text mining, social data, tweeter

Introdução

Donald Trump, como candidato ou atual presidente dos Estados Unidos, tem como hábito se comunicar através do Tweeter (@POTUS / @realDonaldTrump). Recentemente, em março de 2017, o noticiário foi fortemente marcado por denúncias de que o ex-presidente Barack Obama havia interceptado comunicações internas de Trump. Verdadeiras ou não, essas denúncias geraram forte tráfego na mídia social. Este trabalho então se centra numa mineração de textos de tweets relativos a isso.

Objetivo

O trabalho tem como objetivo analisar tweets da segunda semana de março de 2017, referentes a Donald Trump (#TRUMP), através de mineração de texto (*text*

[37] MC15 Consultores – luizsa.lucas@mc15.com.br
[38] PUC-RJ- felipelobodesouza@yahoo.com.br

mining) e técnicas correlatas, abordando as forças e fraquezas da fonte de dados: os tweets.

Material e Métodos:

Mineração de textos (text mining) compreende um vasto campo de abordagens teóricas e métodos com um ponto em comum: texto como fonte de informação (Feinerer et al. 2008). De acordo com o texto citado, e de uma forma geral, text mining corresponde a um campo de atividade que engloba data mining, linguística, estatística computacional e ciência da computação. São técnicas comuns na área, entre outras, (i) a classificação e agrupamento (clustering) de textos e palavras; e (ii) o resumo de documentos. Text mining é: "o uso de uma coleta de um grande volume de textos com o objetivo de descobrir novos fatos ou tendências sobre o que se passa no mundo". (Hearst 1999).

Em *text mining*, a ideia é transformar texto, um dado em formato não estruturado, em uma estrutura, baseada em frequências de termos, e subsequentemente aplicar técnicas padrão de data mining / analytics a essa estrutura. Como se trata de texto não estruturado, essas técnicas se inserem no contexto do que se denomina Big Data.

Aplicações típicas em *clustering* de textos incluem agrupamento de novos artigos ou documentos em geral (Steinbach et al 2005). Em categorização de textos pode-se citar filtros de e-mail ou rotulação (*labeling*) de documentos em bibliotecas de qualquer área (Miller 2005). Outras aplicações, que são o foco deste trabalho, investigam as relações entre os termos principais (Zhao 2013). Denomina-se aqui como ***general keywords*** esses termos principais com os quais se trabalha neste documento. Boas fontes sobre o assunto são também Gries 2009, Liu 2012 e Srivastava e Sahami 2009. Além de *clustering* de documentos, uma importante ferramenta nessa área, também abordada neste trabalho, é a análise de tópicos (Grun e Hornik 2011, Taddy 2013, Blei et al. 2003, e Hofman et al. 2010).

No presente trabalho foram coletados, na segunda semana de março de 2017, pouco menos de cinco mil tweets relativos a #TRUMP, analisando-se o conteúdo dos tweets com o uso do package ***tm*** do software R e de pacotes em R de análise multivariada e de análise de redes sociais. Seguimos as seguintes etapas:

coleta dos ***cinco mil tweets*** on line
transformação dos tweets num ***data frame*** do R

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

transformação do data frame num ***corpus*** (um conjunto de documentos)
tratamento do corpus:
transformação dos termos em letras minúsculas (***tolower***)
remoção de caracteres especiais (\n, por exemplo), espaços desnecessários, números etc.
remoção de palavras que não colaboram com a análise (***stopwords***): 'the', 'as' etc. e termos que tais como 'trump', 'https', 'just' etc.
transformação do corpus em uma ***DocumentTermMatrix*** e em uma ***TermDocumentMatrix***
classificação dos tweets através de clustering hierarquizado
classificação das ***general keywords*** em grupos (clustering hierarquizado)
apresentação das relações entre as *general keywords* e seus grupos através de uma ***rede***
classificação dos documentos em ***tópicos***

Os principais ***packages*** do software R utilizados foram:

twitteR (Gentry 2015) , para a coleta dos tweets

tm (Feinerer e Hornik 2017)**,** para a análise de textos

cluster (Maechler et al. 2017) **,** para o agrupamento dos termos e de documentos

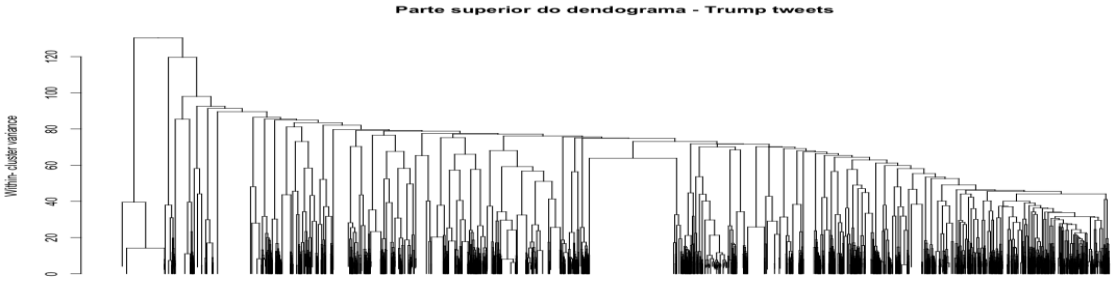
wordcloud (Fellows 2014) **,** para a criação de wordclouds

igraph (Csardi e Nepusz 2006), **,** para a análise de redes de termos

topicmodels (Grun e Hornik 2011) **,** para a análise de tópicos

Resultados e Discussão:

A figura 1 abaixo indica a parte superior do dendograma obtido na clusterização hierarquizada dos tweets / documentos. Nota-se que, de uma forma geral, existem três grandes grupos de documentos (à esquerda, e no centro e à direita, esses dois maiores):



Maiores detalhes sobre os grupos de tweets / documentos serão fornecidos mais adiante, ao se tratar da análise de tópicos.

As figuras a seguir indicam (i) um gráfico de barras com as principais general keywords; (ii) uma nuvem de palavras (***wordcloud***) das mesmas, mas filtrando numa menor frequência, para apresentar mais palavras; (iii) um dendograma de termos chave; e (iv) uma rede desses termos chave identificando dois grupos e suas relações.

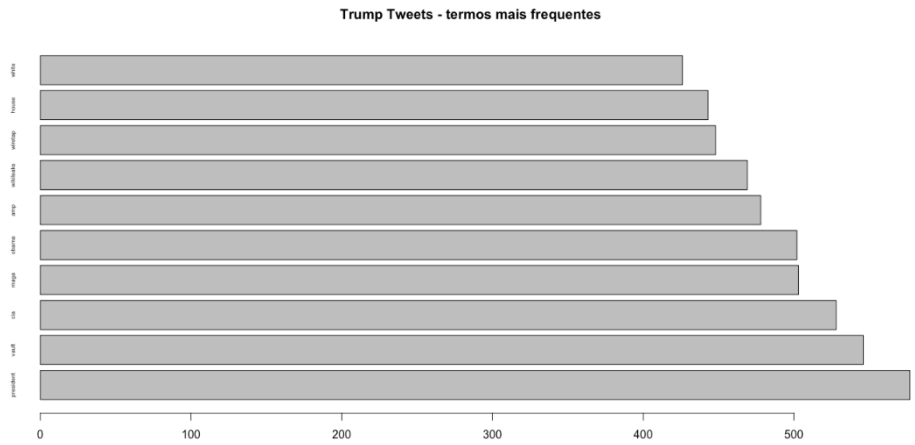


Figura 1 – Termos mais frequentes nos tweets

Os cinco termos mais frequentes nos tweets são ‘president’, ‘vault’, ‘cia’, ‘maga’ e ‘obama’. O termo ‘vault’ parece associado a ‘Trump vault’, um site que avalia citações on line sobre Donald Trump (Trumpvault 2017). ‘Maga’ significa ‘Make America Great Again’.

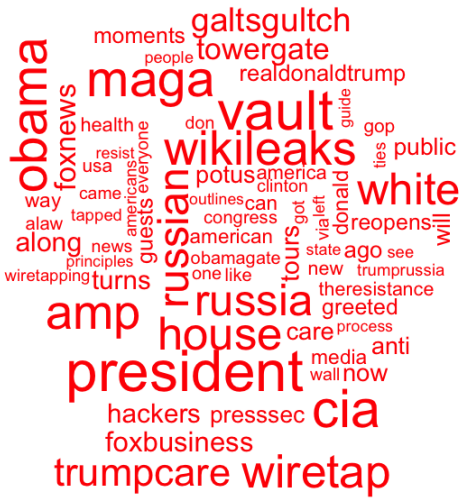


Figura 2 – Wordcloud – Trump tweets

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Além dos termos acima, temos também na Figura 2 muitos outros como ‘amp’, ‘white’ + ‘house’, ‘wikileaks’, ‘galtsgutch’, ‘russia’ + ‘russians’, ‘potus’+ ‘realdonaldtrump’ etc. O termo ‘amp’ se refere ao site AMP (www.amptud.com/tag/trump/), que entre outros assuntos, também trata de Donald Trump. O termo ‘galtsgutch’ refere-se a @galtsgutch. Tanto ‘potus’ como ‘realdonaldtrump’ referem-se a tweets de Trump.

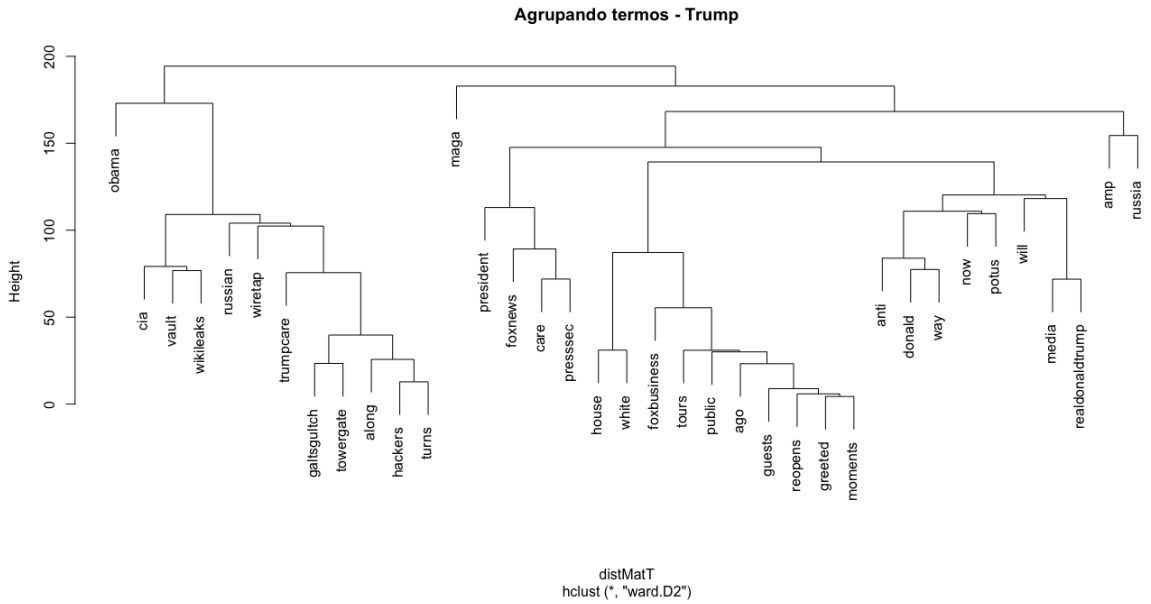


Figura 3 – Wordcloud – Trump tweets

A Figura 3 apresenta um agrupamento hierarquizado dos termos mais frequentes. Nota-se claramente a formação de dois grandes grupos (à esquerda e à direita). É interessante notar que se, com base no dendograma, forem formados três grupos, um deles (no centro), seria constituído apenas pelo termo ‘maga’.

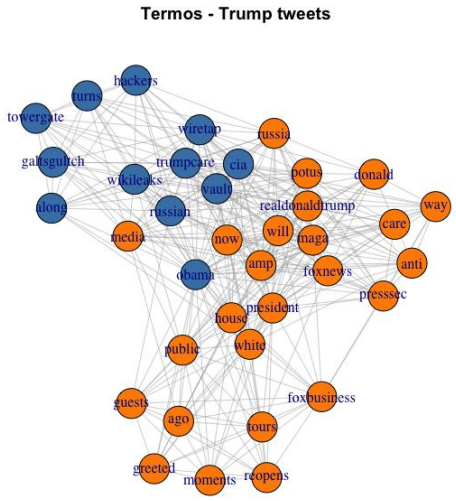


Figura 4 – Rede de Termos – Dois grupos de termos

A Figura 4 apresenta uma rede dos termos, mostrando a conexão entre eles, isto é, a co-ocorrência desses termos nos mesmos documentos / tweets. As cores indicam os dois grupos de termos. O grupo em azul, à esquerda e acima, trata mais especificamente dos vazamentos. O grupo à direita, em laranja, trata da presidência e das notícias sobre ela.

Foi efetuada também uma análise de tópicos. Esse tipo de análise tem semelhança com o agrupamento de texto via cluster analysis, como foi comentado na Figuar 1, que apresenta o dendograma da *clusterização* dos documentos. A análise de tópicos, já citada anteriormente, procura, através de algoritmos como Expectation-Maximization e técnicas bayesianas (ver, p.ex. Grun e Hornik 2011), e modelos como Regressão Multinomial Inversa (Taddy 2013), alocar documentos a tópicos, mas difere da clusterização usual na medida em que pode alocar um documento a mais de um tópico.

A Figura 5 abaixo apresenta os resultados da aplicação do modelo VEM (Grun e Hornik 2011) aos tweets aqui analisados:

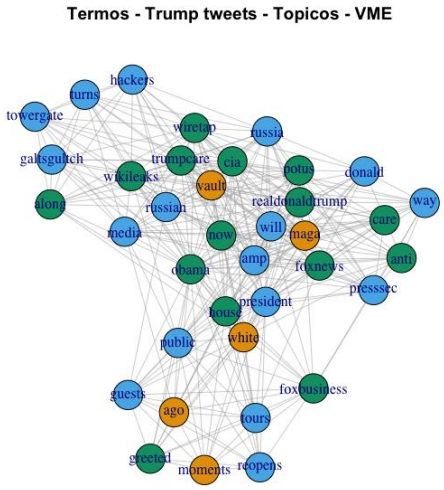
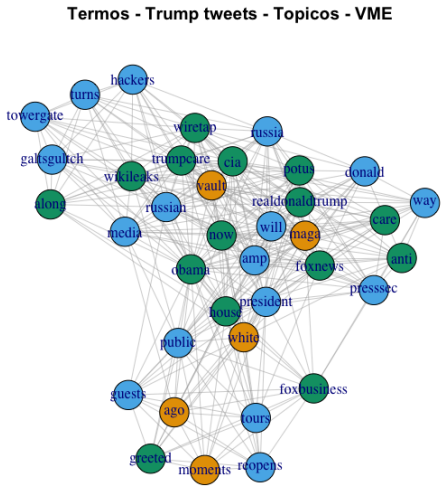


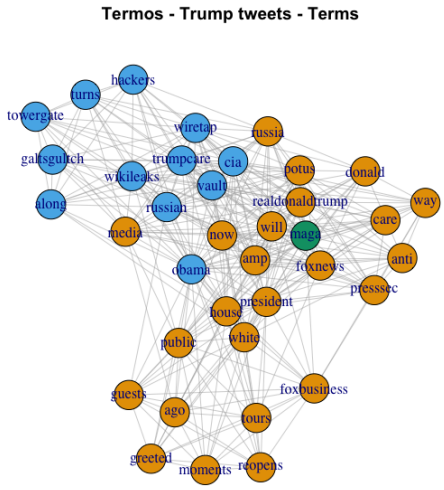
Figura 5 – Rede de Termos – Dois grupos de termos

Foi adotado um número k=3, de três tópicos, sendo então apresentada a rede de termos mais frequentes, onde a cor do nó (termo) identifica o tópico a que ele pertence. O grupo em verde trata de Trump, Obama, ‘house’, ‘president’ e notícias (‘foxnews’), incluindo vazamentos (‘wikileaks’). O grupo em azul se concentra mais nos assuntos relativos à Rússia e também à mídia. Inclui @PressSec, referente a Sean Spicer, Secretário de Imprensa da Casa Branca. O grupo em amarelo trata mais de ‘maga’ e ‘vault’, isto é, parece um tópico mais difuso.

Finalmente, a Figura 6 apresenta uma comparação entre dois agrupamentos de termos:

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017





Se a preferência for pela adoção for três grupos de termos, a adoção do agrupamento de termos decorrente da análise de tópicos (que agrupa documentos) parece mais informativa.

Conclusão:

Os tweets trazem uma grande dificuldade na análise de seu conteúdo: como seu limite é de 140 caracteres, fica difícil extrair dados mais profundos ou consistentes. Identifica-se apenas temas gerais, ou principais temas focados.

Mesmo assim, a análise via ***text mining*** de tweets pode ser extremamente valiosa para identificar temas relevantes numa dada área, e para avaliar a evolução no tempo desses temas relevantes. Fica assim clara sua importância, entre outras, nas áreas de marketing e opinião pública, identificando tendências e aspectos relevantes numa dada época ou ao longo do tempo.

Cabe finalizar indicando que se trata de um projeto em andamento, que pretende aprofundar o uso dessas técnicas em outras formas de mídia social e em textos de maior volume.

Referências:

Blei, D., Ng,A. e Jordan, M. 2003 *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3 (2003) 993-1022.

Csardi G, Nepusz T: *The igraph software package for complex network research*, InterJournal, Complex Systems 1695, 2006.

Feinerer, I. e Hornik, K., *tm: Text Mining Package*. R package version 0.7-1, 2017.

Feinerer, I., Hornik, K. e Meyer, D., *Text Mining Infrastructure in R*, Journal of Statistical Software, 2008. Vol. 25, Issue 5.

Fellows, I., *wordcloud; Word Clouds*. R package version 2.5, 2014.

Gentry, J., *twitteR: Text Mining Package*. R package version 1.1.9, 2015.

Gries, S., *Quantitative Corpus Linguistics with R*, New York: Routledge, 2009.

Grün B. e Hornik K., *topicmodels: a Package for Fitting Topic Models*. Journal of Statistical Software, 40(13), 1–30, 2011

Hearst, M., *Untangling Text Data Mining*, Proccedings of the 37th annual meeting of the Association of Computational Linguistics", 1999. pp.3-10.

Hoffman, M., Blei, D. e Bach, F., *Online Learning for Latent Dirichlet Allocation*, NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems, pp. 856-864, 2010.

Liu, B., *Sentiment Analysis and Opinion Mining*, Toronto: Morgan & Claypool, 2012.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6, 2017.

Miller, T., *Data and Text Mining*. Pearson Educational International. 2005.

Srivastava, A. e Sahami, M., *Text Mining- Classification, Clustering and Applications,* Boca Raton: CRC Press, 2009.

Steinbach, M., Karypis, G. e Kumar, V., *Ä Comparison of Document Clustering*, KDD Workshop on Text Mining, 2000.

Taddy, M., Multinomial Inverse Regression for Text Analysis, Journal of the American Statistical Association, Vol 108, Issue 503, pgs. 755-770, 2013.

Trumpvault.com, http://trumpvault.com/ , acesso em 14 mar 2017.

Zhao,Y., *R and Data Mining –Examples and Case Studies*, San Diego: Elsevier, 2013.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID40 - CREATING AND GRADING LATEX EXAMS WITH RANDOMIZED CONTENT USING RNDTEXEXAMS

Marcelo S. Perlin [39]

Resumo

This paper introduces the functionalities of \rndtexexams{}, an R package designed to minimize visual cheating and facilitate the grading of printed exams. Built on top of \emph{R} and \emph{\LaTeX} exam classes, the package makes it easy to create exams with randomized spatial content by automatically changing the order of questions, their textual content, and the order of answers. It also includes a specialized function that, when combined with a cloud-based service, makes it easy to grade a large number of multiple-choice exams built with \rndtexexams{}. The package can be used to statistically test for cheating based on student answer sheets. In this document, we describe the usage and features of the package and present several examples.

Palavras-Chave: RndTexExams, random exams

Abstract

This paper introduces the functionalities of \rndtexexams{}, an R package designed to minimize visual cheating and facilitate the grading of printed exams. Built on top of \emph{R} and \emph{\LaTeX} exam classes, the package makes it easy to create exams with randomized spatial content by automatically changing the order of questions, their textual content, and the order of answers. It also includes a specialized function that, when combined with a cloud-based service, makes it easy to grade a large number of multiple-choice exams built with \rndtexexams{}. The package can be used to statistically test for cheating based on student answer sheets. In this document, we describe the usage and features of the package and present several examples.

Keywords: RndTexExams, random exams

Introduction

A significant part of the academic system is related to the assessment of knowledge that students retain after completing their coursework. This is usually implemented in the form of collective examinations that give the university some quality control regarding the graduating students. It makes economic sense to implement such a system, as better-prepared graduates will enhance the university's reputation, which will then attract more students, better faculty, and possibly more funding.

Although, in theory, collective examinations are straightforward, in practice, the efficiency of the process can be jeopardized by external factors. A small physical location for a large class can make it easy for students to copy answers from colleagues seated nearby. The technological advances of smartphones have also

[39] UFRGS, marceloperlin@gmail.com

facilitated the sharing of answers in identical exams \citep{teixeira2010cheating,meier2005stealing}.

Despite its negative impacts, cheating is widespread and hard to control. A recent study by \cite{teixeira2010cheating} reports that, out of 7,213 students, 62\% have copied answers during an exam and 90\% have seen other students copying answers from those seated nearby. From the faculty side, the work of \citet{mccabe2005cheating} indicates that 41\% of the university faculty have seen one student copying from another on a test without their knowledge. The empirical evidence of cheating is strongly corroborated by several previous studies on the topic \citep{hsiao2015impact,sideridis2015predicting,richmond2015detection,david2015academic}.

It is also hard to counter act against the possibility of visual cheating. Larger physical locations and exams with mixed students from different classes can help but it significantly increases the economic costs of the exams. Manually creating different exams also help but it becomes a burden for the examiner in the creation and grading of the different tests. This situation motivated the creation of a tool that can assist the examiner. This paper introduces \rndtexexamsCRAN{}, a R package designed to help minimizing visual cheating in exams by making it easy to create different versions of the same exam and grade them without hassle.

The use of software to create dynamic documents such as exams and exercises with random content is relatively new. \cite{dryver2009enhancement} proposed a combination of Javascript and \LaTeX{} to build exams using a random data generating process (DGP). This framework allowed the user to create exams with numerical questions that have different values and results, but which can be solved with the same intrinsic calculations. By creating a large number of exams, the use of a random DGP ensured that students faced a variety of questions, making it difficult for them to learn the answers \textit{by heart}. The idea of building exams with random DGPs was further developed by \citet{Grun_2009} with the R package \CRANpkg{exams}. This package uses Sweave \citep{leisch2002sweave} to identify and dynamically generate parts of the exam that contain random content.

\rndtexexams{} is a specialized package that makes it easy to produce exams with random content. The user's input is based in \LaTeX{} templates \textit{exam} and \textit{examdesign}. An examiner with a test in this format can start using \rndtexexams{} directly, with minor additions to the \LaTeX{} code. The randomization

of the exam requires no superior knowledge of R functions as all the important parts are identified based on the structure of the \LaTeX{} file. \rndtexexams{} mixes the order of the questions, their textual content, and the order of the answers. Similar to the use of Sweave, the user can create textual switches directly in the \LaTeX{} code based on a pre-defined set of symbols. Within a classroom of any size, it is very unlikely that two students seated close together will have the same exam or correct answer key. Essentially, a unique test can be produced for each student. When using a diversified set of versions, the use of \rndtexexams{} makes it very difficult for students to cheat by looking around or sharing answers.

The remainder of this paper is organized as follows. Next, we list the software requirements for using \rndtexexams{}. Practical examples of creating questions, building exams with random content. The paper finishes with the usual concluding remarks.

How to use RndTexExams

Requirements

To use \rndtexexams{}, a computer must have \LaTeX{} and pdflatex installed. There are two main choices of distributions, Miktex and Texlive. Both are good choices and will work well with \rndtexexams{}. Users should also download a suitable \myLatex{} interface such as texstudio. The installation of R is an obvious requirement. The list of requirements and their locations for download are as follows:

\begin{description}
\item[Pdflatex (requirement)] \url{www.miktex.org} or \url{https://www.tug.org/texlive/}
\item[R (requirement)] \url{https://www.r-project.org/}
\item[Texstudio (optional)] \url{www.texstudio.org/}
\item[RStudio (optional)] \url{https://www.rstudio.com/}
\end{description}

All of the code demonstrated in this section was executed in R $3.3.0$ \textit{``Supposedly Educational''}, with version 1.4 of \rndtexexams{}. A working internet connection is required to replicate the examples.

It is also necessary to install the \LaTeX{} packages exam and examdesign, and also the system function \code{texi2dvi}. For Linux users, the following terminal code will ensure that all \LaTeX{} requirements are installed and ready for use:

\begin{Verbatim}
sudo apt-get install texlive-base texlive-latex-extra texinfo
\end{Verbatim}

Be aware that these are heavy dependencies. If texlive is not installed, it may take some time to download it.

\subsection[Writing questions and building exams with RndTexExams]{Writing questions and building Exams with \rndtexexams{}}

The \rndtexexams{} package works by taking as input a file containing an exam based on two possible \LaTeX{} templates: examdesign and exam. Both are mature \LaTeX{} packages for creating exams. Within \rndtexexams{}, the identification of the \LaTeX{} class is automatic, based on the file contents. The user can find examples of \LaTeX{} exam files in the exam and examdesign format in the author \href{https://gist.github.com/msperlin}{gist repository}.

For the rest of this article, we will use the template for examdesign as this exam class has an intuitive structure. Thus, before using \rndtexexams{}, it is necessary to understand how questions are formulated in examdesign.

A standard multiple-choice exam is composed of several questions that have a set of alternative answers to choose from. A simple visual example of a multiple-choice question is:

\begin{Verbatim}[frame=single]
Given the next five options, which one is the correct answer?
a) Choice 1
b) Choice 2
c) Choice 3
d) Choice 4

e) Choice 5 - The CORRECT answer!
\end{Verbatim}

In examdesign, the structure of a multiple-choice question is defined using \myLatex{} commands. The start of the multiple-answer section is encapsulated by the commands \verb|\begin{multiplechoice}| and \verb|\end{multiplechoice}|. Within this environment, all questions begin with \verb|\begin{question}| and end with \verb|\end{question}|. The multiple answers of the questions are marked as \verb|\choice{}|. The \LaTeX{} code for the above example can be written as:

```
\begin{Verbatim}[frame=single]
% LaTeX example of a multiple choice question in examdesign
% The preamble and the rest of the document are omitted for simplification.
% Be aware that this simple code as it is will NOT compile in
% pdflatex, as there are other requirements

\begin{multiplechoice}

\begin{question}

Given the next five options, which one is the correct answer?

\choice{Choice 1}
\choice{Choice 2}
\choice{Choice 3}
\choice{Choice 4}
\choice[!]{Choice 5 - The CORRECT answer!}

\end{question}
\end{multiplechoice}
\end{Verbatim}
```

Note that, in this simple example, each question has a main text and choices. The correct answer to the question is marked with the symbol \samp{[!]}. The correct

answer for each question can later be used to build the correct answer key of each version of the exam.

The \rndtexexams{} package reads the \LaTeX{} file, searches for all occurrences of a multiple choice question, randomly rearranges the order of questions and possible answers, and finally builds a new \LaTeX{} file. This process is repeated for N exams. Therefore, each version of the test will have a different key answer sheet. The \LaTeX{} files are later compiled from R, resulting in a set of \textit{pdf} files that are ready for printing.

Changing the textual content of the exam

The package \rndtexexams{} can use textual switches with specific symbols to define which parts of the questions can change between versions. This is an optional feature for those examiners that wish to create versions of the same questions with different interpretations.

Any change in the text, whether in the main text of the question or the text of the answers, is organized with symbol \verb/@/. To clarify, each version of the test will show the text according to its position. Thus, version one will show the text \verb|text in ver 1|, version two will show the text \verb|text in ver 2|, and so on. The version of the content changes every time R and \LaTeX{} compile a new test.

As we are changing the text of the questions and answers, it is also necessary to change the correct answers in each version. To do this, the symbol \verb|[x]| is added to the text of the answers, where \verb|x| is the version in which the choice is correct. For example, we can make different versions of the previous example question using the following \LaTeX{} code with \rndtexexams{}:

```
\begin{Verbatim}[frame=single]
% Example of multiple choice question in examdesign, with 2 versions
\begin{multiplechoice}

\begin{question}

Given the next five options, which one is the correct answer
in @{version 1}|{version 2}@?
```

\choice{Choice 1 - Incorrect in all versions}
\choice{[2] Choice 2 - @{Incorrect in version 1}|{Correct in version 2}@ }
\choice{Choice 3 - Incorrect in all versions}
\choice{Choice 4 - Incorrect in all versions}
\choice{[1] Choice 5 - @{Correct in version 1}|{Incorrect in version 2}@ }

\end{question}
\end{multiplechoice}
\end{Verbatim}

And that's it! The R code in \rndtexexams{} will look for these symbolic expressions and randomly choose one of them for the final version of each exam.

Once the questions have been coded with the proper syntax for use with \rndtexexams{}, the \LaTeX{} file is simply passed to the functions \code{rte.analyze.tex.file} and \code{rte.build.rdn.test}. This \LaTeX{} file should have all of the other exam components, such as some student identification area, and class name. The easiest way to start using \rndtexexams{} is to modify the example file\footnote{Also available from \href{https://gist.githubusercontent.com/msperlin/}{gist}} in the package. Next, I present an example of the use of \rndtexexams{}.

\begin{example}
R> library(RndTexExams) # from CRAN
R> set.seed(5)
R> ## Get latex file from package
R> f.in <- system.file("extdata", "MyRandomTest_examdesign.tex", package = "RndTexExams")
R> ## Breakdown latex file into a a list
R> list.out <- rte.analyze.tex.file(f.in)
\end{example}

\begin{example}
rte: Changing LaTeX file into dataframe... Done
\end{example}

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

The function \code{rte.analyze.tex.file} analyses the \LaTeX{} file and produces an R list with all of the details of the exam. This finds and separates all of the multiple choice questions. An example of the resulting dataframe in the output list is as follows.

\begin{example}
R> print(sapply(list.out,class))
\end{example}

\begin{example}
df.questions df.answers my.begin.mchoice.line
"data.frame" "data.frame" "character"
my.preamble my.last.part examclass
"character" "character" "character"
\end{example}

The item \code{df.questions} contains all of the questions in the exam, \code{df.answers} provides all of the answers in a dataframe, and the remaining character items are the additional \LaTeX{} code found in the \LaTeX{} file. The exam file is broken down by \code{rte.analyze.tex.file}, and can be reassembled without any loss of information based on \code{list.out}.

We proceed by creating the exams with \code{rte.build.rdn.test}.

\begin{example}
R> # Options for build.rdn.test
R> list.in <- list.out # output from rte.analyze.tex.file
R> f.out <- 'MyRandomTest_' # pattern of pdfs (MyRandomTest_1,.)
R> n.test <- 5 # number of random tests
R> n.question <- 4 # number of questions
R>
R> # Builds pdfs
R> list.build.rdn.exam <- rte.build.rdn.test(list.in = list.in,
f.out = f.out,
n.test = n.test,

```
    n.question = n.question)
```

\end{example}

\begin{example}

```
    rte: Checking for error in inputs... Done
    rte: pdflatex flavor: miktex
    rte: Type of OS: Windows
    rte: Latex compile function: texi2pdf
    rte: Type of exam template: examdesign
    rte: Number of mchoice questions: 4
    rte: Building Test #1...Done
    rte: Building Test #2...Done
    rte: Building Test #3...Done
    rte: Building Test #4...Done
    rte: Building Test #5...Done
    rte: FINISHED - Check folder PdfOut for pdf files
```

\end{example}

The function \code{rte.build.rdn.test} uses the output from \code{rte.analyze.tex.file}, randomizes all of the content of the multiple choice questions, and, at the end of each iteration, pastes together a new \LaTeX{} file that is used to compile the pdf files of the exam. We can check the existence of the pdf files by running the following code:

\begin{example}

```
    R> print(list.files(path = 'PdfOut', pattern = '*.pdf'))
```

\end{example}

\begin{example}

```
    [1] "MyRandomTest_1.pdf"  "MyRandomTest_2.pdf"
    [4] "MyRandomTest_3.pdf"  "MyRandomTest_4.pdf"  "MyRandomTest_5.pdf"
```

\end{example}

The correct answer sheet for all versions is available in \code{list.build.rdn.exam\$answer.matrix}, where each row is the version of the test and the columns are the answers. A dataframe version of the answer key is also available in \code{list.build.rdn.exam\$df.answer.wide}.

\begin{example}

R> print(list.build.rdn.exam$answer.matrix)

\end{example}

\begin{example}

1 2 3 4

Version 1 "c" "b" "e" "a"

Version 2 "b" "d" "e" "e"

Version 3 "b" "e" "e" "a"

Version 4 "d" "a" "c" "b"

Version 5 "e" "a" "a" "a"

\end{example}

The list \code{list.build.rdn.exam} should be locally saved using the function \code{save} for later use in grading the exams. It is advisable to use the function \code{set.seed} before \code{rte.build.rdn.test} to ensure that every execution of the code results in the same random answer sheet. If there are any problems in the exam, it is possible to rerun the code with the same configuration.

Conclusion

This paper has described the use of the \rndtexexams{} package, an R package designed to create exams with randomized content. We demonstrated how users can create their own exams based on template examdesign and form different versions with \rndtexexams{}. We also explained how to use cloud-based services to easily grade exams created with the package. The use of \rndtexexams{} is straightforward. Anyone familiar with \LaTeX{} and R will be able to save a considerable amount of time that would previously be spent writing different versions and grading exams.

The \rndtexexams{} package has been used in real classrooms with very positive feedback. It was used in five different classes in the author's university, and the teachers were especially pleased with the possibility of minimizing cheating and digitally grading the exams. The main difficulty faced by users was writing the original exam in \LaTeX{}. None of the instructors were familiar with \LaTeX{} or examdesign, and so constant interventions were required by the developer. A more user-friendly and visual interface to \rndtexexams{} could facilitate the popularization of the package.

References:

David, L. T. (2015). Academic cheating in college students: Relations among personal values, self-esteem and mastery. Procedia-Social and Behavioral Sciences, 187:8892.

Dryver, A. et al. (2009). The enhancement of teaching materials for applied statistics courses by combining random number generation and portable document format les via latex. Journal of Statistical Software, 31(3):19.

Grün, B. and Zeileis, A. (2009). Automatic generation of exams in r. Journal of Statistical Software, 29(1):114.

Hsiao, C.-H. (2015). Impact of ethical and aective variables on cheating: Comparison of undergraduate students with and without jobs. Higher Education, 69(1):5577.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In Compstat, pages 575580. Springer.

McCabe, D. L. (2005). Cheating among college and university students: A north american perspective. International Journal for Educational Integrity, 1(1).

Meier, B. (2005). Stealing the Future: Corruption in the Classroom; Ten Real World Experiences. TI.

Richmond, P. and Roehner, B. M. (2015). The detection of cheating in multiple choice examinations. Physica A: Statistical Mechanics and its Applications, 436:418429.

Sideridis, G. D., Tsaousis, I., and Al Harbi, K. (2015). Predicting academic dishonesty on national examinations: The roles of gender, previous performance, examination center change, city change, and region change. Ethics & Behavior, pages 123.

Teixeira, A. A. and Rocha, M. F. (2010). Cheating by economics and business undergraduate students: An exploratory international assessment. Higher Education, 59(6)

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 41 - IMPACTO DE CARACTERÍSICAS ESCOLARES NAS NOTAS DO ENEM: UM ESTUDO COM METADADOS

Vinícius do Carmo Oliveira de Lemos[40]
Bruno Figueiredo Damásio[41]

Resumo

O Exame Nacional do Ensino Médio (ENEM) avalia a qualidade do ensino médio através de um exame aplicado nacionalmente todo ano, desde 1998, com reformulações em 2009. Além de ser uma forma de avaliar a educação, o ENEM também serve como forma de ingressos em cursos superiores. Outra importante forma de avaliação da educação nacional é o Censo da Educação. Ele se refere a uma avaliação a nível nacional que coleta dados sobre educação em todas as escolas públicas e privadas, de todos os níveis de educação. Este estudo tem o objetivo de investigar a relação entre os escores do ENEM com a porcentagem de disciplinas ensinadas nas escolas e a infraestrutura das escolas de todos os municípios brasileiros. Os dados foram retirados do ENEM e Censo da Educação de 2014 e concatenados no nível dos municípios. Este procedimento permitiu que fossem combinados dados de características escolares dos municípios com os escores do ENEM. Resultados demonstram que as relações entre o número de disciplinas e a estrutura das escolas, em geral, não se correlacionam significativamente com os escores do ENEM.

Palavras-Chave: Educação, ENEM, Censo da Educação

Abstract

The Exame Nacional do Ensino Médio (ENEM) evaluates the quality of high school education in Brazil through a standard exam applied nation-wide every year, since 1998, with reformulations in 2009. Besides education evaluation ENEM also serves as the individuals' form of ingress in colleges. Another important national evaluation is the Censo da Educação. It refers to a nationwide evaluation which gather data about education in all public and private schools, from all levels of basic education. This study has the objective of investigate the relation between ENEM's scores with percentage of disciplines taught in schools and infrastructure of schools from all Brazilian municipalities. Data were gathered from 2014's ENEM and Censo da Educação concatenating then on municipalities' levels. This procedure allowed to match municipalities' school's characteristics with its ENEM´s grade. Results show that the relations between number of disciplines and the structure of the schools generally do not correlate with ENEM's scores.

Keywords: Education, ENEM, Censo da educação

Introdução

No Brasil, existe grande variabilidade entre escolas em relação a recursos físicos e humanos. Diferentes escolas podem não dispor dos mesmos recursos

[40] Instituto de Psicologia, Departamento de Psicometria, Laboratório de Psicometria e Psicologia Positiva, LP3, UFRJ, lemos.vncs@gmail.com
[41] Instituto de Psicologia, Departamento de Psicometria, Laboratório de Psicometria e Psicologia Positiva, LP3, UFRJ, bf.damasio@gmail.com

básicos para funcionamento. Deste modo, a presença destes recursos torna-se um fator de eficácia escolar (Franco et al., 2007). De acordo com Franco et al. (2007), em 2007, 39% da variância da proficiência de matemática em alunos de 4ª série foi explicada por fatores entre relacionados às diferenças escolares, enquanto em países europeus e nos EUA, a variância entre escolas está em torno de 20%. Isso é um indicativo de que nosso sistema educacional apresenta alta desigualdade. Nesse contexto, este estudo busca investigar quais variáveis do Censo Escolar de 2014 impactam as diferentes notas do Exame Nacional do Ensino Médio (ENEM) de 2014.

O ENEM foi criado em 1998 com o intuito de avaliar o desempenho de estudantes ao final do ensino básico, a fim de melhorar a qualidade de escolarização no Brasil. Em 2004, o ENEM passou a ser, também, uma forma de ingresso no Ensino Superior, sendo combinado com outras formas de processo seletivo nas universidades, ou sendo a única forma de seleção de novos estudantes, de acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2017a). Isso é, de acordo com o instituto, uma forma de democratizar "as oportunidades de acesso as vagas oferecidas por Instituições Federais de Ensino Superior (IFES) ".

Os resultados do ENEM são utilizados para: "compor a avaliação de medição da qualidade do Ensino Médio no País", "subsidiar a implementação de políticas públicas", "criar referência nacional para o aperfeiçoamento dos currículos do Ensino Médio", "desenvolver estudos e indicadores sobre a educação brasileira", estabelecer critérios de acesso do participante a programas governamentais", "constituir parâmetros para a autoavaliação do participante, com vista à continuidade de sua formação e à sua inserção no mercado de trabalho" (INEP, 2014).

De acordo com o Edital do ENEM (INEP, 2014), o exame é constituído de uma redação e quatro provas objetivas, cada uma contendo 45 questões de múltipla escolha e abordando quatro áreas de conhecimento: Ciências Humanas, Ciências da Natureza, Linguagens, Códigos e Redação e Matemática. Sendo estas realizadas em dois dias de exame. No primeiro dia são realizadas as provas de Ciências Humanas e Ciências da Natureza. No segundo dia são realizadas as provas de Linguagens, Códigos, Redação e Matemática. Após as provas corrigidas é realizado um cálculo de proficiências nas provas objetivas utilizando-se da Teoria de Resposta ao Item (TRI). A TRI é um modelo matemático que visa avaliar cada item de um teste separadamente, e que permite ter parâmetros de dificuldades sem variação

dependente da amostra. A redação é corrigida com base em 4 critérios, variando de 0 a 1000 pontos.

O Censo Escolar é, de acordo com o INEP (2017b), o principal instrumento de coleta de dados da educação básica, tendo por finalidade prover dados para a compreensão da situação educacional brasileira, através de indicadores como o Índice de Desenvolvimento da Educação Básica (IDEB), taxas de rendimento, fluxo escolar, entre outros. É realizado de acordo com o decreto nº 6.425, de 4 de abril de 2008, coletando dados de forma descentralizada de todos os estabelecimentos de ensino públicos e privados da educação básica, adotando alunos, turmas, escolas e profissionais da educação como unidades de informação. O Censo Escolar abrange o ensino regular (educação infantil, ensino fundamental e médio), além de educação especial, Educação de Jovens e Adultos (EJA) e educação profissional (INEP, 2017b).

Objetivo

Este estudo busca investigar quais, dentre as variáveis do Censo Escolar, tem relação com as notas do ENEM, ambos aplicados no ano de 2014, separando-as de acordo com as quatro categorias administrativas das escolas brasileiras: privadas, municipais, estaduais e federais, para alunos do ensino médio.

Material e Métodos:

Para análise foram utilizados três bancos de dados: 1) microdados sobre escolas do Censo Escolar 2014; 2) microdados sobre turmas do Censo Escolar 2014; 3) microdados do ENEM 2014. O primeiro contém informações sobre a estrutura das escolas, como presença de laboratórios de informática, internet banda larga, refeitórios, etc. O segundo contém informações sobre as disciplinas oferecidas nas escolas. O terceiro contém informações sociodemográficas dos participantes do ENEM (sexo e raça), e as notas do ENEM. A quantidade de casos analisados para cada categoria administrativa e banco de dados se encontra na tabela 1. As informações utilizadas neste estudo encontram-se disponibilizadas em bases de dados nacionais, de acesso livre, e podem ser acessadas através do site *dados.gov.br*.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Tabela 1

Número de casos por categorias administrativas por banco de dados

Categoria Administrativa	ENEM	Escolas	Turmas
Particular	335.848	10.813	53.750
Municipal	13.182	362	2.423
Estadual	997.302	19.132	231.958
Federal	29.995	514	6.776
Total:	1.376.327	30.821	294.907

Como foram utilizados três diferentes bancos de dados, os casos foram agrupados pelos municípios das escolas ou município de residência do participante do ENEM. As variáveis contínuas (como as notas do ENEM) foram transformados em escores z para cada município, para possibilitar as suas comparações.

As variáveis discretas foram transformadas em porcentagem por município. Ou seja, foi verificada a porcentagem de escolas ou turmas num determinado município que continham determinada característica. Por exemplo, a porcentagem de turmas que têm aulas de matemática no município de Paracambi. Ao final, foram elaborados dois bancos de dados, um com as características das escolas e notas do ENEM e outro com dados sobre as turmas e notas do ENEM.

Em relação ao banco do ENEM, foram excluídos participantes que não atenderam a alguma das provas, ou tiveram suas provas ou notas inválidas por qualquer motivo. Só foram selecionados participantes que haviam concluído o Ensino Médio no ano de 2014 ou o iriam concluir no mesmo ano. Também só foram aceitos dados de escolas na modalidade de ensino regular e na etapa “ensino médio”.

As bases de dados permitiram que fosse avaliado a i. porcentagem de turmas em um município que tinham as seguintes disciplinas: física, matemática, inglês, espanhol, francês, outras línguas, língua indígena, artes, educação física, história, geografia, sociologia, filosofia e ensino religioso e ii. a porcentagem de escolas em um município que continham as seguintes características de infraestrutura: presença de biblioteca, sala de leitura, dependências e vias adequadas a alunos com deficiência ou mobilidade reduzida, refeitório, pátio coberto, pátio descoberto, área verde,

internet, internet banda-larga, fornecimento de alimentação escolar, se tem atendimento educacional especializado, atividades complementares, materiais didáticos específicos para atendimento à diversidade sociocultural, proposta pedagógica de formação por alternância, se possui água filtrada e esgoto. Também foram calculados para as escolas o escore z, descrito acima, em relação ao número de equipamentos de TV, DVD, copiadora, retroprojetor, equipamentos de som e equipamentos multimídia.

Todas as variáveis foram correlacionais através da correlação bi-variada com o escore z das seguintes áreas do ENEM: Ciências da Natureza, Ciências Humanas, Linguagens e Códigos e Matemática, e com o escore z da nota de Redação. Consideraremos fracas as relações a partir de 0,3; moderadas as relações a partir de 0,5 e fortes as relações a partir de 0,7; de acordo com os critérios de Mukaka (2012). As análises foram realizadas em separado para escolas e turmas de escolas particulares, municipais, estaduais e federais, gerando uma matriz de relação 5 por 24, para escolas particulares e 5 por 16 para o banco de dados de turmas. As outras categorias administrativas tiveram algumas variáveis retiradas por inconsistência nos dados (escolas federais, por exemplo, não apresentam ensino de língua indígena). Portanto, escolas municipais, estaduais e federais, não tem a variável alimentação, e escolas federais apresentam, também, não apresentam as variáveis sobre material especial para diversidade sociocultural e esgoto. Em relação às turmas, escolas particulares possuem todas as variáveis, escolas estaduais e federais não possuem "língua indígena", e turmas de escolas federais não possuem a variável "ensino religioso".

As análises foram realizadas no software RStudio, utilizando-se da linguagem *R* em sua versão 3.3.2 (R Core Team, 2016). Foram utilizados os pacotes *psych* (REVELLE, 2016) para realização das correlações bi-variadas e cálculo da significância. Também foi utilizado o pacote *qgraph* (EPSKAMP, 2012) para análise de rede. O código utilizado pode ser encontrado em: https://github.com/lemosvncs/educa-br-ufrj.

Resultados e Discussão:

De forma a tornar as informações obtidas mais claras, serão analisados somente relações significativas ($p < 0,05$) e com coeficiente de correlação maior que 0,2. As correlações podem ser encontradas nas tabelas 2, 3, 4 e 5 para correlações

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

entre escores do ENEM e turmas particulares, municipais, estaduais e federais, respectivamente.

Em relação às turmas de escolas particulares, nenhuma relação significante e relevante foi encontrada para as variáveis sobre turmas. Isto indica que as escolas não impactam as notas do ENEM, nem as disciplinas que estão diretamente ligadas às áreas do conhecimento. Das 80 relações observadas, apenas 24 obtiveram *p* menor que 0,05, indicando que não há diferença significativa entre o município ter uma alta ou baixa porcentagem de aulas e a nota do ENEM. Além disso, nas relações significativas, a correlação mais relevante foi entre Ciências Humanas e matemática, com *r* = -0,120.

Tabela 2

Correlações bivariadas entre turmas de escolas particulares e escores do ENEM, N = 1638

Turmas	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação
Física	-0.019	**-0.116*** **	-0.019	**-0.082*** **	-0.003
Matemática	-0.011	**-0.120*** **	0.026	**-0.073** **	0.005
Biologia	-0.017	**-0.117*** **	0.000	**-0.074** **	0.002
Português	-0.007	**-0.117*** **	-0.009	**-0.079** **	0.001
Inglês	-0.047	**-0.109*** **	-0.035	**-0.080** **	-0.007
Espanhol	-0.040	**-0.049***	**-0.082*** **	-0.041	-0.017
Francês	0.026	0.001	0.027	-0.011	-0.035
Outras línguas	-0.007	0.002	-0.015	0.012	0.046
Língua indígena	0.000	-0.004	-0.010	-0.006	-0.004
Artes	0.035	**-0.070** **	**0.068** **	-0.035	-0.042
Ed. Física	-0.016	**-0.090*** **	-0.017	**-0.054***	0.032
História	-0.013	**-0.119*** **	0.007	**-0.075** **	-0.001
Geografia	-0.015	**-0.117*** **	0.004	**-0.083*** **	0.001
Sociologia	-0.012	**-0.090*** **	-0.011	-0.061*	0.010
Filosofia	-0.021	-0.098***	-0.016	**-0.067** **	0.009
Ens. Religioso	0.046	0.007	-0.006	0.024	0.030

*Nota: *** p < 0,001; ** p < 0,01; * p < 0,05*

Em escolas municipais, somente 3 das 75 correlações observadas foram significativas, e duas delas relevantes: entre Linguagens e Códigos e ter outras línguas como disciplina (*r* = 0,37, *p* < 0,001) e Redação e ensino religioso (*r* = 0,23, *p* < 0,01).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Tabela 3

Correlações bivariadas entre turmas de escolas municipais e escores do ENEM, N = 156

Turmas	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação
Física	-0.077	0.038	0.023	-0.147	-0.023
Matemática	0.029	-0.035	0.116	0.068	-0.036
Biologia	-0.068	0.034	0.008	0.031	0.002
Português	0.028	-0.047	0.109	0.070	-0.031
Inglês	-0.138	-0.061	**-0.191***	-0.077	-0.025
Espanhol	-0.097	0.027	-0.114	-0.044	0.152
Francês	-0.019	-0.049	-0.015	-0.032	-0.032
Outras línguas	-0.025	0.044	**0.369*****	0.147	-0.021
Artes	0.075	0.020	0.102	-0.049	-0.073
Ed. Física	-0.117	0.080	-0.076	-0.073	0.006
História	-0.077	0.021	-0.037	-0.038	-0.001
Geografia	-0.071	0.011	0.005	0.032	0.003
Sociologia	-0.029	-0.033	0.089	0.120	-0.007
Filosofia	-0.036	-0.062	0.061	0.113	-0.042
Ens. Religioso	-0.026	0.061	-0.001	0.128	**0.235****

*Nota: *** $p < 0,001$; ** $p < 0,01$; * $p < 0,05$*

Nas escolas estaduais 40 das 80 relações observadas com $p < 0,05$. Sendo destas, a de maior relevância entre a porcentagem de aulas de sociologia e nota de redação com $r = 0,10$, indicando novamente, que não somente não há diferença entre a porcentagem de aulas ministradas e as notas do ENEM, como quando há, a diferença é diminuta.

Correlacionando-se os escores do ENEM com disciplinas diretamente relacionadas a estes, foi obtida correlação significante somente no escore de Linguagens e Códigos e a porcentagem de turmas de outras línguas, em escolas municipais ($r = 0,37$, $p < 0,001$). Outras correlações entre disciplinas diretamente ligadas às áreas de conhecimento do ENEM ou não foram significantes, ou não foram relevantes.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Tabela 4

Correlações bivariadas entre turmas de escolas estaduais e escores do ENEM, N = 5552

Turmas	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação
Física	**0.062*** **	0.007	**0.038** **	0.011	-0.020
Matemática	**0.055*** **	-0.014	**0.047*** **	-0.011	**-0.048*** **
Biologia	**0.063*** **	-0.016	**0.047*** **	-0.008	**-0.045*** **
Português	**0.063*** **	-0.016	**0.064*** **	-0.010	**-0.051*** **
Inglês	-0.015	**0.056*** **	-0.022	**0.069*** **	**0.093*** **
Espanhol	0.022	0.008	0.004	**-0.027***	**-0.032***
Francês	0.010	-0.010	0.022	-0.012	-0.008
Outras línguas	**0.040** **	-0.026	**0.039** **	-0.021	**-0.081*** **
Língua indígena	0.004	**-0.031***	0.009	**-0.034***	**-0.033***
Artes	-0.013	0.010	0.004	-0.005	**0.074*** **
Ed. Física	0.026	0.009	**-0.037** **	0.015	**0.034***
História	**0.071*** **	-0.005	**0.067*** **	-0.002	**-0.042** **
Geografia	**0.069*** **	**-0.029***	**0.070*** **	-0.019	**-0.056*** **
Sociologia	0.006	**0.049*** **	-0.011	**0.040** **	**0.098*** **
Filosofia	**-0.034***	0.024	**-0.050*** **	**0.035** **	**0.092*** **
Ens. Religioso	**0.061*** **	**0.063*** **	**0.029***	**0.029***	0.022

*Nota: *** p < 0,001; ** p < 0,01; * p < 0,05*

Tabela 5

Correlações bivariadas entre turmas de escolas federais e escores do ENEM, N = 331

Turmas	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação
Física	-0.026	**-0.167** **	**-0.181*** **	**-0.142** **	-0.062
Matemática	-0.047	**-0.144** **	**-0.211*** **	**-0.151** **	-0.092
Biologia	-0.007	**-0.219*** **	**-0.145** **	**-0.135***	-0.046
Português	-0.050	**-0.134***	**-0.225*** **	**-0.149** **	-0.096
Inglês	0.017	**-0.154** **	-0.102	-0.079	-0.063
Espanhol	-0.087	**-0.130***	**-0.130***	-0.064	**-0.160** **
Francês	-0.017	-0.023	-0.076	-0.025	-0.018
Outras línguas	-0.016	-0.026	-0.073	-0.029	-0.029
Artes	0.070	**-0.125***	-0.093	-0.087	0.013
Ed. Física	-0.014	-0.102	**-0.130***	-0.089	-0.028
História	-0.001	**-0.185*** **	**-0.138***	**-0.117***	-0.001

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Geografia	-0.008	**-0.203*** **	**-0.186*** **	**-0.118***	-0.044
Sociologia	-0.019	**-0.111***	**-0.181*** **	-0.087	-0.031
Filosofia	-0.018	**-0.156****	**-0.186*** **	**-0.109***	-0.096

*Nota: *** p < 0,001; ** p < 0,01; * p < 0,05*

Os resultados das correlações entre escores do ENEM e características de infraestrutura das escolas se encontram nas tabelas 6, 7, 8 e 9, com dados sobre escolas particulares, municipais, estaduais e federais, respectivamente. Em relação às escolas particulares e dados sobre infraestrutura, foi encontrada uma correlação significativa e relevante entre o escore matemática do ENEM e presença de internet (r = -0,21, p < 0,001).

Tabela 6

Correlações bivariadas entre infraestrutura de escolas particulares e escores do ENEM

Características de infraestrutura	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação	Número de casos
Biblioteca	-0.018	-0.006	**-0.052***	0.012	-0.019	
Sala de Leitura	-0.040	0.001	-0.001	**-0.055***	-0.005	
PNE	-0.004	-0.035	**-0.062***	-0.043	-0.032	
Refeitório	**0.076****	**0.049***	0.002	0.046	**0.058***	1661
Pátio Coberto	**-0.081*****	-0.042	-0.009	**-0.107*****	0.008	
Pátio descoberto	-0.039	-0.043	-0.047	-0.035	**0.080****	
Área verde	-0.001	0.034	-0.008	**-0.055***	**0.079****	
Internet	**-0.151*****	0.013	0.021	**-0.208*****	**0.086*****	1655
Banda larga	**-0.056***	0.000	-0.011	**-0.142*****	-0.003	1618
Alimentação	**0.080****	-0.010	-0.030	**0.083*****	-0.018	
Atendimento Educacional Especializado	0.005	0.007	0.003	0.000	-0.012	
Atividade complementar	-0.007	0.006	0.020	-0.026	**0.048***	
Não possui material especial	-0.022	-0.003	0.017	-0.044	**0.050***	1661
Alternância	**0.100*****	**0.080****	0.011	**0.189*****	0.004	
Água filtrada	**0.100*****	0.009	-0.004	0.008	**0.060***	
Esgoto inexistente	0.031	0.003	0.004	0.044	**0.061***	
TV	**-0.070****	-0.010	-0.009	-0.056*	**0.089*****	
DVD	-0.001	-0.057	**-0.083****	**0.088****	**-0.122*****	978
Copiadora	0.007	-0.008	-0.038	0.079*	-0.017	922

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Retroprojetor	0.012	-0.001	-0.011	0.027	-0.042	886
Impressora	0.034	-0.010	0.014	**0.107*** **	**-0.078***	995
Som	**0.099****	0.001	-0.021	**0.078***	-0.042	983
Multimedia	0.038	-0.004	-0.013	0.040	-0.012	963
Computador	0.022	0.010	-0.016	0.053	-0.003	1054

*Nota: *** $p < 0,001$; ** $p < 0,01$; * $p < 0,05$*

Nas escolas municipais, as correlações relevantes foram negativas entre Ciências da Natureza e escore z de retroprojetor (r = -0,41, $p < 0,05$) e equipamento multimídia (r = -0,44, $p < 0,05$).

Tabela 7

Correlações bivariadas entre infraestrutura de escolas municipais e escores do ENEM

Características de infraestrutura	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação	Número de casos
Biblioteca	-0.025	0.001	0.080	-0.025	0.029	
Sala de Leitura	0.055	-0.144	-0.103	0.006	0.009	
PNE	-0.083	-0.143	-0.088	0.037	0.034	
Refeitório	-0.101	**0.169***	0.044	0.045	0.115	164
Pátio Coberto	-0.131	**-0.166***	-0.072	-0.075	-0.146	
Pátio descoberto	-0.070	0.039	0.006	-0.050	0.151	
Área verde	-0.097	0.009	-0.032	0.030	0.040	
Internet	0.000	-0.001	-0.037	**-0.158***	0.091	
Banda larga	0.089	0.042	0.047	0.009	-0.041	155
Atendimento Educacional Especializado	-0.082	-0.048	-0.062	-0.071	**0.174***	
Atividade complementar	0.031	-0.138	-0.089	-0.062	0.032	
Não possui material especial	0.033	0.009	0.022	0.004	0.022	164
Alternância	-0.022	**0.153***	-0.015	0.116	-0.096	
Água filtrada	-0.003	0.017	0.034	**0.172***	**0.243****	
Esgoto inexistente	-0.010	0.023	-0.029	0.001	-0.023	
TV	0.000	0.046	-0.025	-0.091	0.012	
DVD	0.020	-0.142	-0.126	0.138	-0.023	
Copiadora	-0.049	0.093	-0.119	-0.107	0.004	36
Retroprojetor	**-0.408***	0.244	-0.124	-0.338	-0.155	29
Impressora	-0.148	0.122	-0.068	-0.202	-0.221	25
Som	-0.226	0.152	-0.180	-0.095	-0.229	41

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Multimedia	**-0.438***	0.060	-0.172	-0.239	-0.229	39
Computador	-0.117	0.148	-0.008	-0.120	0.070	33

*Nota: *** p < 0,001; ** p < 0,01; * p < 0,05*

45

Nas escolas estaduais não houveram relações relevantes e nas escolas federais elas foram negativas entre Ciências da Natureza e presença de biblioteca (r = -0,25, p < 0,001) e positivas entre Ciências da Natureza e educação por alternância (r = 0,20, p < 0,001) e equipamento multimídia (r = 0,66, p < 0,001).

Nas variáveis "presença de água filtrada" e "ausência de esgoto", não era esperado que houvesse diferença significativa entre nenhuma das categorias administrativas, com relação a nenhuma das notas, o que indicaria que estas características, básicas, não difeririam entre escolas. Mas a variável "água filtrada" foi significativa e relevante nas escolas municipais (em relação à Redação, r = 0,24, p < 0,001). Diferenças significativas, mas não relevantes ocorreram em todas as categorias administrativas.

Tabela 8
Correlações bivariadas entre infraestrutura de escolas estaduais e escores do ENEM

Características de infraestrutura	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação	Número de casos
Biblioteca	-0.006	0.005	-0.016	**-0.029***	-0.003	
Sala de Leitura	**-0.069*****	0.001	**0.047*****	**-0.061*****	**0.097*****	
PNE	-0.010	0.022	0.014	**-0.058*****	**0.057*****	
Refeitório	**-0.036****	**0.102*****	**0.106*****	**-0.111*****	**0.180*****	5552
Pátio Coberto	**-0.040****	-0.002	0.000	**-0.055*****	**0.067*****	
Pátio descoberto	**-0.032***	**0.113*****	**0.082*****	**-0.096*****	**0.181*****	
Área verde	**-0.088*****	**0.087*****	**0.084*****	**-0.120*****	**0.191*****	
Internet	-0.009	**0.029***	**0.049*****	**-0.080*****	**0.075*****	5549
Banda larga	-0.021	**0.031***	0.024	**-0.032***	**0.042****	5464
Atendimento Educacional Especializado	-0.018	**0.033***	0.021	**-0.093*****	**0.086*****	
Atividade complementar	**-0.058*****	0.024	0.023	**-0.094*****	**0.085*****	5552
Não possui material especial	-0.016	**0.039****	**0.043****	-0.014	**0.067*****	
Alternância	-0.002	-0.008	-0.019	0.018	-0.009	
Água filtrada	0.017	**0.075*****	0.010	-0.012	**0.094*****	

Esgoto inexistente	0.009	-0.018	-0.016	0.017	**-0.030***	
TV	**-0.060*****	0.001	-0.007	**-0.066*****	**0.094*****	
DVD	-0.013	-0.017	0.006	-0.024	**0.044***	2289
Copiadora	-0.047*	0.025	**0.050***	**-0.058****	**0.100*****	2194
Retroprojetor	-0.043	**0.114*****	**0.106*****	**-0.159*****	**0.198*****	1955
Impressora	-0.026	0.004	**0.086*****	**-0.130*****	**0.087*****	2317
Som	**-0.054***	**0.079*****	**0.123*****	**-0.148*****	**0.164*****	2179
Multimedia	**-0.059****	-0.036	0.018	**-0.107*****	0.018	2227
Computador	-0.004	0.019	**0.049***	**-0.053****	**0.063****	2480

*Nota: *** p < 0,001; ** p < 0,01; * p < 0,05*

Tabela 9
Correlações bivariadas entre infraestrutura de escolas federais e escores do ENEM

Características de infraestrutura	Ciências da Natureza	Ciências Humanas	Linguagem e Códigos	Matemática ENEM	Redação	Número de casos
Biblioteca	**-0.254*****	0.035	-0.008	0.023	0.011	
Sala de Leitura	-0.022	0.088	0.068	0.031	0.027	
PNE	-0.091	-0.037	-0.085	-0.032	-0.027	
Refeitório	-0.044	-0.094	-0.085	-0.061	-0.079	339
Pátio Coberto	-0.030	-0.068	0.018	-0.012	-0.058	
Pátio descoberto	-0.099	0.049	-0.031	-0.038	-0.008	
Área verde	-0.079	-0.042	-0.029	-0.063	0.039	
Internet	0.013	-0.044	0.018	**-0.109***	0.007	
Banda larga	**-0.182*****	**0.164****	0.066	0.029	0.087	334
Atendimento Educacional Especializado	-0.011	-0.012	0.002	-0.017	-0.021	
Atividade complementar	-0.016	-0.024	-0.035	-0.056	-0.037	339
Alternância	**0.201*****	**-0.147****	-0.039	-0.058	-0.021	
Água filtrada	-0.013	0.008	0.069	-0.044	-0.041	
TV	0.016	-0.044	-0.024	0.009	-0.022	
DVD	0.054	0.035	-0.075	-0.066	-0.075	43
Copiadora	-0.024	0.017	-0.071	-0.095	-0.047	
Retroprojetor	-0.121	0.031	0.269	-0.156	-0.086	39
Impressora	-0.128	-0.100	-0.091	-0.206	0.056	
Multimedia	-0.144	0.106	0.132	-0.269	**0.660*****	43
Computador	-0.123	-0.113	-0.038	-0.024	-0.043	

*Nota: *** p < 0,001; ** p < 0,01; * p < 0,05*

Conclusão:

As análises centraram-se em duas vertentes: 1) a porcentagem de aulas existentes sobre determinadas temáticas em um determinado município e seu impacto nas notas do ENEM e 2) A infraestrutura escolar e seu impacto nas notas do ENEM. Em relação à porcentagem de aulas, os resultados demonstram que seu impacto nas notas do ENEM é praticamente nulo. A maior parte das relações não foi estatisticamente significativa, indicando que não há relação entre as variáveis observadas. Isso pode ser explicado pela qualidade extremamente baixa da educação brasileira, como apontado pelo INEP (2003) e por Franco et al. (2007).

Somente as escolas estaduais demonstraram ter relações significativas em metade das disciplinas observadas, entretanto, o coeficiente de relação destas variáveis foi baixo demais para considerá-las fracas. Significando que mesmo nas relações estatisticamente significantes, o impacto das disciplinas é muito baixo para ter um impacto efetivo nos escores do ENEM.

Era esperado ter relações significativas entre todas as aulas e notas do ENEM, com plausíveis ausências pontuais, indicando um possível desfalque na educação. Também era esperado que tivéssemos relações medianas a fortes. Mas o que encontramos foi um quadro onde não há diferença entre um município com alta porcentagem de escolas que ofereçam as disciplinas e a nota do ENEM, essa diferença não existe nem em relação a disciplinas que são diretamente ligadas aos campos de conhecimento do ENEM.

Em relação à infraestrutura, era esperado que os dados nos mostrassem disparidade de infraestrutura entre os municípios. Os resultados demonstram que há algum impacto da infraestrutura sobre as notas, indicando que a hipótese de que fatores intra-escolares são mais relevantes quando não há uma base mínima adequada à educação é plausível. Isso se faz evidente quando pegamos, por exemplo, a presença de água filtrada nas escolas municipais, com relação positiva nas escolas municipais. A presença de água filtrada seria irrelevante se todas as escolas a possuíssem, só se fazendo notar pois há um número significativo de escolas sem água-filtrada disponível aos alunos. Nas escolas municipais a presença de retroprojetores e equipamento multimídia tem um impacto negativo na nota do ENEM.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Isso indica que há uma diferença de infraestrutura entre as escolas municipais. Mas a presença destes equipamentos impacta negativamente as notas de Ciências da Natureza. É possível que, nestes casos, a qualidade do ensino esteja modulando a presença dos equipamentos.

Sobretudo, os dados não representam uma diferença clara entre municípios, visto que as únicas variáveis relevantes em infraestrutura foram a presença de retroprojetor, equipamento multimídia e internet. Estas são variáveis que indicam diferença estrutural nas escolas, mas outras variáveis que deveriam representar grande diferença, como banda-larga, área-verde e sala de leitura não tem diferenças significativas. Isto pode ser um indicativo de que há mais fatores contribuindo para as diferenças e indiferenças educacionais, que não a qualidade do ensino ou infraestrutura. Também é possível que dados do questionário do Censo da Educação sejam enviesados, através de preenchimentos realizados de maneira não adequada, ou que o ENEM não seja uma maneira adequada de se avaliar a educação, gerando relações espúrias.

Referências:

ALEKSANDAR, B. Corstar.r. Disponível em: <https://gist.github.com/aL3xa/887249>. Acesso em 09 de março de 2017.

BRASIL. *Constituição da República Federativa do Brasil*. 1988. Disponível em: <https://legislacao.planalto.gov.br/legisla/legislacao.nsf/viwTodos/509f2321d97cd2d203256b280052245a?OpenDocument&Highlight=1,constitui%C3%A7%C3%A3o&AutoFramed>. Acesso em 09 de março de 2017.

BRASIL. DECRETO Nº 6.425, DE 4 DE ABRIL DE 2008.

EPSKAMP, S.; COSTANINI, G.; HASLBECK, J. ANGELIQUE, O. J. C.; WALDORP, L. J.; SCHIMITTMANN, V. D.; BORSBOOM. qgraph: Network Visualizations of Relationships in Psychometric Data. Journal of Statistical Software, v. 48, n. 4, p.1-18. 2012. Disponível em: <http://www.jstatsoft.org/v48/i04/>. Acesso em: 09 mar. 2017.

FRANCO, C.; ORTIGÃO, I.; ALBERNAZ, A.; BONAMINO, A.; AGUIAR, G.; ALVEZ, F.; SÁTYRO, N. Qualidade e equidade em educação: reconsiderando o significado de

"fatores intra-escolares". *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v.15, n. 55, p.277-298, abr./jun. 2007.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Microdados do Exame Nacional do Ensino Médio 2014. Disponível em: <http://dados.gov.br/dataset/microdados-do-exame-nacional-do-ensino-medio-enem/resource/94731b73-e9f1-4262-b8d6-479b6d02a6f0>. Acesso em 28 de fevereiro de 2017.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Microdados do Censo Escolar 2014. Disponível em: <http://dados.gov.br/dataset/microdados-do-censo-escolar/resource/809f2de5-1d3b-448b-8485-eb238492919a>. Acesso em 28 de fevereiro de 2017.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Edital nº 12, de 8 de Maio de 2014. Exame Nacional do Ensino Médio. 2014.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. ENEM. Disponível em: <http://portal.inep.gov.br/web/guest/enem>Acesso em 27 de fevereiro de 2017a.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo Escolar. Disponível em: <http://portal.inep.gov.br/web/guest/censo-escolar> Acesso em 27 de fevereiro de 2017b.

MUKAKA, M. M. Statistics Corner: A guide to appropriate use of Correlation Coefficient in medical research. Malawi Medical Journal, 2012.; v. 24(3), p. 69-71, set. 2012.

NAKANO, T. C.; PRIMI, R.; NUNES, C. H. S. Análise de itens e Teoria de Resposta ao Item. In: HUTZ, C. S.; BANDEIRA, D. R.; TRENTINI, C. M. *Psicometria*. Porto Alegre: Artmed, 2015. p. 97-123.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria: 2016. Disponível em:< https://www.R-project.org/>. Acesso em: 09 mar. 2017.

RAVELLE, W. Psych: Procedures for Personality and Psychological Research. Northwestern University, Evanston, Illinois: 2016. Disponível em: < http://CRAN.R-project.org/package=psych>. Acesso em: 09 mar. 2017.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 46 - QUALITATIVE *PANEL* – DESENVOLVIMENTO DE UMA APLICAÇÃO EM *SHINY* PARA ANÁLISE INTERATIVA DE DADOS SENSORIAIS.

Adson Costanzi Filho[42]
Flaviane Peccin Brevi[43]
Gabriel Martins Brock[44]
Rodrigo Oliveira da Fontoura[45]

Resumo

Após a criação e disponibilização do pacote *shiny* o uso de interfaces gráficas para a geração e visualização de informações tem sido amplamente utilizada tanto no meio acadêmico quanto no empresarial. Outra *library* denominada *plotly* vem tornando a experiência gráfica ainda mais interativa através de uma sinergia entre o usuário e os resultados. Tendo em vista tais avanços, este estudo propõe demonstrar os principais conceitos para a construção de uma ferramenta utilizando estes pacotes, além de sua utilização no meio corporativo, especificamente na análise de dados sensoriais. A ferramenta apresentada neste artigo mostrou-se eficaz, pois contribuiu para agilizar o processamento de dados e a disponibilização de informações de suma importância para a tomada de decisão na corporação.

Palavras-Chave: Shiny, Plotly, Interface, Análise Sensorial

Abstract

After its creation and availability, *shiny* package has been widely used for graphic interfaces generation and information visualization, in academia and business. Another library called *plotly* has been making the graphic experience even more interactive through a synergy between users and results. In view of such advances, this study proposes to demonstrate the main concepts for the construction of a tool using those packages and their use in a corporate environment, specifically in sensory data analysis. The tool presented in this paper has shown to be effective, because it contributed to streamline data processing and the availability of important information for corporation decision making.

Keywords: Shiny, Plotly, Interface, Sensory analysis

Introdução:

O surgimento da linguagem R no ano 1993[1][2][3], vem aproximando o uso da estatística vinculada à computação. Esse avanço proporciona uma maior interação para interessados no assunto, além de maior velocidade e praticidade na obtenção e entrega de resultados.

[42] British American Tobacco – adson_filho@souzacruz.com.br
[43] British American Tobacco – flaviane_brevi@souzacruz.com.br
[44] British American Tobacco – gabriel_brock@souzacruz.com.br
[45] British American Tobacco – rodrigo_fontoura@souzacruz.com.br

O IDE (*integrated development environment*) *R-Studio* criado em 2011[4], é um advento que contribuiu ainda mais para tal. O desenvolvimento deste ambiente trouxe uma interface de programação mais amigável por apresentar diversos atalhos, funções de auto completar, identação e sinalizações de erros e inconsistências antes mesmo da execução do programa.

Em 2012, os mesmos criadores do *R-Studio* contribuíram mais uma vez para a difusão da estatística no mundo através da disponibilização do pacote *shiny*[5]. Este, integra a programação HTML com funções da linguagem R, com uma sinergia que proporciona a criação de interfaces visuais intuitivas, fluidas e dinâmicas para usuários que não precisam ter, necessariamente, conhecimento sobre programação ou estatística.

Através da interface criada a partir do *shiny*, também é possível apresentar gráficos interativos ao usuário como os gerados pelo pacote *plotly*, desenvolvido pela companhia de mesmo nome em 2015[6]. Estes gráficos podem ser facilmente manuseados, possibilitando uma experiência visual ainda melhor.

Ao integrar as funcionalidades das ferramentas descritas acima, foi possível a criação de uma interface denominada *Qualitative Panel* para ser utilizada na tomada de decisão no meio corporativo. Essa ferramenta possibilitou a aproximação da estatística ao negócio aumentando a acurácia e a velocidade da obtenção de resultados.

Objetivo:

Este artigo tem como objetivo apresentar uma aplicação criada através da linguagem R, principalmente utilizando os pacotes *shiny* e *plotly*, no meio corporativo. Será demonstrado como uma interface visual pode auxiliar e agilizar a tomada de decisão no negócio.

Material e Métodos:

Aplicações em *shiny* são usualmente compostas por duas partes. Uma contém as informações sobre a interface gráfica que será visualizada pelo usuário (*user interface – ui*) e outra que é composta de funções e rotinas de programação (*server*). A *ui* é responsável por receber *inputs* do usuário, como dados e ações. Posteriormente devolverá *outputs* (gráficos, tabelas e resultados de análises) gerados a partir das rotinas e funções executadas pelo *server*.

A ferramenta *Qualitative Panel*, apresentada neste artigo, tem como objetivo processar informações sensoriais sobre avaliações de atributos em produtos para depois quantificar e ilustrar a distância existente entre eles. Os atributos avaliados podem ser os mais diversos, incluindo sensações (aroma, identidade visual, sabor, toque ou som) e são medidos em escalas ordinais representados numericamente[7]. Uma vez obtidas essas informações, torna-se possível computar as distâncias euclidianas entre as matrizes, além das medidas de similaridade e dissimilaridade dos produtos. Estes valores são representados graficamente através de diagramas de dispersão de duas ou três dimensões.

Desta forma, o usuário deve fornecer como *inputs*: uma base de dados com um formato pré-determinado e informações sobre quais atributos avalidos irão compor cada eixo. Como *output*, a ferramenta entrega os gráficos que ilustram a distância sensorial entre produtos. O pacote plotly torna possível criar estes gráficos de dispersão de uma maneira que esses possam ser facilmente manuseados pelo usuário, com vantagens que incluem: aplicar *zoom*, girar a visualização e obter informações sobre os pontos do gráfico ao apontar o cursor sobre eles.

Essa ferramenta está disponibilizada em um servidor que pode ser acessado dentro do domínio da companhia, tornando sua utilização e manutenção descomplicada.

Resultados e Discussão:

Ao executar o programa a interface apresentada ao usuário é a que segue abaixo:

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

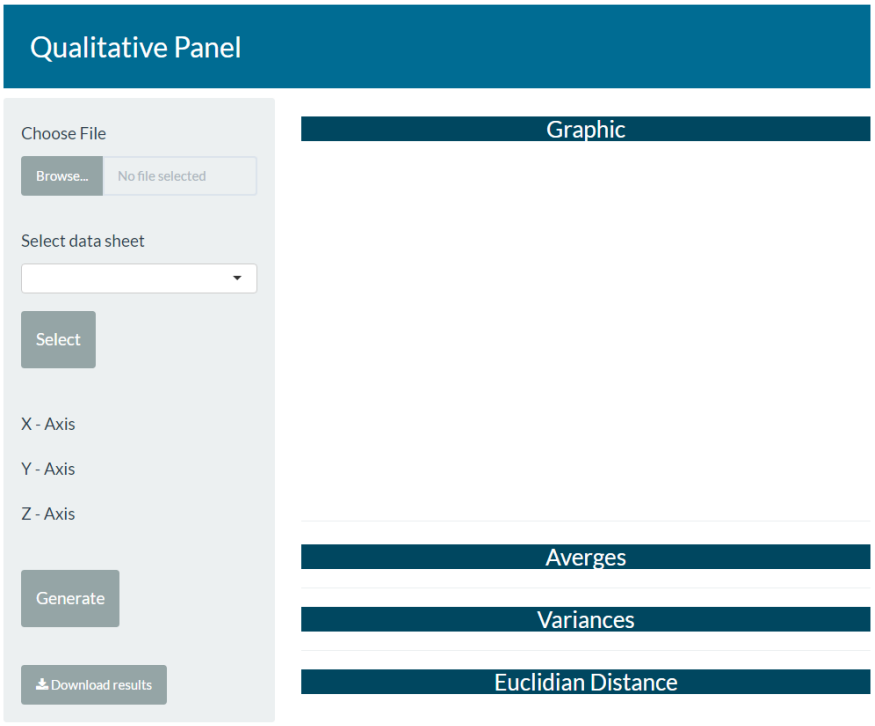


Figura 1 – Layout inicial do *Qualitative Panel*

Existem, basicamente, dois campos distintos: um menu direcionado a interação usuário-interface (barra lateral esquerda) e outro campo que apresentará os resultados (quadro à direita).

A ferramenta é *user friendly*, com botões intuitivos ao usuário que, de uma maneira sequencial, auxilia como proceder com os *inputs* requeridos. Primeiramente é necessário que uma base de dados em formato *excel* seja carregada. Esta base deverá ter uma formatação específica, acordada previamente com o usuário. Após carregada a base de dados, *checkboxes* surgirão no menu à esquerda com as opções de atributos que posteriormente farão parte dos eixos, conforme figura 2.

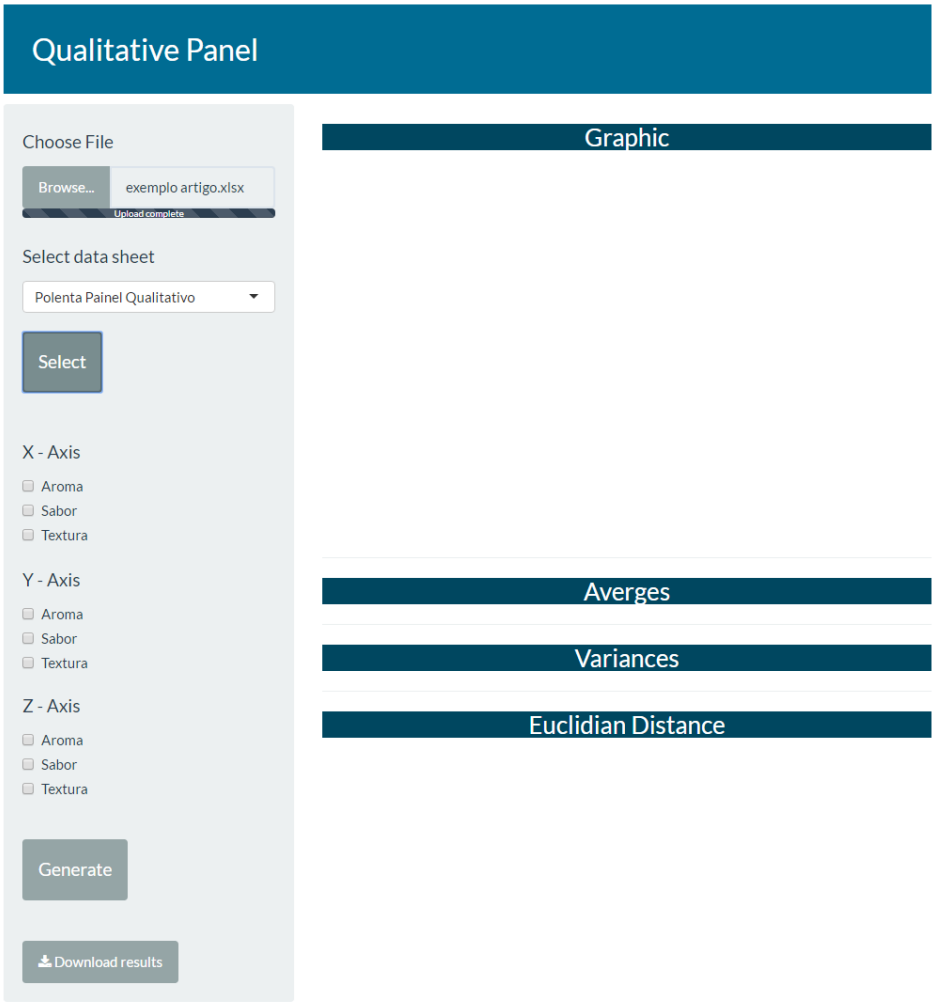


Figura 2 – *Checkboxes* na interface do *Qualitative Panel*

É possível ao usuário selecionar múltiplos atributos em um mesmo eixo, que serão combinados para o posicionamento dos produtos no espaço dimensional. A partir das seleções feitas para compor cada eixo, as rotinas contidas no *server* executarão analises especificas. Exemplificando, se o usuário selecionar opções apenas para o eixo X e Y a rotina retornará um gráfico em duas dimensões. De outra forma, se forem selecionadas opções para os eixos X, Y e Z o *output* será um gráfico em três dimensões.

Como já descrito anteriormente, a utilização do pacote plotly para a construção gráfica desta interface permite ao usuário interações com apenas um passo.

Junto ao gráfico, são apresentadas medidas descritivas (média e variância) das variáveis sensoriais avaliadas em cada produto, além de um diagrama representando as distâncias euclidianas entre os pontos considerando todos os atributos presentes na base de dados.

Para exemplificar o uso do *Qualitative Panel,* será utilizada uma base de exemplo contendo informações sensoriais relativas a polenta frita. Foram avaliados sete produtos quanto a sua aparência, sabor e textura. Na figura 3 está representado o gráfico da dispersão espacial dos produtos em relação aos atributos avaliados.

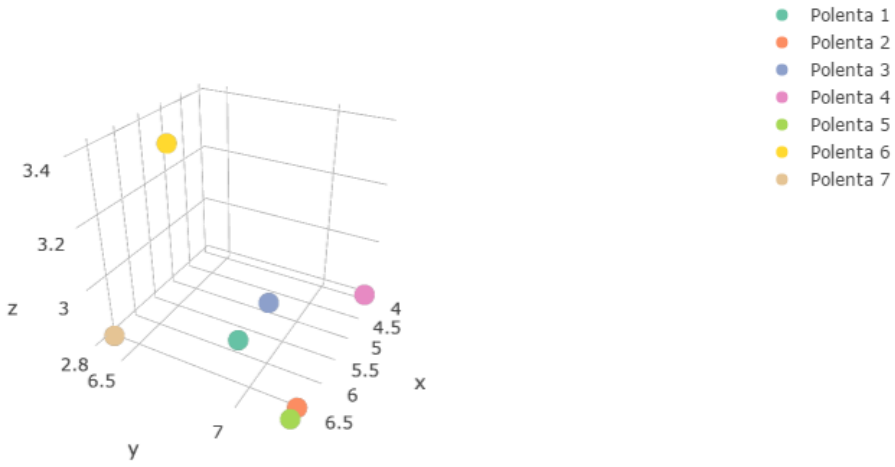


Figura 3 – Diagrama de dispersão *3D* das marcas de Polenta

No gráfico em três dimensões, os eixos X, Y e Z representam respectivamente as variáveis aroma, sabor e textura. Pode-se ver que os produtos quatro e seis, marcados em amarelo e marrom respectivamente, se destacaram dos demais, dado que apresentaram distâncias maiores. Os produtos dois e cinco (laranja e verde) mostraram um comportamento muito semelhante em relação às variáveis medidas.

A figura 4 confirma o descrito anteriormente.

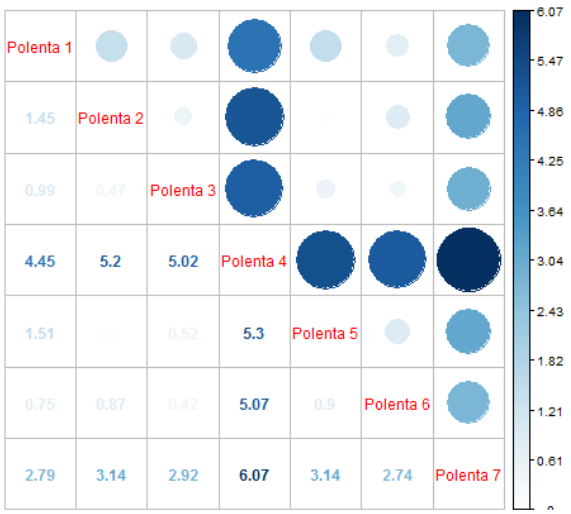


Figura 4 – Distâncias euclidianas entre produtos

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

A tabela de médias e variância (tabela 1) auxilia a avaliar o desempenho geral de cada produto. Por exemplo, se o objetivo for lançar um produto com a melhor textura decisão recomendada seria lançar o produto seis.

Tabela 1 – Medidas Descritivas das polentas

Produto	Média			Variâncias		
	Aroma	Textura	Sabor	Aroma	Textura	Sabor
Polenta 1	6.8	3	7	1.2	1.5	1.5
Polenta 2	6.6	2.8	7.2	0.3	0.7	0.7
Polenta 3	6	3	7	0.5	1	1
Polenta 4	4	2.8	7.2	5.5	0.7	0.7
Polenta 5	6.8	2.8	7.2	0.2	0.7	0.7
Polenta 6	6	3.4	6.6	0.5	1.3	1.3
Polenta 7	6.6	2.8	6.4	0.3	0.2	3.8

No apêndice se encontra um panorama dos *outputs* da ferramenta em execução.

Conclusão:

O entendimento dos atributos sensoriais é estratégico para assegurar o sucesso do lançamento de novos produtos no mercado, bem como avaliar o posicionamento dos mesmos em relação a concorrência.

A ferramenta demonstrada neste estudo contribuiu de forma relevante para o setor de analises sensoriais, pois agilizou o processamento de dados e a disponibilização de informações indispensáveis para a tomada de decisão na corporação. Isso só foi possível pela eficiente integração dos pacotes *shiny* e plotly, que tornam a experiência fluida e interativa.

Referências:

[1] STIGLER, Stephen M. ***The History of Statistics: The Measurement of Uncertainty Before 1900.*** The Belknap Press of Harvard University Press. Cambridge, USA, 1986.

[2] CORDEIRO, Gauss M. **O Amadurecimento da Pesquisa e Ensino de Estatística no Brasil**. arScientia, 2006.

[3] HUBER, Peter J.; RONCHETTI, Elvezio M. ***Robust Statistics.*** 2. ed. Hoboken, USA: Wiley, 2009. 38 p. Disponível em:

<http://samples.sainsburysebooks.co.uk/9780470434680_sample_410695.pdf>. Acesso em: 15 mar. 2017.

[4] ***About – RStudio.*** Disponível em: < https://www.rstudio.com/about/> Acesso em: 15 mar. 2017.

[5] ***R Project: Available CRAN Packages By Date of Publication.*** Disponível em: <https://cran.r-project.org/web/packages/available_packages_by_date.html> Acesso em: 15 mar. 2017.

[6] ***R Project: Available CRAN Packages By Date of Publication.*** Disponível em: <https://cran.r-project.org/src/contrib/Archive/plotly> Acesso em: 15 mar. 2017.

[7] Dutcosky, S. D. (2013). Análise Sensorial de alimentos (4 ed.). Curitiba: PUCPRess.

Apêndice:

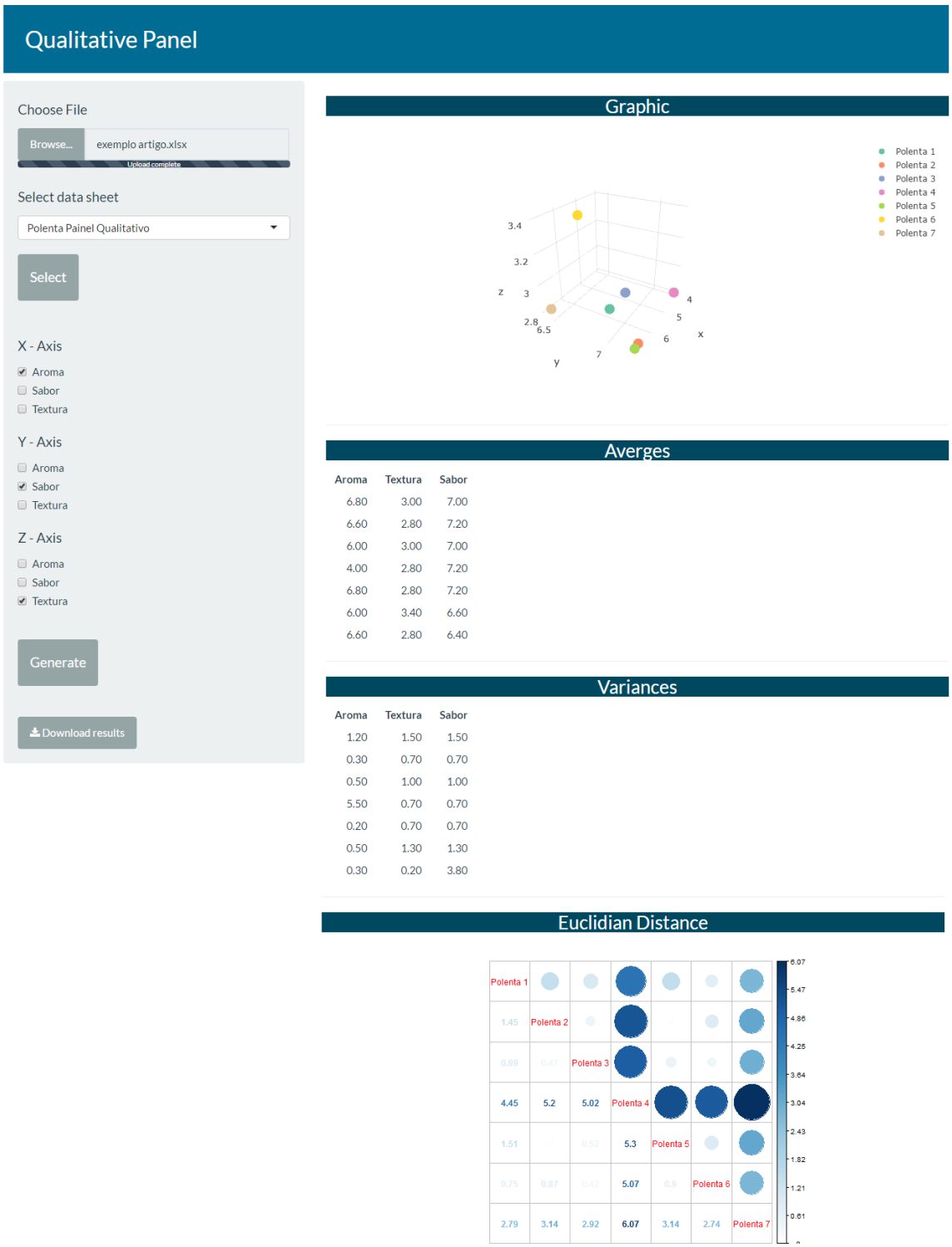


Figura 6 - *Overview*

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 47 - HIERARQUIZAÇÃO DOS FATORES DE ATRASO EM OBRAS PÚBLICAS NA REGIÃO SUL FLUMINENSE COM BASE NA OPINIÃO DOS GESTORES

Alessandra Simão[46]
Luciane Ferreira Alcoforado[47]
Ariel Levy[48]
Leonardo Filgueira[49]

Resumo

A problemática de atrasos em obras públicas é um fenômeno global. A identificação dos fatores de atraso favorece o estabelecimento de ações corretivas para os pontos identificados como os mais críticos e possibilita as ações corretivas permanentes. Dessa forma, questiona-se: Quais são os fatores de atraso mais críticos, na percepção dos gestores, de acordo com o método AHP para escolha de medidas mitigatórias em obras públicas? Assim, especificamente este artigo objetiva: Identificar na perspectiva de especialistas, gestores públicos e privados da construção civil a importância das categorias e seus perspectivos fatores de atrasos na construção de obras públicas com o método AHP. Como procedimento metodológico adotou-se pesquisa exploratória e abordagem quantitativa, com aplicação do questionário de DOLOI ET AL (2012). Como principal resultado, s causas mais significativas dos atrasos identificadas são relacionadas aos fatores R26 Má gestão/supervisão do local; R 4. Retrabalho devido à mudança de desempenho ou desacordo de ordem, R40. Má produtividade do trabalho, fator R18. Demora na aprovação do trabalho completo por parte do cliente (isto é, mudança de estágio), R5. Condições severas de tempo/clima e R16. Especificações imprecisas da condição do local e R34 Mudanças em leis ou em regulamentação governamental.

Palavras-Chave: Obra Públicas, Atrasos, AHP

Abstract

The problem of delays in public works is a global phenomenon. The identification of the delay factors favors the establishment of corrective actions for the points identified as the most critical and enables the permanent corrective actions. Thus, we ask: What are the most critical delay factors, in the managers' perception, according to the AHP method to choose mitigation measures in public works? Thus, specifically, this article aims to: Identify from the perspective of specialists, public and private managers of the construction industry the importance of the categories and their perspectives factors of delays in the construction of public works with the AHP method. As a methodological procedure we adopted exploratory research and quantitative approach, with application of the questionnaire DOLOI ET AL (2012). As the main result, the most significant causes of delays identified are related to factors R26 Poor site management / supervision; R 4. Rework due to change of performance or order disagreement, R40. Poor labor productivity, factor R18. Delays in the approval of the complete work by the client (that is, change of stage), R5. Severe weather / climate conditions and R16. Inaccurate location condition specifications and R34 Changes in law or government regulation.

[46] UFF – Programa de Pós Grad. Eng. Civil - alessandra_simao@id.uff.br
[47] UFF - Programa de Pós Grad. Eng. Civil / Dep. Estatística – lucianealcoforado@gmail.com
[48] UFF – Programa de Pós Grad. Administração / Dep. Administração – alevy@id.uff.br
[49] UFF – Grad. Estatística – leo-filgueira@hotmail.com

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Keywords: Public Works, Delays, AHP

Introdução

O problema de atrasos em projetos de construção civil é um fenômeno global. A necessidade de identificar os fatores de atraso favorece ao estabelecimento de ações corretivas para os pontos identificados como os mais críticos e possibilita a obtenção de ações corretivas permanentes (AL-KHARASHI e SKITMORE, 2009).

As empresas construtoras precisam aperfeiçoar os sistemas de gestão de forma a minimizar o problema de não cumprir o prazo estabelecido, otimizando dois importantes aspectos: o tempo e o custo. Neste contexto, é de extrema importância o mapeamento dos fatores de atraso em obras públicas, uma vez que somente é possível desenvolver planos de mitigação para os fatores identificados.

A utilização do AHP (*Analytic Hierarchy Process)* para hierarquizar os fatores de atraso mais relevantes pode possibilitar medidas mais eficientes para reduzir o problema, tornando este processo mais eficiente, racional e claro, pois assim, as decisões são tratadas de forma matemática, minimizando os erros agregados, e não simplesmente de forma subjetiva (SHIMIZU, 2006).

Dessa forma, levanta-se o seguinte questionamento: Quais são os fatores de atraso mais críticos, na percepção dos gestores, de acordo com o método AHP para escolha de medidas mitigatórias em obras públicas?

Objetivo

Identificar na perspectiva de especialistas, gestores públicos e privados da construção civil a importância das categorias e seus perspectivos fatores de atrasos na construção de obras públicas com o método AHP.

Material e Métodos:

O estudo utilizou-se de pesquisa exploratória com abordagem quantitativa, sendo desenvolvida com a aplicação dos fatores de atraso apontados por (DOLOI ET AL., 2012).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Quadro 1 – Identificação dos atributos e suas referências

Categoria	Atributos que afetam o atraso	Referências bibliográficas
Relacionada ao projeto	R1. Aumento do escopo do trabalho	Semple *et al*. (1994); Sambasivan e Soon (2007);
	R2. Ambiguidade nas especificações e interpretações conflitantes das partes	Satyanarayana e Iyer (1996)
	R3. Relatório de investigação do solo falho	
	R4. Retrabalho devido à mudança de desenho ou desacordo de ordem	
	R5. Cronograma fora da realidade imposto no contrato	
	R6. Indisponibilidade de desenho/*design* em tempo hábil	
	R7. Retrabalho devido a erro de execução	
Relacionada ao local	R8. Acesso restrito ao local	Aibinu e Odeyinka (2006); Lo *et al*. (2006); Satyanarayana e Iyer (1996)
	R9. Condições severas de tempo/clima	
	R10. Tomada de decisão lenta por parte do dono (ou contratante?)	
	R11. Demora na entrega de material por parte do fornecedor	
	R12. Acidentes no local devido à negligência	
	R13. Acidentes no local devido à falta de medidas de segurança	
	R14. Condições imprevistas do terreno	
	R15. Condições políticas hostis	
	R16. Especificações imprecisas da condição do local	
Relacionada ao processo	R17. Demora na entrega de suprimentos por parte do dono (ou contratante?)	Iyer e Jha (2005); Satyanarayana e Iyer (1996);
	R18. Demora na aprovação do trabalho completo por parte do cliente (isto é, mudança de estágio)	
	R19. Demora na aquisição do material por parte da empreiteira	
	R20. Demora na compra de amostras e desenhos	
	R21. Demora no pagamento de contas à empreiteira	
	R22. Demora na entrega do local	
	R23. Demora em finalizar taxas para itens extras	
	R24. Condições impróprias de estocagem, danificando os materiais	
Relacionada a recursos humanos	R25. Resistência à mudança por parte do arquiteto ou consultor	Iyer e Jha (2005); Satyanarayana e Iyer (1996);
	R26. Má gestão/supervisão do local	Sambasivan e Soon (2007)
	R27. Conflito entre donos e outras partes	
	R28. Falta de mão-de-obra qualificada para equipamentos específicos	
	R29. Má coordenação entre grupos	
	R30. Mudança frequente de sub-empreiteiros(?)	
Relacionada a autoridades	R31. Obter permissão das autoridades locais	Assaf *et al*. (1995); Iyer e Jha (2005); Satyanarayanae Iyer (1996);
	R32. Burocracia na organização do cliente	
	R33. Má estrutura organizacional do cliente ou consultor	
	R34. Mudanças em leis ou em regulamentação governamental	
	R35. Falta de controle em relação à sub-empreiteira	
	R36. Formas de contratação inadequadas	
Questões técnicas	R37. Falta de motivação para os empreiteiros terminarem antes do prazo	Chan e Kumaraswamy (1997); Sambasivan e Soon (2007);
	R38. Planejamento impróprio da empreiteira durante o estágio de licitação	Faridi e El-Sayegh (2006)
	R39. Impasses financeiros dos empreiteiros	
	R40. Má produtividade do trabalho	
	R41. Experiência inadequada do empreiteiro	
	R42. Mudanças nos preços dos materiais ou no levantamento dos preços	
	R43. Uso ineficiente dos equipamentos	
	R44. Uso de métodos de construção obsoletos ou impróprios	
	R45. Métodos impróprios de inspeção e testes propostos no contrato	

Fonte: Doloi et al (2012).

Inicialmente, foi aplicado um questionário com os fatores de atraso apontados por de (DOLOI ET AL., 2012) aos gestores públicos nas Secretarias de Obras dos

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

municípios da Região Sul Fluminense (Volta Redonda, Barra Mansa, Resende, Porto Real, Quatis e aos gestores de empresas construtoras sediadas na mesma região. Utilizou-se a escala de Likert* com a pontuação nos valores de 1 (menos importante) até 5 (muito importante), para verificar o grau de importância atribuída aos fatores de atraso nas obras públicas da região, cujas respostas foram representadas graficamente em (SIMÃO ET AL.,2016), destacando-se especialmente o fator R40: Má produtividade do trabalho entre os fatores mais importantes nos atrasos em obras, segundo os especialistas.

Em um segundo momento, aplicou-se um novo questionário a especialistas do programa de Pós-Graduação em Engenharia Civil – UFF, para verificar qual categoria de atributos apontados por (DOLOI ET AL., 2012) que afetam o atraso em obras é mais relevante sob a percepção deles.

Após a coleta dos dados, estes foram analisados pelo software R adotando o método AHP para hierarquização dos fatores de atraso mais relevantes na opinião dos especialistas e dos gestores através da implementação de uma função auxiliar para transformar as notas dos especialistas na matriz de comparação dos critérios:

Resultados e Discussão:

A modelagem do problema conforme a metodologia AHP envolve, a hierarquização do objetivo global, critérios, subcritérios e as alternativas (GOMES ET AL., 2004). Contudo, o objetivo deste trabalho é hierarquizar os fatores de atraso em obras públicas, e dessa forma, o foco de análise do problema é representado na Figura 1, destacado na parte tracejada, onde é estabelecido a hierarquia entre os critérios e subcritérios.

* Escala de Likert – Desenvolvido por Rensis Likert (1932) para mensurar atitudes no contexto das ciências comportamentais. A escala de verificação de Likert consiste em tomar um construto e desenvolver um conjunto de afirmações relacionadas à sua definição, para as quais os respondentes emitirão seu grau de concordância.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

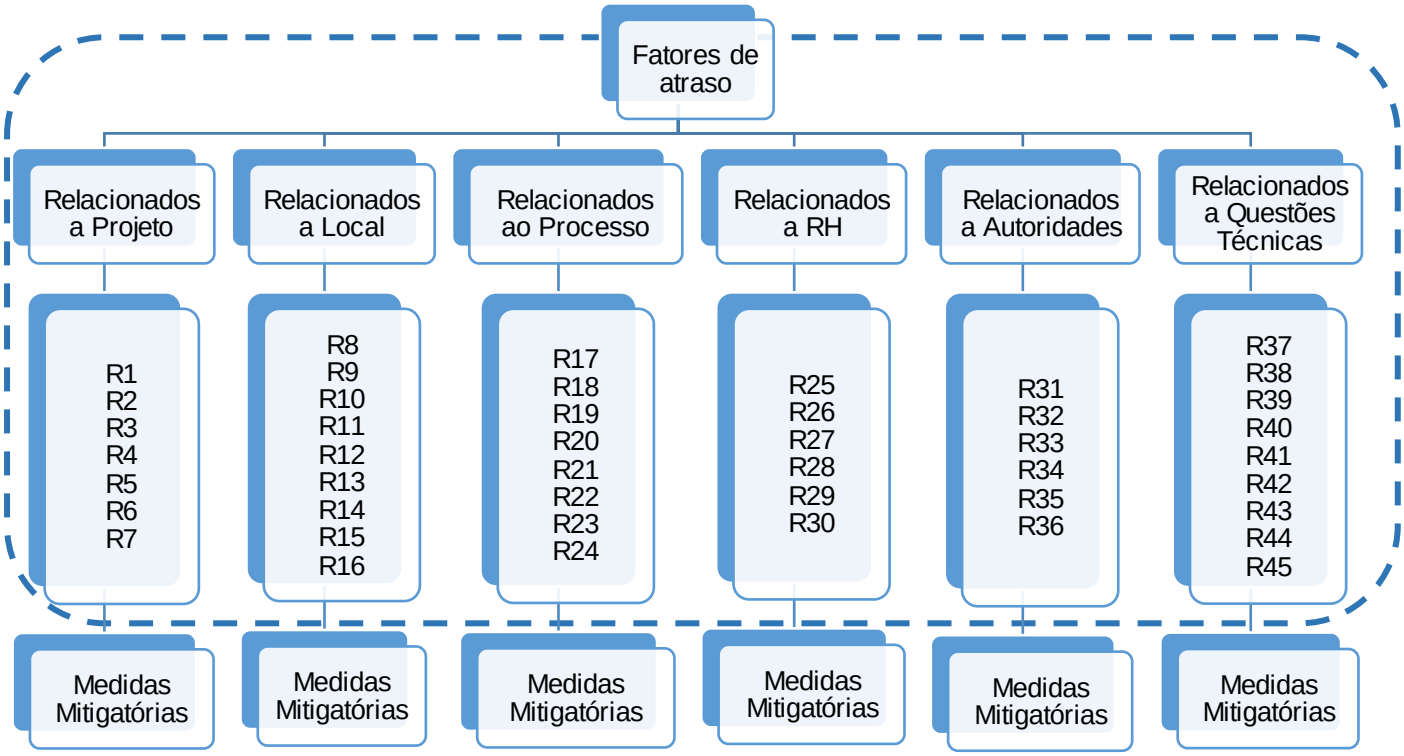


Figura 1: Estrutura Hierárquica do problema Fonte: Autores (2017)

Foi elaborada a matriz de comparação entre os critérios (categorias de atraso), utilizando a Escala Fundamental de Saaty (SAATY, 2008), de acordo com a percepção dos especialistas. E posteriormente o teste de consistência, no qual foi calculado autovalor da matriz de comparação, λ_{max}; o índice de consistência, *IC* e a taxa de consistência, *CR*. Conforme (SHIMIZU,2006) “com uma taxa de consistência de 0,10 ou menos é considerada aceitável”.

$$\mathrm{IC} = \frac{(\lambda\ \mathrm{max} - \mathrm{n})}{\mathrm{n} - 1} \tag{12}$$

A matriz de comparação dos 6 grupos de fatores (critérios) apresentou consistência aceitável, conforme Tabela 2.

Tabela 2 – Verificação da Consistência da Matriz

lambda	IC	RC
6.081125	0.016225	0.0130846

Fonte: Autores (2017)

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

A Tabela 3 apresenta o Vetor dos Pesos (que representa o grau de importância ou ordenamento) de cada grupo de fatores de atraso. Conforme a Tabela 3, o critério Recursos Humanos (RH) foi o de maior relevância.

Tabela 3 – Pesos dos Grupo de Fatores de atraso

Critérios	**pesos**
Projeto	0.1994323
Local	0.1130566
Processo	0.1622495
RH	0.2261131
Autoridades	0.0997162
Quest_Técnicas	0.1994323

Fonte: Elaborado pelos autores

Para cada subcritério (fator de atraso), foi seguido os mesmos passos: i) A construção da matriz de comparação; ii) Obtenção da prioridade relativa; iii) Verificação da consistência das prioridades relativas; iv) Obtenção do vetor de consistências; e v) Cálculo da taxa de consistência.

Para simplificação do problema, apresenta-se resumidamente, os resultados da Taxa de Consistência (CR) de cada fator de atraso, conforme a Tabela 4.

Tabela 4 – Verificação da consistência das matrizes de subcritérios

Grupos de Fatores	lambda	IC	RC
Relacionada ao projeto	7.227659	0.0379431	0.0287448
Relacionada ao local	9.138093	0.0172617	0.0119046
Relacionada ao processo	8.253963	0.0362804	0.0257308
Relacionada a recursos humanos	6.05679	0.011358	0.0091597
Relacionada as autoridades	6.107243	0.02114486	0.0172973
Relacionada as Questões técnicas	9.083822	0.0104778	0.0072261

Fonte: Autores (2017)

Os dados foram considerados consistentes em todas as comparações de pares para os fatores de atraso, pois os quocientes de consistência encontrados foram menores que 0,1, ou seja, menor que 10%.

A Figura 2 sintetiza os resultados do estudo com as posições do critério e subcritérios de maior relevância, na perspectiva dos especialistas, gestores do setor público e privado.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

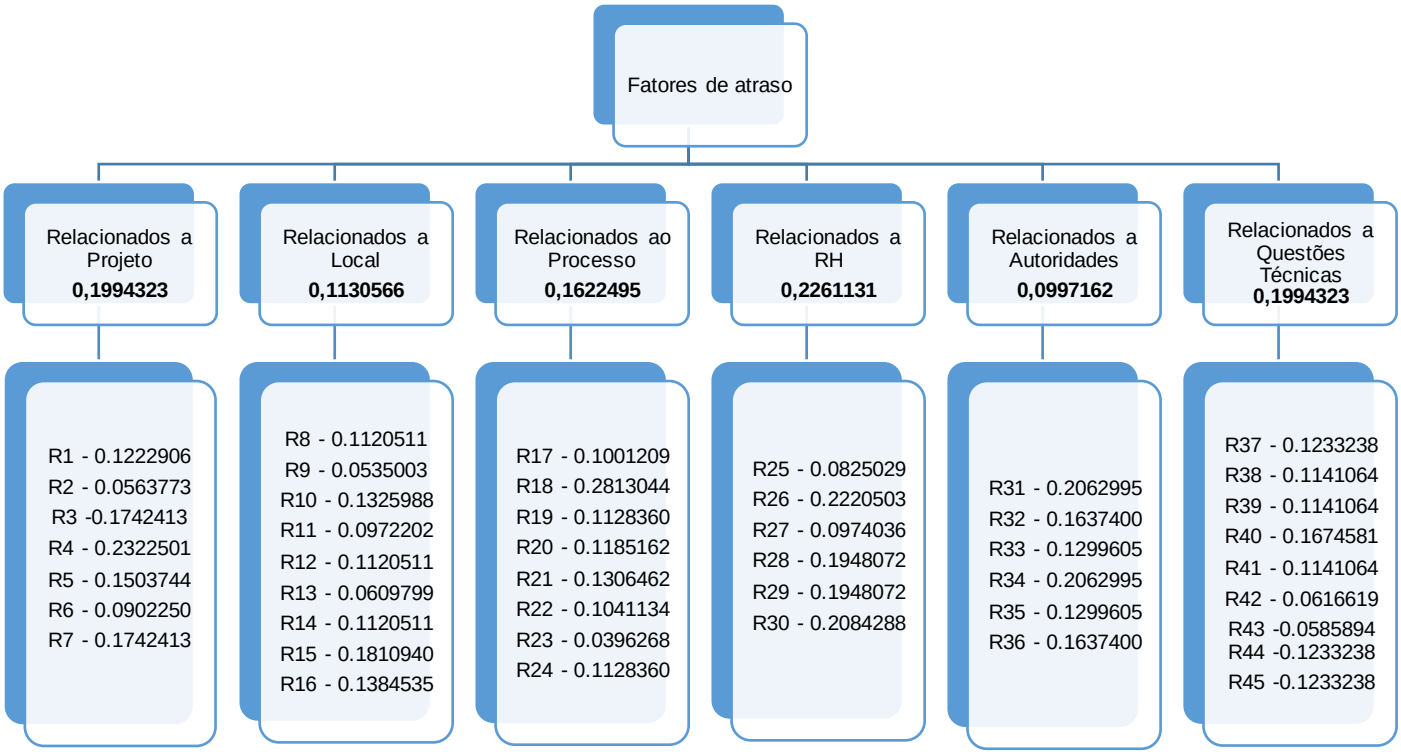


Figura 2: Resultado Global – Critérios e subcritérios com a importância
Fonte: Elaborada pelos autores

A ordem de importância das categorias de atraso, conforme a Figura 2 é: Recursos Humanos > Projeto e Questões Técnicas> Processo > Local > Autoridades. O resultado demonstra que na percepção dos especialistas e gestores, de acordo com o método AHP, o critério do Recursos Humanos é o mais importante no problema de atrasos em obras de construção.

Nos Fatores relacionados a Recursos Humanos, a classificação obtida foi R26> > R28 e R29> R30> R27> R25. Este resultado evidencia que o fator de atraso R26. Má gestão/supervisão do local foi o que possui a maior importância. Este resultado está de acordo com (SIMÃO ET AL, 2016).

Um ponto que pode ser destacado é a baixa qualificação dos profissionais do setor de construção. Contudo, a adequada qualificação dos gestores e profissionais pode proporcionar melhoria e reduzir a má gestão/supervisão do local, assim como falta de mão-de-obra qualificada para equipamentos específicos.

Em relação aos Fatores relacionados a Projeto, a classificação obtida foi R4 > R3 e R7 > R5 > R1 > R6 > R2. Este resultado demonstra que R4. Retrabalho devido à mudança de desenho ou desacordo de ordem é o fator indicado como o mais

importante. As falhas nas especificações técnicas ou mesmo modificações no escopo, podem ocasionar em aditamento do prazo e consequentemente em aumento nos custos gerais da obra.

Como medida redutora do problema, pode-se destacar o planejamento com o acompanhamento do projeto tanto do executor como pelo contratante.

Já, nos Fatores relacionados a Questões Técnicas, a classificação obtida foi R40 >R44, R45 e R37> R38, R39 e R41> R42 > R43. Este resultado aponta o fator R40. Má produtividade do trabalho como o mais importante. Como medidas de redução do problema, pode-se destacar a utilização de tecnologia nos processos de construção e melhoria das técnicas utilizadas para aumento da produtividade, como também a melhor qualificação dos profissionais, o que pode proporcionar a redução inclusive de desperdícios.

Nos Fatores relacionados ao Processo, a classificação alcançada foi: R18> R21 > R19, R20 e R24 > R22 e R17 > R23. Sendo o fator R18. Demora na aprovação do trabalho completo por parte do cliente (isto é, mudança de estágio) o que apresenta maior relevância, o que pode ser reduzido com o planejamento prévio.

Dentro dos Fatores relacionados ao Local, o subcritério R9. Condições severas de tempo/clima e R16. Especificações imprecisas da condição do local são os fatores com maior importância para os gestores. A classificação obtida neste critério é: R15> R10 e R16 > R14, R12 e R8> R11 > R13.

Quanto aos Fatores relacionados a Autoridades, a classificação obtida foi R34 e R31> R32 e R36> R33 e R35, sendo o fator de atraso R34. Mudanças em leis ou em regulamentação governamental o mais importante, o que demonstra a preocupação com os aspectos legais que regem o processo de construção de obras. A forma que pode ser reduzido o problema é a consulta a especialista jurídico no que se refere as leis, ainda no desenvolvimento do projeto e o acompanhamento durante a execução da obra.

Conclusão:

A utilização do método AHP, como uma ferramenta de auxílio na hierarquização dos fatores de atraso em obras públicas, pode tornar este processo mais eficiente, racional e claro, pois são adotadas decisões com base matemática, reduzindo os erros agregados com melhoria no processo de tomada de decisão nas medidas mitigatórias nos atrasos em obras públicas.

O resultado demonstra que na percepção dos especialistas e gestores o critério Recursos Humanos é o mais importante no problema de atrasos em obras de construção. Seguido por Projeto e Questões Técnicas, Processo, Local e Autoridades.

As causas mais significativas dos atrasos identificadas são relacionadas aos fatores *R26 Má gestão/supervisão do local; R 4. Retrabalho devido à mudança de desempenho ou desacordo de ordem, R40. Má produtividade do trabalho*, fator *R18. Demora na aprovação do trabalho completo por parte do cliente (isto é, mudança de estágio), R5. Condições severas de tempo/clima* e *R16. Especificações imprecisas da condição do local e R34 Mudanças em leis ou em regulamentação governamental.*

Verifica-se a necessidade de investimento em qualificação em gestores e profissionais da construção, em melhores sistemas, metodologias e técnicas para planejamento e controle de obras. Evidenciando que é preciso entender e planejar Recursos Humanos, Questões Técnicas, Projeto, Autoridades, Processo e Local, e como estes fatores são relacionados para a produtividade e cumprimento do prazo das obras públicas.

Como sugestão para futuros trabalhos, sugere-se a aplicação completa do método AHP para selecionar as alternativas que possibilitem mitigar os fatores de atraso, assim como trabalhos com a utilização de outros métodos a fim de realizar comparações entre os resultados encontrados e selecionar as melhores medidas para minimizar o problema de atrasos em obras públicas.

Referências:

AL-KHARASHI, A.; SKITMORE, M. Causes of delays in Saudi Arabian public sector construction projects. **Construction Management and Economics**, v. 27, n. 1, p. 3–23, 2009.

DOLOI, Hemanta; SAWHNEY, Anil; IYER, K. C.; RENTALA, Sameer. **Analysing factors affecting delays in Indian construction projects**. International Journal of Project Management Volume 30, Issue 4, May 2012, Pages 479–489.

GOMES, L. F. A. M., ARAYA, M. C. G., CARIGNANO, C. **Tomada de decisões em cenários complexos.** São Paulo: Pioneira Thompson Learning, 2004.

R CORE TEAM. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2016.

SAATY, T. L. **Método de análise hierárquica**. São Paulo: McGraw-Hill; Makron, 1991.

SHIMIZU, T. **Decisão nas organizações.** São Paulo: Atlas, 2006.

SIMÃO, A., ALCOFORADO, L.F., LONGO, O.C. Visualização de respostas dos gestores do setor público e privado sobre os atrasos em obras públicas usando o pacote sjplot do software R. In: ALCOFORADO, L.F. e LONGO, O.C. (Orgs**.) Seminário Internacional de Estatística com R: Inovação e Atuação do profissional no mercado.** Juiz de Fora, Ed. Templo, 2016.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 55 - GOOGLE TRENDS NO R

Charles Albano Coutinho[50]

Resumo

A expansão da data mining trouxe uma atenção à ferramentas que fazem extração desse tipo de conteúdo. Entre as ferramentas gratuitas, o Google Trends é uma das mais famosas, pois é possível obter, através de palavras – chaves, as principais consultas que os usuários têm feito para solver suas dúvidas através do site Google, e verificar a tendência das consultas de acordo com cada assunto escolhido. Ainda, o Google Trends possibilita aos seus usuários que filtre os seus resultados em função de um período de tempo, localidade (país e estado), categoria (esportes, política, economia, entre outros) e fontes de busca (google news, google shopping, youtube, google imagens). A pesquisa ainda pode ser realizada através de um conjunto de termos de busca, sendo possível assim obter padrões e correlações com o intuito de proporcionar vantagens estratégicas na tomada de decisões comerciais.O presente artigo busca apresentar algumas de suas inúmeras utilidades em consonância ao uso do R através do pacote gtrendsR. Suas utilidades são apresentadas através de exemplos comerciais e voltadas à área de marketing, e por fim utilizamos a buscamos os resultados do termo corrupção e verificamos como podemos construir uma *proxy* da percepção dessa variável de forma inovadora

Palavras-Chave: Data mining, Google Trends, gtrendsR

Abstract

The expansion of data mining has brought attention to the tools that extract this type of content. Among the free tools, Google Trends is one of the most famous, because it is possible to obtain, through keywords, the main queries that users have done to solve their doubts through the Google site, and verify the trend of the consultations according to each subject chosen. In addition, Google Trends enables users to filter their results based on time, location (country and state), category (sports, politics, economy, etc.) and search sources (google news, google shopping Youtube, google images). The research can still be performed through a set of search terms, thus obtaining patterns and correlations in order to provide strategic advantages in making commercial decisions.The present article seeks to present some of its numerous utilities in consonance with the use of R through the gtrendsR package. Its utilities are presented through commercial and marketing-related examples, and we finally look at the results of the term corruption and see how we can construct a proxy for the perception of this variable in an innovative way.

Keywords: Data mining, Google Trends, gtrendsR

Introdução

O Google Trends é uma ferramenta do Google que fornece, em tempo real, os dados da evolução do número de pesquisas de uma determinada palavra - chave realizada no Google. Os dados estão disponíveis desde 2004 até o presente momento

[50] Mestrando em economia na Universidade Federal Fluminense (UFF), charlescoutinho85@gmail.com

e são apresentados na forma de um índice que varia entre 0 e 100, onde 100 é um ponto em um período de tempo em que a pesquisa teve seu valor máximo. Essa escala é formada a partir do volume de dados pesquisados naquele período de tempo em relação ao total de pesquisas daquele termo desde o início das operações do Google Trends, ou seja, os dados são normalizados para oscilar em uma escala de 0 à 100, onde 100 refere-se ao(s) ponto(s) máximo(s) de procura daquele termo. As utilidades do Google Trends são inúmeras, sendo bastante utilizado em pesquisas de mercado, na economia, pesquisas eleitorais e de opinião, marketing, pois através de seus resultados podemos ter uma *proxy* da tendência de buscas daquele termo, e com isso inserir em análises de demanda de produtos, por exemplo, poupando custos que poderiam vir a ser dispendidos em alguns de tipos de pesquisa.

Objetivo

O presente artigo busca apresentar os diversos recursos possíveis de serem explorados por essa ferramenta, com o auxílio do pacote 'gtrendsR' desenvolvido por Philippe Massicotte e Dirk Eddelbuettel (2016). A colaboração proposta por esse trabalho é apresentar três formas distintas de análise utilizando o Google Trends no R, com a finalidade de exemplificar sua ampla utilização.

Metodologia

Ferramentas de análise de mercado e análise exploratória das séries dos termos pesquisados através do pacote 'gtrendsR'.

```
> #Instalar do devtools as versoes mais atualizadas do GitHub
> ############################################################
> if (!require("devtools")) install.packages("devtools")
> devtools::install_github("PMassicotte/gtrendsR")
```

Posteriormente será utilizada a função 'gtrends' para a pesquisa dos termos em razão dos seguintes fatores: localização geográfica, categoria e período temporal da análise, frequência para apresentação dos dados (diária, mensal, anual, trimestral) principalmente. A visualização gráfica dos dados do(s) termo(s) pesquisado(s) é feita através da função 'plot' ou através de outra função auxiliar, como exposto abaixo:

```
>   + ggplot2::theme(legend.position='top',
+                    legend.text = element_text(face = "bold",size=20))
```

Para o **primeiro exemplo** de utilização temos a análise dos termos de 'compra TV LED' e 'compra smart TV'.

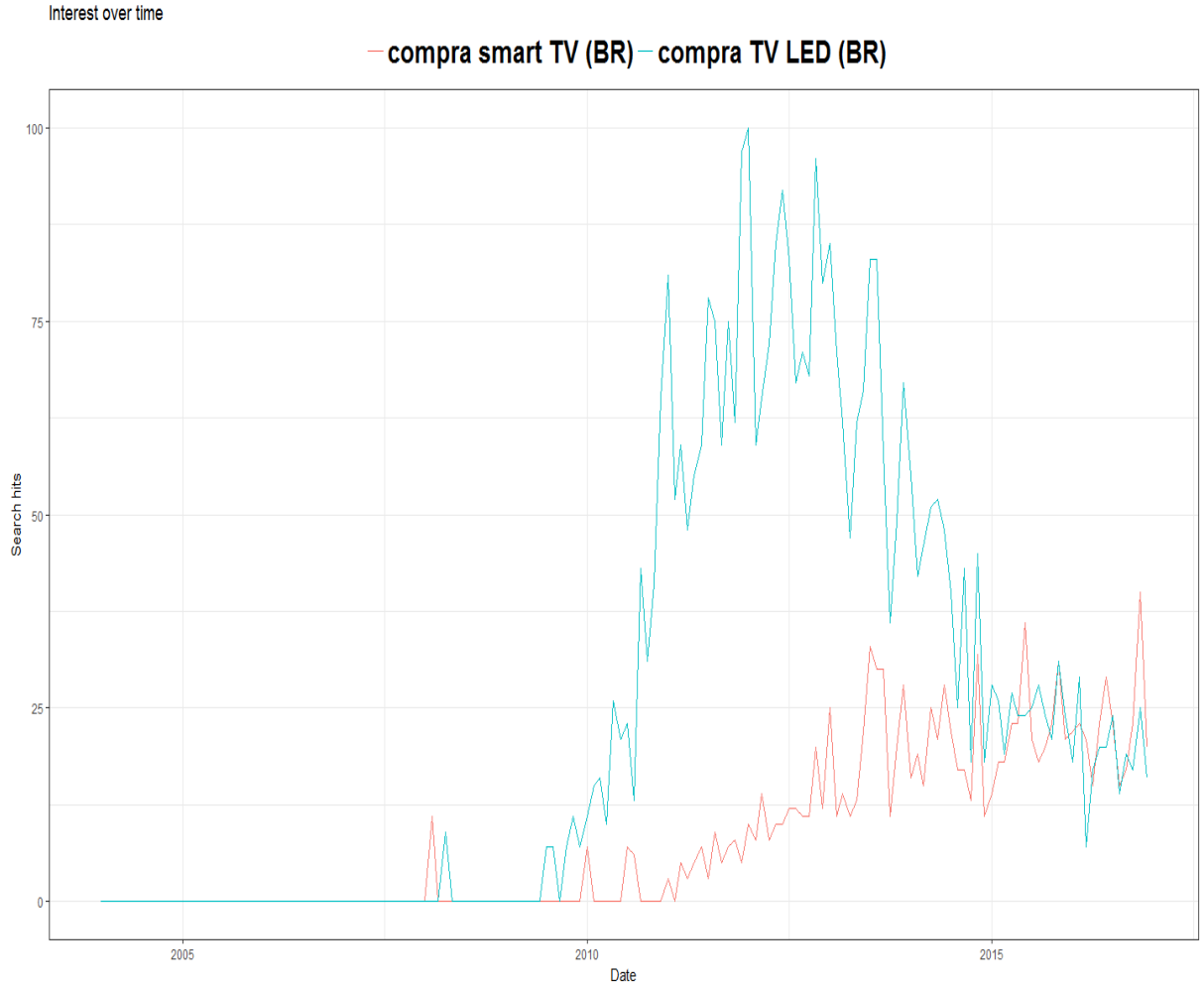


Gráfico 1: Pesquisas do termo 'compra smart TV' e 'compra TV LED' no Google. Período: 01/2004 – 12/2016.

Podemos verificar no gráfico que ambas as tecnologias eram pouco difundidas e exploradas em pesquisas pelos consumidores brasileiros na internet, porém a partir de 2009 ocorreu um marketing voltado à aquisição de novas televisões em razão da Copa do Mundo, e como isso a demanda por TV LED foi alavancada. Após a Copa, a demanda por tal produto foi gradativamente reduzindo-se a partir de 2012 em decorrência de uma expansão das TV Smart (que receberam diversos novos modelos e integrações à dispositivos a partir desse ano). Já a análise exploratória básica dos dados pode ser obtida da seguinte forma:

```
> ddply(df1,'keyword', summarise,N=length(hits),
+                min=min(hits),
+                primeiro_quartil=quantile(hits,0.25),
+                media=mean(hits),
+                mediana=median(hits),
+                terceiro_quartil=quantile(hits,0.75),
+                maximo=max(hits),
+                sd=sd(hits),
+                se=sd/sqrt(N))
```

keyword	N	min	primeiro_quartil	media	mediana	terceiro_quartil	maximo	sd	se
compra smart TV	156	0	0	8,012820513	0	15	40	10,22330493	0,818519472
compra TV LED	156	0	0	24,87179487	15,5	47,25	100	29,11262179	2,33087519

Após a criação do objeto trend também é possível analisar os principais temas que estão relacionados aos termos de pesquisa.

```
> head(trend$related_queries) #temas relacionados
       subject related_queries value geo
1 compra facil             top   100  BR
2    tv led 32             top    45  BR
3    tv led 42             top    30  BR
4    tv led 40             top    30  BR
5   americanas             top    25  BR
6  casas bahia             top    20  BR
```

O **exemplo 2** é baseado no mercado de cervejas e vinhos.

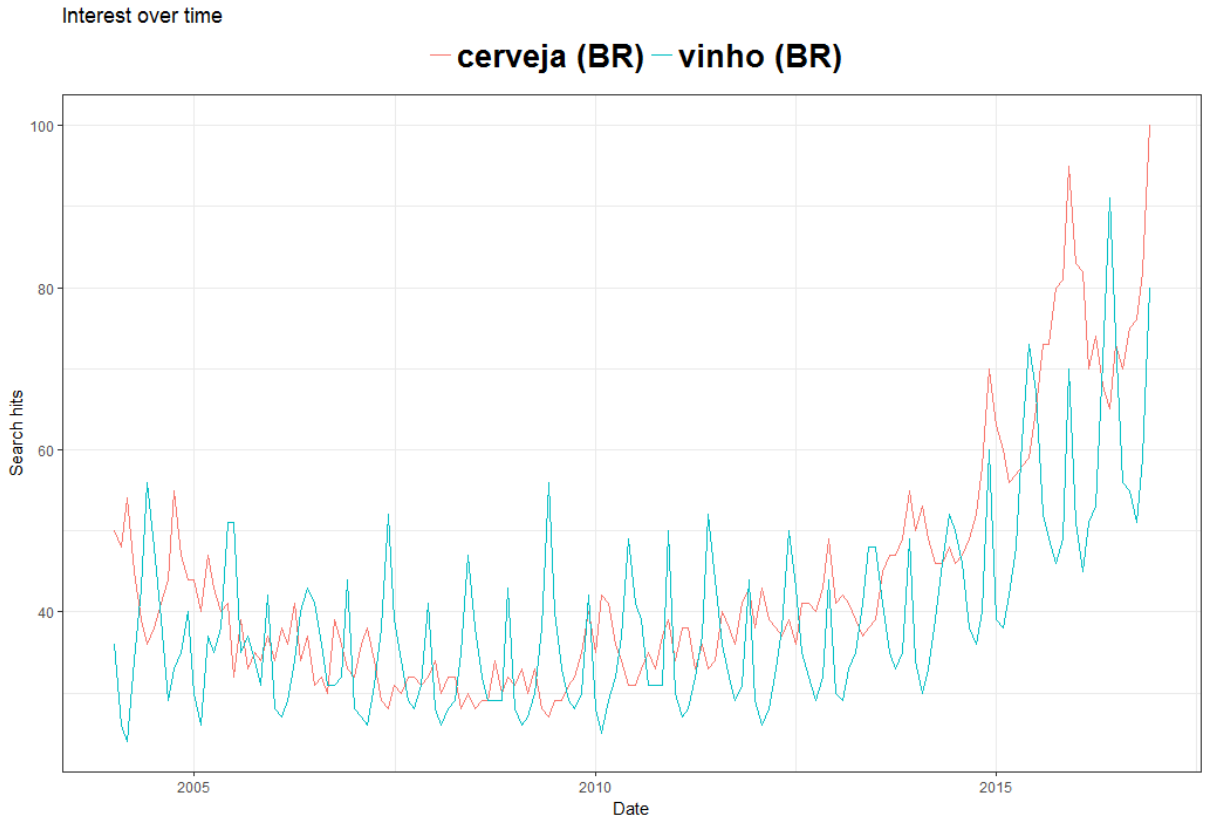


Gráfico 2: Pesquisas do termo 'cerveja' e 'vinho' no Google. Período: 01/2004 – 12/2016.

Notem que o comportamento das pesquisas sobre vinhos possuem uma sazonalidade associada ao final do ano (os picos são apresentados nos meses de dezembro), ao passo que a pesquisa sobre o termo cerveja é tem comportamento

mais suavizado em relação ao vinho. Notem que a partir de dezembro de 2014 a pesquisa por cerveja no período de festas de fim de ano superou os de vinhos, e o comportamento replicou-se para os anos subsequentes. Partindo desse pressuposto temos indícios, por inspeção gráfica, que estamos em uma transição dos costumes de bebidas em festas de fim de ano, substituindo gradualmente o tradicional vinho por cerveja.

keyword	N	min	primeiro_quartil	media	mediana	terceiro_quartil	maximo	sd	se
cerveja	156	27	33	43,49358974	39	47,25	100	14,77683313	1,183093504
vinho	156	24	30	39,14102564	36	45,25	91	11,7980035	0,94459626

Por fim, é apresentada uma proxy da variável corrupção. A literatura, inclusive econômica e de ciências sociais, apresenta recorrentemente a necessidade de introduzir tal variável em suas análises. Será apresentado seu resultado e verificaremos a relação desse termo com algumas variáveis econômicas e uma breve aplicação.

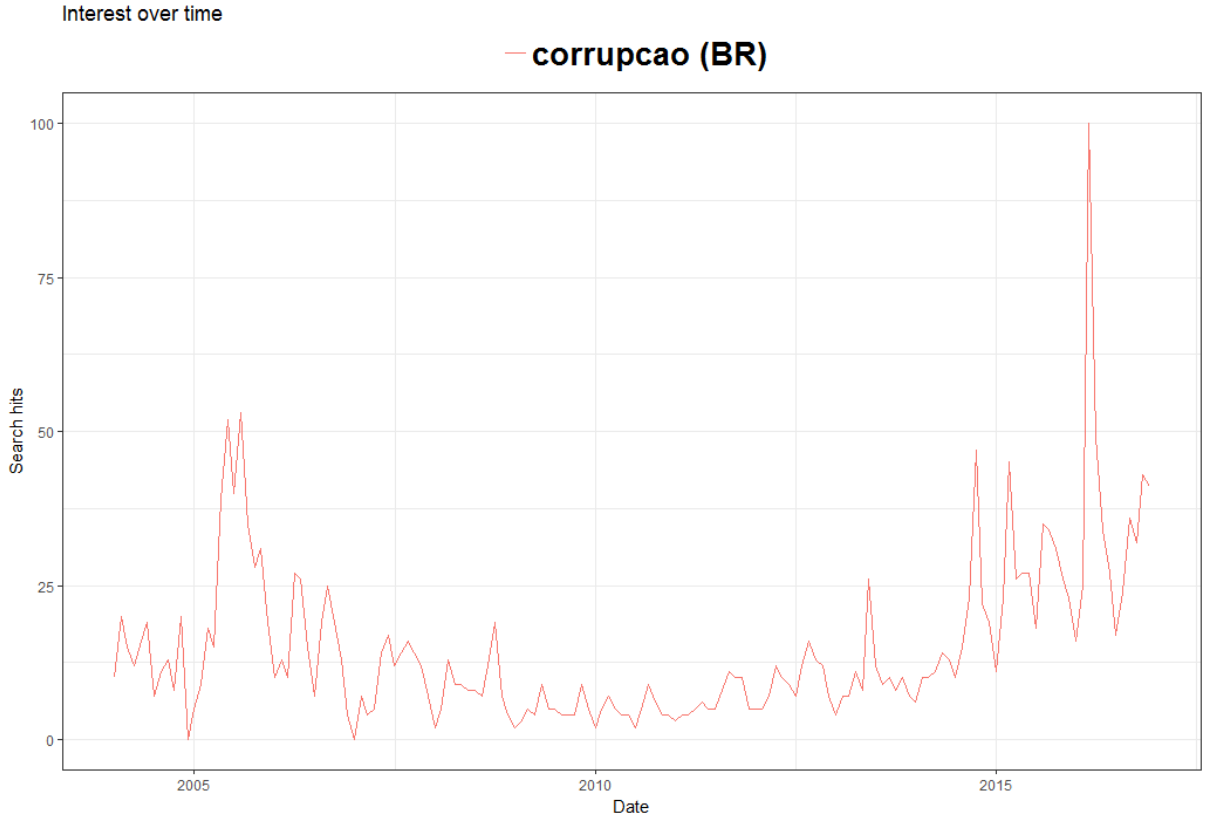


Gráfico 3: Pesquisas do termo 'corrupcao' no Google. Período: 01/2004 – 12/2016.

keyword	N	min	primeiro_quartil	Media	mediana	terceiro_quartil	maximo	sd	se
corrupcao	156	0	6	14,76923077	10	19	100	13,26691065	1,062202954

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Resultados e Discussão

Como exemplo de aplicação, segue abaixo uma regressão linear múltipla utilizando os dados obtidos no Google Trends para a percepção de corrupção (período 2004.01 à 2016.11). A variável dependente é a Dívida Pública Bruta.

```
> modreg=lm(data=dadosr,Div.Publ.Bruta~gtcorrup+IPCA_acum+
+             tx_cambio_ef_real+IBC_BR)
> summary(modreg)

Call:
lm(formula = Div.Publ.Bruta ~ gtcorrup + IPCA_acum + tx_cambio_ef_real +
    IBC_BR, data = dadosr)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4596 -2.0431 -0.1684  1.3465  8.7874

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       72.00671    4.40225  16.357  < 2e-16 ***
gtcorrup           0.06978    0.02089   3.341 0.001055 **
IPCA_acum          0.74296    0.20340   3.653 0.000358 ***
tx_cambio_ef_real  0.05865    0.02206   2.658 0.008711 **
IBC_BR            -0.15085    0.02524  -5.976  1.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.835 on 150 degrees of freedom
Multiple R-squared:  0.6119,   Adjusted R-squared:  0.6015
F-statistic: 59.11 on 4 and 150 DF,  p-value: < 2.2e-16
```

Tabela 1: Variáveis utilizadas na regressão múltipla 'modreg'

Variável	Descrição	Fonte
Div.Publ.Bruta	Dívida bruta do governo geral (% PIB) - Metodologia utilizada até 2007 - %	SGS BACEN - código 4537
gtcorrup	variável de percepção de busca sobre o termo 'corrupcao' no Brasil	Google Trends
IPCA_acum	Índice nacional de preços ao consumidor - amplo (IPCA) - em 12 meses - %	SGS BACEN - código 13522
tx_cambio_ef_real	Índice da taxa de câmbio efetiva real (IPCA) - Jun/1994=100 - Índice	SGS BACEN - código 11752
IBC_BR	Indice de Atividade Econômica	SGS BACEN - código 24363

Fonte: autoria própria.

O sinal das variáveis estão de acordo com o esperado pela literatura, em especial vale ressaltar o resultado do coeficiente estimado para a percepção de corrupção (gtcorrup).

Conclusão

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

O pacote 'gtrends' permite a análise, via percepção de tendência, de termos que podem ser utilizados para os mais distintos campos de conhecimento, porém são de difícil cálculo ou sem uma metodologia de obtenção consagrada, como é o caso da proxy da variável 'corrupcao'. Sua restrição está no campo de previsões, pois como já alertado por Choi e Varian (2012) *"We are not claiming that Google Trends data can help in predicting the future. Rather we are claiming that Google Trends may help in predicting the present"*.

Referências:

BANCO CENTRAL DO BRASIL. **SGS – Sistema Gerenciador de Séries Temporais – v.2.1.** Disponível em < https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=preparar TelaLocalizarSeries >. Acesso em 15 de março de 2017.

BANDUCCI, S. A., KARP, J. A. **The electoral consequences of scandal and reapportionment in the 1992 House elections**. American Politics Quarterly, 22 (1): 3–26. 2004.

BOWLER, S., KARP, J. A. **Politicians, scandals and trust in Government.** Political Behavior, 26 (3): 271-287.2004.

CHOI, Hyuyoung, VARIAN, Hal. **Predicting the Present with Google Trends.** The Economic Record, Vol. 88, Special Issue, Junho/2012.

FANTAZZINI, Dean, TOKTAMYSOVA, Zhamal. **Forecasting German car sales using Google data and multivariate models**. Internation J. Production Economics. 2015.

HAMID, Alain. HEIDEN, Moritz. **Forecasting volatility with empirical similarity and Google Trends**. Journal of Economic Behavior & Organization. Elsevier.2015.

IPEADATA. Disponível em < ipeadata.gov.br >. Acesso em 15 de março de 2017.

LI, Xin, et.all. **How does Google search affect trader positions and crude oil prices?.** Economic Modelling. Elsevier.2015.

MASSICOTTE, Philippe. EDDELBUETTEL, Dirk. gtrendsR: Perform and Display Google Trends Queries. R package version 1.3.5. Disponível em < https://cran.r-project.org/web/packages/gtrendsR/ >. Acesso em 15 de março de 2017.

TECMUNDO. **LED TV, a sensação de 2010**. Disponível em < https://www.tecmundo.com.br/led/2262-led-tv-a-sensacao-de-2010.htm >. Acesso em 15 de março de 2017.

VLASTAKIS, Nikolaos, MARKELLOS, Raphael N. **Information demand and stock market volatility.** Elsevier. 2012.

VOSEN, Simeon, SCHMIDT, Torsten. **A monthly consumption indicator for Germany based on Internet search query data**. Applied Economics Letters. 19:7, pág.: 683 – 687. 2012.

ZEYBEK, Ömer, UGURLU, Erginbay. **Nowcasting credit demand in Turkey with Google Trends**. Journal of Applied Economic Sciences. Volume X. Issue 2(32). 2015.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 61 – AVALIAÇÃO DAS CARACTERÍSTICAS FÍSICO-QUÍMICA DO VINHO COM RELAÇÃO À SUA QUALIDADE UTILIZANDO ANÁLISE DE COMPONENTES PRINCIPAIS

Iasmin da Silva Ferreira[51]
Karinne Novaes de Moraes[52]
Vinicius Sampaio Andrade[53]
Marcus Vinicius Pereira de Souza[54]

Resumo

O objetivo deste artigo é investigar a qualidade do vinho com relação às suas características fisico-química. Para tal, utilizou-se 178 amostras de vinho oriundos de uma mesma região da Itália, mas produzidos a partir de três solos de vinhedos distintos. Estes dados estão disponibilizados no pacote *rattle* do *software* estatístico R. A técnica estatística multivariada denominada Análise de Componentes Principais (PCA) foi aplicada com o propósito de identificar quais atributos eram responsáveis pela qualidade de cada tipo de vinho. É digno registrar que esta metodologia visa explicar a estrutura de variância e covariância de um vetor aleatório, composto de p variáveis aleatórias, através da construção de combinações lineares da variância original. Estas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si.

Palavras-Chave: análise de componentes principais, vinho italiano, autovetores, autovalores.

Abstract

The objective of this article is to investigate a quality of wine in relation to its physicochemical characteristics. For a total of 178 samples of wine from a region of the same world, the product has three levels of different vineyards. This data is available in the rattle package of statistical software R. The multivariate statistical technique called Principal Component Analysis (PCA) was applied with the purpose of identifying which attributes were controlled by the quality of each type of wine. It is worth noting that this methodology aims to explain the variance and covariance structure of a random vector, composed of p random variables, through the construction of combinations of original variance. These linear combinations are called principal components and are not correlated with each other.

Keywords: principal componente analysis, Italian wine, eigenvectors, eigenvalues.

[51] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ) Campus Valença - ferreira.ias@gmail.com
[52] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ) Campus Valença - novaes.karinne@gmail.com
[53] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ) Campus Valença - viniciusampaios@gmail.com
[54] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ) Campus Valença - marcus.souza@cefet-rj.br

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

O uso de rolhas de cortiça natural, como vedante de garrafas, é secular. Referindo-se especificamente ao vinho, a literatura especializada descreve que ao se colocar esta bebida em contato direto com rolhas de cortiça, é possível ocorrer significativas alterações sensoriais. De fato, se ocorrer a migração de 2,4,6-tricloroanisol (TCA) da rolha de cortiça para o vinho, este apresentará um sabor de mofo. Vale explicar que este TCA é resultante da atividade de microrganismos (fungos). Continuando, Trace e Skaalen (2008) mencionam que existem outros compostos que também podem causar alterações organolépticas no produto como, por exemplo: TBA, TeCA e PCA.

Isto posto, diversas pesquisas relacionadas com a análise sensorial têm sido desenvolvidas com o objetivo de minimizar o risco de desvios sensoriais em vinhos engarrafados. De acordo com Teixeira (2009), a análise sensorial é definida pela Associação Brasileira de Normas Técnicas (ABNT, 1993) como a disciplina científica utilizada para evocar, medir, analisar e interpretar reações das características dos alimentos e materiais; como são percebidos pelos sentidos da visão, olfação, gustação, tato e audição. Além disso, Teixeira (2009) descreve que no setor alimentar, a análise sensorial é de grande importância por avaliar a aceitabilidade mercadológica e a qualidade do produto, sendo parte inerente ao plano de controle de qualidade de uma indústria. Continuando, é importante registrar que os métodos sensoriais podem ser divididos em 3 tipos (Meilgaard, Civille e Carr, 1999): 1) métodos discriminativos; 2) métodos analíticos ou descritivos; e, 3) métodos afetivos.

Uma outra maneira que pode ser utilizada para investigar a qualidade do vinho baseia-se em análises físico-química. Segundo Diniz et al. (sem data), tais análises envolvem a investigação do teor de álcool, acidez, pH, açúcar, dentre outros.

Objetivo

O objetivo deste trabalho é examinar a qualidade do vinho a partir da técnica denominada análise de componentes principais (PCA). De acordo com Mingoti (2005), esta metodologia visa explicar a estrutura de variância e covariância de um vetor aleatório, composto de p variáveis aleatórias, através da construção de combinações

lineares da variância original. Estas combinações lineares são chamadas de componentes principais e são não correlacionadas entre si.

Material e Métodos:

Este estudo usa os dados físico-químicos [1]de 178 amostras de vinho oriundos de uma mesma região da Itália, mas produzidos a partir de três solos de vinhedos distintos. Diante do exposto, é válido informar que cada amostra está classificada em tipos de vinho: a) vinho tipo 1, b) vinho tipo 2, e c) vinho tipo 3). Além disso, as mesmas estão caracterizadas por 13 atributos, a saber: 1) Alcohol, 2) Malic, 3) Ash, 4) Alcalinity, 5) Magnesium, 6) Phenols, 7) Flavanoids, 8) Nonflavanoids, 9) Proanthocyanins, 10) Color, 11) Hue, 12) Dilution, e 13) Proline.

A metodologia utilizada foi, inicialmente, separar a amostra em 3 agrupamentos (de acordo com o tipo de vinho) e, a seguir, aplicar PCA em cada um dos agrupamentos.

Resultados e Discussão:

Neste tópico, vamos apresentar os resultados obtidos a partir da análise de componentes principais (PCA). Antes, porém, foi realizado uma análise descritiva dos atributos e os resultados (médias e os desvios-padrão) estão registrados nas Tabelas 1 e 2 abaixo.

Tabela 1: Média dos atributos de acordo com os vinhos tipo 1, tipo 2 e tipo 3.

	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proant.	Color	Hue	Dilution	Proline
V1	13,74	2,01	2,46	17,04	106,30	2,84	2,98	0,29	1,90	5,53	1,06	3,16	1115,70
V2	12,28	1,93	2,25	20,24	94,55	2,26	2,08	0,36	1,63	3,09	1,06	2,79	519,50
V3	13,15	3,33	2,44	21,22	99,31	1,68	0,78	0,45	1,16	7,40	0,68	1,68	629,90

Tabela 2: Desvios-padrão dos atributos de acordo com os vinhos tipo 1, tipo 2 e tipo 3.

	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proant.	Color	Hue	Dilution	Proline
V1	0,46	0,69	0,23	2,55	10,50	0,34	0,40	0,07	0,41	1,24	0,12	0,36	221,52
V2	0,54	1,02	0,32	3,35	16,75	0,55	0,71	0,12	0,60	0,92	0,20	0,50	157,21
V3	0,53	1,09	0,18	2,26	10,89	0,36	0,29	0,12	0,41	2,31	0,11	0,27	115,10

Por inspeção visual dos resultados acima, é possível verificar que as médias de V1, V2 e V3 estão muito próximas, com exceção do atributo Proline. Já com relação aos desvios-padrão, salta aos olhos a variabilidade dos seguintes atributos:

[1] Os dados estão disponibilizados no pacote *rattle* (*https://cran.r-project.org/web/packages/rattle/index.html*)

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Magnesium (V2), Color (V3) e Proline (V2 e V3). Como a variância dos atributos é muito diferente, é interessante usar uma PCA de correlação para que a variável com maior variância não “domine” a análise.

De posse dessas informações, o próximo passo consiste em estimar o número de componentes. Para tal, em todas as análises, adotamos modelos que expliquem pelo menos 70% da variância total. Com relação ao vinho tipo 1, a primeira componente principal responde por cerca de 28% da variância total dos dados padronizados. Assim sendo, se considerarmos as cinco primeiras componentes atingimos cerca de 75% da variância total. Além disso, o Gráfico 1 indica que o número de componentes a reter é 5.

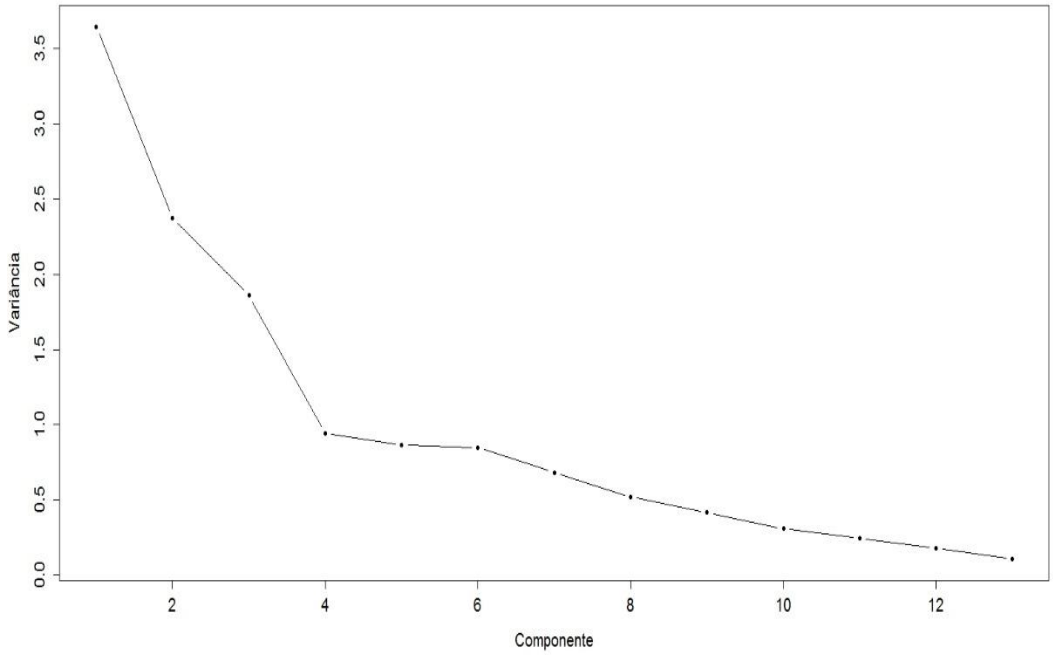


Gráfico 1: *Screeplot* dos autovalores das componentes principais (vinho tipo 1)

Analisando os dados relacionados com o vinho tipo 2, verifica-se que a primeira componente principal responde por cerca de 24% da variância total dos dados padronizados. Assim sendo, se considerarmos as cinco primeiras componentes atingimos cerca de 73% da variância total. Além disso, o Gráfico 2 indica que o número de componentes a reter é 5.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

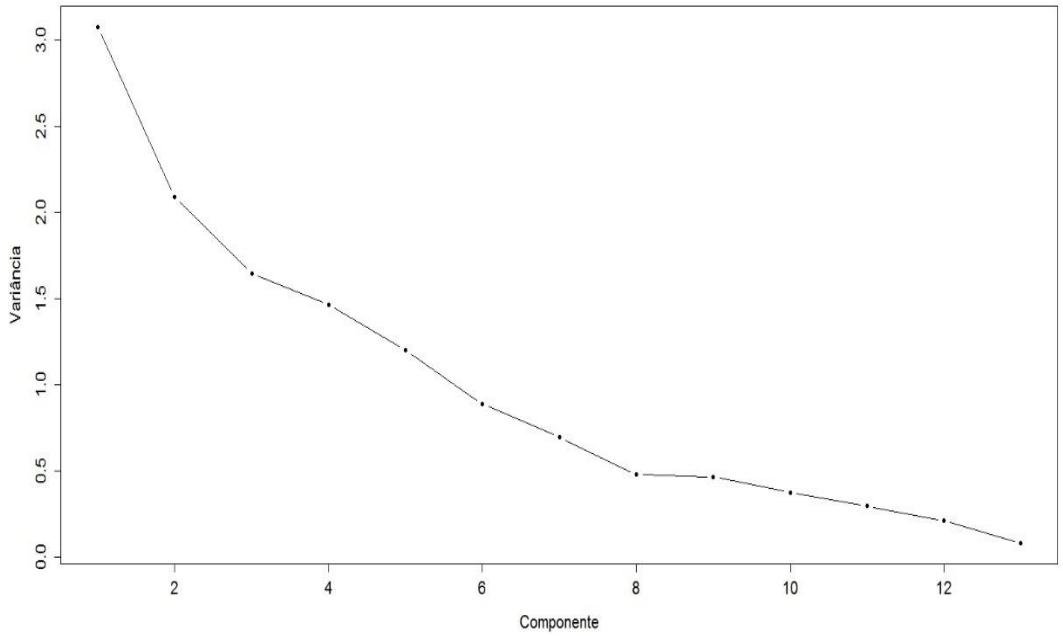


Gráfico 2: *Screeplot* dos autovalores das componentes principais (vinho tipo 2)

Com relação ao vinho tipo 3, é possível verificar que a primeira componente principal responde por cerca de 25% da variância total dos dados padronizados. Assim sendo, se considerarmos as cinco primeiras componentes atingimos cerca de 77% da variância total. Além disso, o Gráfico 3 indica que o número de componentes a reter é 5.

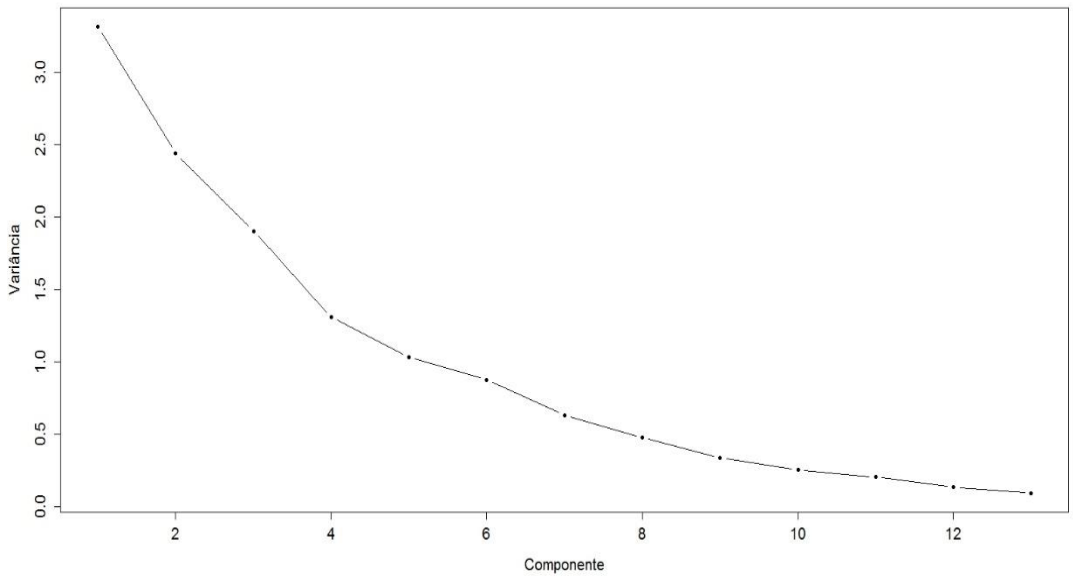


Gráfico 3: *Screeplot* dos autovalores das componentes principais (vinho tipo 3).

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

De acordo com a literatura especializada, a primeira componente é um índice global relacionada com a qualidade.

Em linha com esse raciocínio, podemos conjecturar que as características físico-química responsáveis pela qualidade do vinho tipo 1 são (em grau de importância): i) Flavanoids.; ii) Color; iii) Phenols, iv) Alcohol, v) Proline, vi) Proanthocyanins, vii) Magnesium e viii) Hue.

Com relação ao vinho tipo 2, temos: 1) Nonflavanoids, 2) Alcool, 3) Hue, e 4) Proline.

Para o vinho tipo 3, temos: a) Hue, b) Malic, c) Nonflavanoids, d) Dilution, e e) Proline.

Conclusão:

Este estudo mostrou que a técnica PCA pode ser muito útil na análise multivariada.

Agradecimentos:

Os autores agradecem ao PIBIC CEFET/RJ pelo apoio financeiro dado a este projeto.

Referências:

TRACY, R.; SKAALEN, B. Next '2,4,6 – TCA' in U.S. wine industry. Practical Winery & Vineyard Journal, p. 1-2, 2008. Disponível em: < https://www.practicalwinery.com/novdec08/page1.htm >. Acesso em: 01 de mar. 2017.

MEILGAARD, M.; CIVILLE, G.V.; CARR, B.T. Sensory evaluation techniques. 3rd. New York: CRC, 1999.

MINGOTI, S. A. Análise de dados através de métodos de estatística multivariada – uma abordagem aplicada. 1ed. Belo Horizonte: Editora UFMG, 2005.

TEIXEIRA, L. V. Análise sensorial na indústria de alimentos. Revista Instituto de Laticínios Cândido Tostes, v. 64 ,n.366, p. 12-21, 2009.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Apêndice A - Código fonte do programa em R.

```
#-------------------------------------------------------------------------------------------------
---
# Autores:
# Marcus Vinicius Pereira de Souza
# marcus.souza@cefet-rj.br
# Iasmin da Silva Ferreira
# ferreira.ias@gmail.com
# Karinne Novaes de Moraes
# novaes.karinne@gmail.com
# Vinicius Sampaio Andrade
# viniciusampaios@gmail.com
# Data: 23/04/2017
#-------------------------------------------------------------------------------------------------
# O objetivo deste programa é analisar a qualidade de vinho utilizando Análise de
Componentes Principais (PCA)
# A base de dados contém 178 amostras de vinho italiano.
# Os vinhos estão classificados em 3 tipos.
#-------------------------------------------------------------------------------------------------
# carregando os pacotes
# 1) rattle: possui a base da dados 'wine'.
# 2) vegan: the functions in the vegan package contain tools for diversity analysis,
ordination methods and
#    tools for the analysis of dissimilarities. Together with the labdsv package, the
vegan package
#    provides most standard tools of descriptive community analysis. function
prcomp.
# crtl + L --> limpa o console
# crtl + S --> salva o arquivo
# crtl + A --> seleciona tudo
# crtl + R --> executa o script

library(rattle)
```

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

```
library(vegan)

# carregando os dados

data(wine)  # base de dados
head(wine)  # nome dos atributos

#-----------------------------------------------------------------------------------------------------
amostra_vinho1 <- wine[c(1:59),c(2:14)]  # amostras do vinho tipo 1
stat_amostra_vinho1<- summary(amostra_vinho1)  # análise descritiva dos dados referentes ao tipo de vinho 1

amostra_vinho2 <- wine[c(60:130),c(2:14)]  # amostras do vinho tipo 2
stat_amostra_vinho2<- summary(amostra_vinho2)  # análise descritiva dos dados referentes ao tipo de vinho 2

amostra_vinho3 <- wine[c(131:178),c(2:14)]  # amostras do vinho tipo 3
stat_amostra_vinho3<- summary(amostra_vinho3)  # análise descritiva dos dados referentes ao tipo de vinho 3

#-----------------------------------------------------------------------------------------------------
resul1.pca  <- prcomp(amostra_vinho1,scale=T)

stat_resul1 <- summary(resul1.pca)
# sum(stat_resul1$sdev^2)  --> número de atributos
cinco_primeiros_componentes_resul1 <-- resul1.pca$rotation[, 1:5]
ordem1 <- order(resul1.pca$rotation[, 1], decreasing = FALSE)
print(resul1.pca$sdev[1:5] * t(resul1.pca$rotation[, 1:5]), digits = 3)

plot(1:ncol(amostra_vinho1), resul1.pca$sdev^2, type = "b", xlab = "Componente",
     ylab = "Variância", pch = 20, cex.axis = 1.3, cex.lab = 1.3)

biplot(resul1.pca)
```

```
screeplot(resul1.pca)

#-----------------------------------------------------------------------------------------------------
resul2.pca  <- prcomp(amostra_vinho2,scale=T)

stat_resul2 <- summary(resul2.pca)
# sum(stat_resul1$sdev^2)   --> número de atributos
cinco_primeiros_componentes_resul2  <-- resul2.pca$rotation[, 1:5]
ordem2 <- order(resul2.pca$rotation[, 1], decreasing = FALSE)
print(resul2.pca$sdev[1:5] * t(resul2.pca$rotation[, 1:5]), digits = 3)

plot(1:ncol(amostra_vinho2),  resul2.pca$sdev^2,  type = "b", xlab = "Componente",
     ylab = "Variância", pch = 20, cex.axis = 1.3, cex.lab = 1.3)

biplot(resul2.pca)
screeplot(resul2.pca)

#-----------------------------------------------------------------------------------------------------
resul3.pca  <- prcomp(amostra_vinho3,scale=T)

stat_resul3 <- summary(resul3.pca)
# sum(stat_resul1$sdev^2)   --> número de atributos
cinco_primeiros_componentes_resul3  <-- resul3.pca$rotation[, 1:5]
ordem3 <- order(resul3.pca$rotation[, 1], decreasing = FALSE)
print(resul3.pca$sdev[1:5] * t(resul3.pca$rotation[, 1:5]), digits = 3)

plot(1:ncol(amostra_vinho3),  resul3.pca$sdev^2,  type = "b", xlab = "Componente",
     ylab = "Variância", pch = 20, cex.axis = 1.3, cex.lab = 1.3)

biplot(resul3.pca)
screeplot(resul3.pca)
```

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 62 - UTILIZAÇÃO DO PACOTE MCMC4EXTREMES PARA ANÁLISE ESTATÍSTICA DE VALORES EXTREMOS DE DADOS AMBIENTAIS DAS CAPITAIS DA REGIÃO NORDESTE

Zeferino Gomes da Silva Neto[55]
Fernando Ferraz do Nascimento[56]

Resumo

O estudo da Teoria de Valores extremos (TVE) consiste em duas abordagens distintas. Uma delas, no qual se refere este trabalho, consiste em pegar dados diários de precipitação de chuva da Região Nordeste, e analisar valores altos, acima de um determinado limiar u, considerada distribuição dos excessos. Para esta distribuição específica, a TVE diz que para um limiar u grande, a distribuição adequada é conhecida como distribuição de Pareto generalizada (GPD). Esta distribuição possui três parâmetros, que vão definir o comportamento da distribuição de excessos de dados ambientais, e nos permite calcular probabilidades futuras e prever em um horizonte futuro com qual frequência um evento raro, que resulta em danos ou catástrofes, poderá ocorrer. Portanto, além de encontrar a estimativa dos parâmetros do modelo, também é muito importante encontrar uma forma para determinar os quantis altos, além do limiar, de tal forma que se X, possui distribuição GPD, é importante saber com qual probabilidade ocorre um evento maior ou igual a q, ou seja, P(X>q)=1-p. Com o cálculo destes quantis podemos prever por exemplo, qual o maior nível de chuva esperado que ocorra em Teresina nos próximos 20 anos, ou qual a maior vazão que o Rio Parnaíba terá nos próximos 10 anos.

Palavras-Chave: Valores Extremos, Precipitação, Região Nordeste, Predição.

Abstract

The study of Extreme Values Theory (EVA) consists of two distinct approaches. One of them, in which this work is concerned, consists of collecting daily rainfall data from the Northeast Region, and analyzing high values, above a certain threshold u, considered distribution of excesses. For this specific distribution, the EVA says that for a large u-threshold, the proper distribution is known as Generalized Pareto Distribution (GPD). This distribution has three parameters that will define the behavior of the distribution of environmental data excesses, and allows us to calculate future probabilities and predict in a future horizon how often a rare event that results in damages or catastrophes may occur. Therefore, in addition to finding the estimation of the model parameters, it is also very important to find a way to determine the high quantiles in addition to the threshold, so that if X has a GPD distribution, it is important to know with which probability a larger event occurs or equal to q, that is, P (X> q) = 1-p. With the calculation of these quantiles we can predict, for example, the highest level of rainfall expected in Teresina over the next 20 years, or what the higher flow that the Parnaíba River will have in the next 10 years.

Keywords: Extreme Values. Rainfall. Northeast Region. Prediction.

[55] Universidade Federal do Piauí-UFPI, zeferinon@gmail.com
[56] Universidade Federal do Piauí-UFPI, fernandofn@ufpi.edu.br

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

Fenômenos pouco ocasionais têm grande significância em várias áreas do conhecimento, dentre elas, a climatologia e outras diversas áreas ambientais, fenômenos assim tem muito destaque também na economia e ajudam a pressupor danos ou benefícios, e suas estimativas probabilísticas são fundamentais para o planejamento e prosseguimento de estudos sujeitos a consequências adversas. Na região Nordeste do Brasil as chuvas ocorrem poucas vezes durante o ano, e essa região recebe pouca influência de massas de ar vindas principalmente da região Sul, com isso o conhecimento da regularidade com que esses eventos ocorrem é de muito interesse para a população e para os órgãos representantes. A Teoria de Valores Extremos é essencial neste caso para a modelagem destes eventos. Diversos recursos tecnológicos são utilizados pra se trabalhar com essa Teoria, em especial o software R é um dos principais meios computacionais voltado pra estatística utilizado em todo o mundo, por meio dele foi criado o pacote MCMC4Extremes baseado na teoria de Valores Extremos, este pacote busca por meio de um uso prático, facilitar ao pesquisador que seja feita a modelagem de eventos raros.

Objetivo

O presente trabalho tem como objetivo principal utilizar o pacote MCMC4Extremes como ferramenta eficaz para a modelagem de valores extremos, além disso, criar um modelo que estime a probabilidade de ocorrência de precipitação de chuva muito elevada em determinado período nas capitais da região Nordeste e utilizar medidas de ajuste e gráficos de predição para verificar o modelo.

Material e Métodos:

Os princípios da teoria de Valores Estremos foram desenvolvidos por Fisher e Tippett (1928), que estabeleceram três tipos de distribuições assintóticas de valores extremos, conhecidas como de Gumbel (tipo I), Fréchet (tipo II) e Weibull (tipo III). Em 1955, Jenkinson apresentou a distribuição generalizada de valores extremos (GEV), declarada como uma família de distribuições, pois esta é suficiente para presumir os três tipos de distribuições assintóticas de valores extremos.

Porém, em alguns procedimentos podem acontecer alguns problemas envolvendo valores extremos, dependendo do caso, não são eficientes para modelar o comportamento dos extremos. Em determinados conjuntos de dados, pode acontecer de um conjunto ter mais valores extremos do que outros, com isso,

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

devemos traçar um limiar, considerando tudo que for maior do que esse limiar um evento extremo, assim podemos descrever a distribuição de Pareto Generalizada (GPD) desenvolvida por Pickands (1975), que consiste em representar a distribuição limite dos excessos além de um limiar consideravelmente alto. A distribuição de Pareto Generalizada (GPD) aplica-se com a distribuição limite dos excessos além de um limiar alto, Os excessos além do limiar u, denotados por $X1, \ldots, X_{(Nu)}$, são os valores $Yi - u \geq 0$.

A distribuição de Pareto Generalizada foi desenvolvida por Pickands [1975] e se constitui no seguinte teorema:

Teorema 1: *Se X é uma variável aleatória com função distribuição $F(x)$, que pertence ao domínio de atração de uma distribuição GEV, então quando $u \to \infty$; $F(x|u) = P(X \leq u + x|X > u)$ possui distribuição GPD, possui a seguinte função de distribuição*

$$G(x|\xi,\sigma,u) = \begin{cases} 1 - \left(1 + \xi\frac{(x-u)}{\sigma}\right)^{-1/\xi}, se\ \xi \neq 0 \\ 1 - \exp\left\{-\frac{(x-u)}{\sigma}\right\}, se\ \xi = 0 \end{cases}$$

Onde $x - u > 0$ para $\xi \geq 0$ e $0 \leq x - u < -\sigma/\xi$ para $\xi < 0$. O caso $\xi = 0$ é interpretado como sendo o limite quando $\xi \to 0$, é a distribuição exponencial de parâmetro $1/\sigma$. Os parâmetros são ξ, σ e u e representam a fórma, escala e limiar da distribuição. Os métodos de estimação desses parâmetros não possuem uma forma analítica, sendo necessárias aproximações numéricas. Estas aproximações podem ser feitas a partir da programação, utilizando softwares estatísticos. Entre estes softwares, o mais usado para a realização de programação é o R, pois tem uma linguagem utilizada para cálculos estatísticos e gráficos.

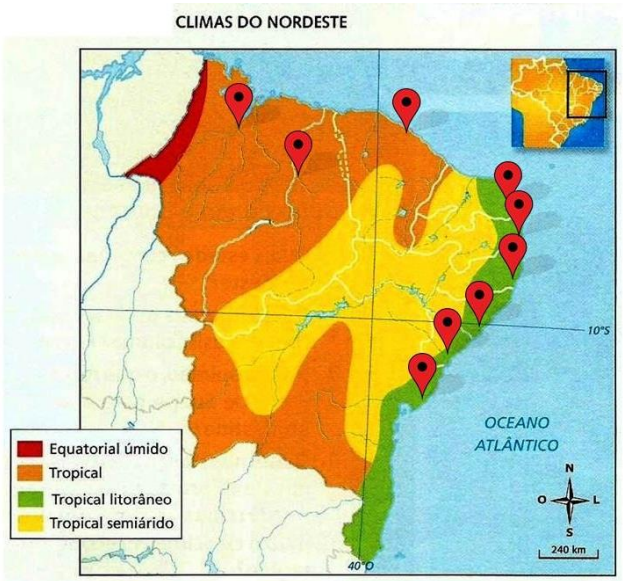
No R, foi trabalhado inicialmente com a Biblioteca POT, esta biblioteca foi criada por Mathieu Ribatet, e se utilizou dela para obter estimativas do limiar, através do DIP plot, pois o DIP é uma técnica gráfica utilizada na escolha do limiar, de acordo com o trabalho de Cunnane[1979](Citado por De Nascimento[2012]), que diz que o número de excessos sobre um limiar alto em um determinado período (geralmente meses ou anos), pode ser distribuído através um processo de Poisson. Assim, a razão entre a variância e a média é igual a 1. Como principal objetivo do trabalho, o uso do Pacote MCMC4Extremes se deu a partir de gráficos de retorno, distribuição preditiva dos dados e histogramas. Este pacote foi criado em Abril de 2015 por Nascimento e Moura e Silva e atualizado em Julho de 2016, ele fornece funções para estimação de

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

distribuições posteriores dos parâmetros das distribuições GEV e GPD. Além disso, o pacote também retorna medidas de ajuste AIC, BIC, DIC.

Para trabalhar com o R foram obtidos dados meteorológicos de precipitação de chuva de todas as capitais do Nordeste, onde as informações foram coletadas por estatísticas diárias, entre os anos de 1961 á 2015. Para fazer as análises, foram obtidos apenas os dados referentes aos meses que apresentavam temperatura mais elevada em cada Capital. A Figura 1 mostra o mapa climático da Região Nordeste, com as capitais utilizadas no estudo.

Figura 1: Mapa Climático da Região Nordeste

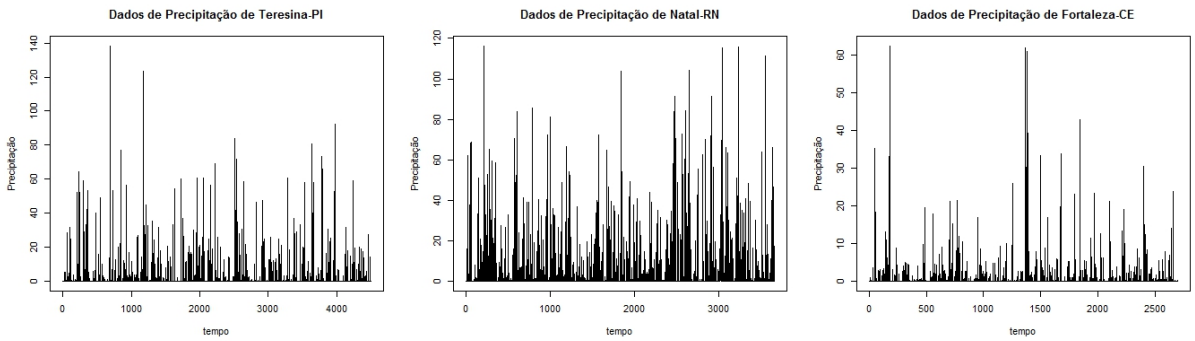


Fonte: FERREIRA, Graça M. L. *Atlas geográfico:* espaço mundial. 3. ed. São Paulo: Moderna, 2010. p. 123.

Resultados e Discussão:

Na Figura 2 temos as séries temporais dos dados diários dos meses mais quentes da cidade de Teresina-PI, Natal-RN e Fortaleza-CE, respectivamente. O comportamento das séries é sazonal, podemos perceber estacionariedade em todos os gráficos.

Figura 2: Observações de Precipitação Máxima Diária



As estações meteorológicas onde foram coletados os dados se encontram operantes, e registram diariamente o acúmulo de precipitação de chuva das capitais do nordeste, no presente artigo analisaremos as cidades de Teresina-PI, Natal-RN e Fortaleza-CE.

Teresina-PI

Teresina é a capital e o município mais populoso do estado brasileiro do Piauí. Localiza-se no Centro-Norte Piauiense a 353 km do litoral, sendo, portanto, a única capital da Região Nordeste que não se localiza as margens do Oceano Atlântico.(Wikipédia[a],[2016]). Teresina se caracteriza com duas estações: o período das chuvas (que ocorrem no verão e outono) e o período seco (que ocorre no inverno e primavera). De janeiro a maio, devido às chuvas, o clima é quente e úmido (porém, há possibilidade de ocorrer neblina nas manhãs); de junho a agosto o clima começa a ficar mais seco com noites relativamente frias; de setembro a dezembro o clima se torna mais quente e abafado, podendo começar a ocorrer algumas pancadas de chuva a partir de novembro. (Wikipédia[a],[2016]).

A análise estatística foi feita apenas com os meses que apresentaram temperatura mais quente em Teresina durante o ano. No total, foram analisadas 4.504 observações, compreendidas entre os meses de setembro a dezembro dos anos de 1961 à 1967, de 1976 à 1984, de 1993 e de 1995 até o ano de 2015.

Natal-RN

Natal é um município brasileiro, capital do estado do Rio Grande do Norte, Região Nordeste do país. Pertence à Mesorregião do Leste Potiguar e à Microrregião de Natal. (Wikipédia[a],[2016]). Segundo dados do Instituto Nacional de Meteorologia (INMET), o clima de Natal é o tropical chuvoso quente com verão seco, com temperatura média anual de 26 ^{o}C.

As precipitações acontecem sob a forma de chuva, que podem vir acompanhadas de raios e trovoadas e ainda serem de forte intensidade. O índice pluviométrico médio é de 1 465 mm/ano, concentrados entre os meses de março e julho (Wikipédia[a],[2016]).

A análise estatística foi feita apenas com os meses que apresentam temperatura mais quente em Natal durante o ano. No total, foram analisadas 3.658 observações, compreendidas entre os meses de janeiro a março dos anos de 1961 à 1970, de 1984 e de 1986 até o ano de 2015.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Fortaleza-CE

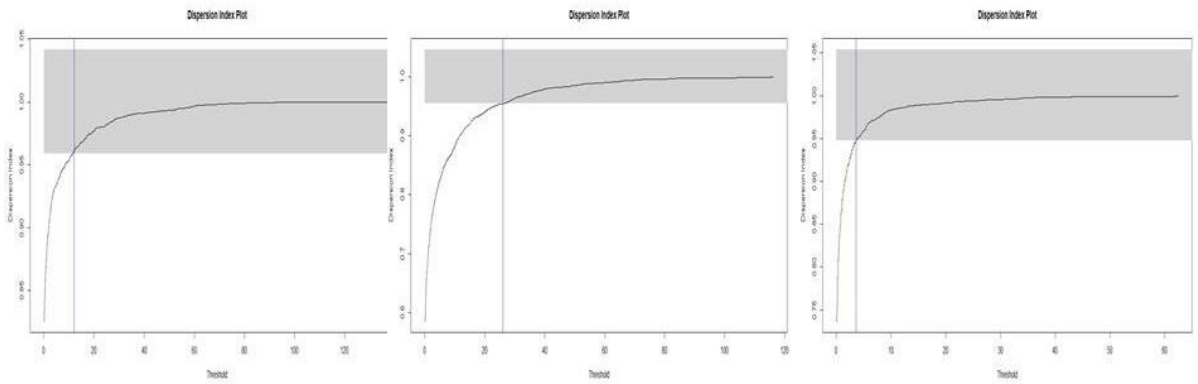
Fortaleza é um município brasileiro, capital do estado do Ceará. Está localizada no litoral Atlântico. (Wikipédia[a],[2016]). Segundo dados do Instituto Nacional de Meteorologia (INMET), Fortaleza possui clima tropical semiúmido, com temperatura média anual de 26,5 ºC.

Sem ter exatamente definidas as estações do ano, há a estação das chuvas, de janeiro a junho (verão e outono), julho é a transição da estação chuvosa para a seca e a estação seca, de agosto a dezembro (inverno e primavera). A média pluviométrica anual é de aproximadamente 1 600 milímetros (mm), concentrados entre fevereiro e maio, sendo abril o mês de maior precipitação (356 mm). Sua localização, entre serras próximas, faz com que as chuvas de verão ocorram com mais frequência na cidade e entorno do que no resto do estado (Wikipédia[a],[2016]).

A análise estatística foi feita apenas com os meses que apresentam temperatura mais quente em Fortaleza durante o ano. No total, foram analisadas 2.701 observações, compreendidas entre os meses de novembro e dezembro dos anos de 1961 à 1985 e de 1994 até o ano de 2015.

A Figura 3 mostra os gráficos de determinação do limiar para os dados de precipitação de algumas capitais do Nordeste (Teresina-PI, Natal-RN e Fortaleza-CE). Pelo DIP plot, está curva de aceitação da GPD no limiar a partir de aproximadamente 12, 26 e 3.6. Para este limiar, colhemos 166 observações maiores, o que diz que a escolha deste limiar pode ser ajustada para o conjunto de dados.

Figura 3: Método Gráfico de determinação do limiar

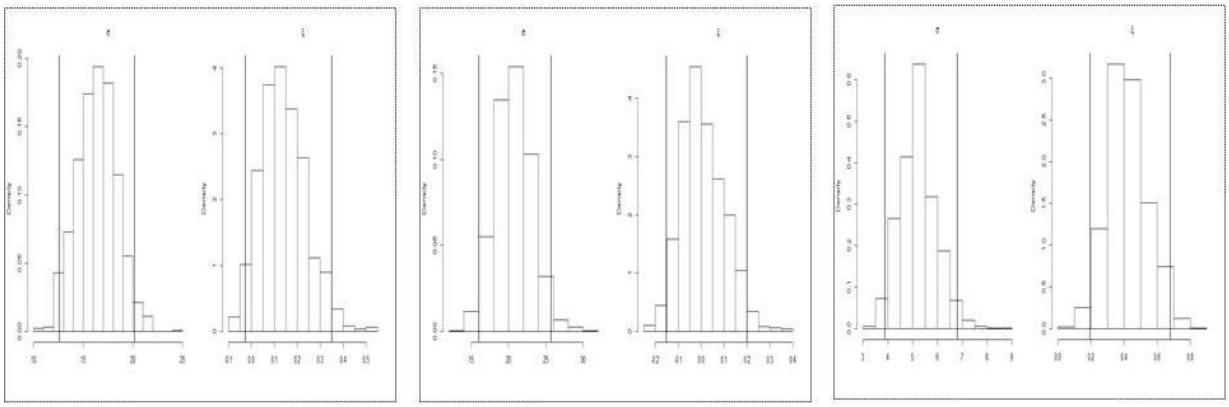


A Figura 4 mostra a distribuição posteriori dos parâmetros da GPD para os dados de precipitação diária dos meses mais quentes, da cidade de Teresina-PI, Natal-RN e Fortaleza-CE, considerando um limiar encontrado pelo DIP plot, respectivamente igual a 12, 26 e 3.6. Mutuamente para essas cidades os pontos no gráfico da média

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

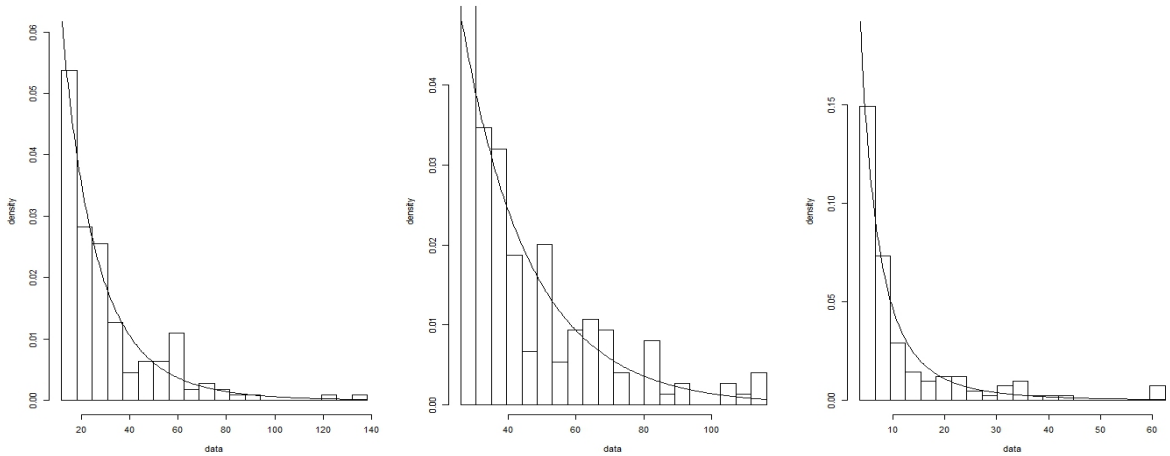
posteriori dos parâmetros são σ= 16.38 e ε= 0.14, σ= 20.67 e ε= 0.011, σ= 5.18 e ε= 0.4265 com IC de 95\%.

Figura 4: Histograma da distribuição posteriori dos parâmetros da GPD, com intervalo de credibilidade de 95%



A Figura 5 mostra a curva da distribuição preditiva ajustada dos dados e o histograma dos dados acima do limiar. Podemos perceber que os modelos estão bem ajustados, pois, a curva está muito próxima das frequências do histograma, logo, os parâmetros estão bem estimados.

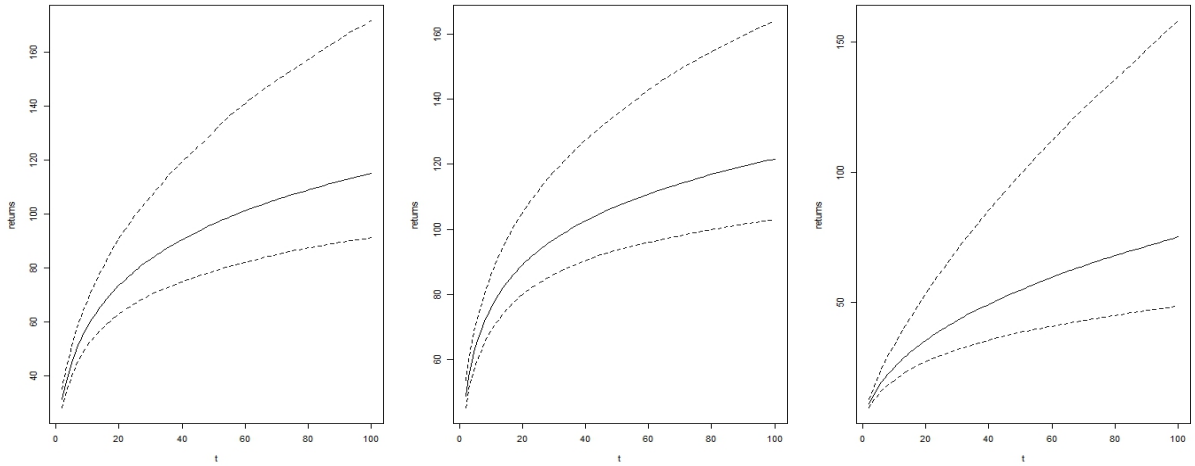
Figura 5: Distribuição preditiva ajustada dos parâmetros da distribuição GPD



A figura 6 mostra o gráfico de retorno esperado em t períodos de tempo, como estes dados são trabalhados com valores mensais, pode-se dizer que a cada t=12 meses, é esperado que pelo menos uma vez as precipitações máximas mensais nas cidades de Teresina-PI, Natal-RN e Fortaleza-CE sejam maiores ou iguais a 63.51338 mm, 90.06142 mm e 35.96544 mm, respectivamente. De modo relativo a cada 100 meses é esperado que pelo menos uma vez as precipitações máximas mensais sejam

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

maiores ou iguais a 116.00690 mm, 121.73228 mm e 75.81288 mm, reciprocamente nas mesmas cidades



Conclusão:

A Teoria dos Valores Extremos tem contribuído bastante na solução de diversos problemas práticos, e a previsão de eventos de estiagem ou de grandes tempestades é muito importante. Com isso, o projeto foi muito enriquecedor para o conhecimento sobre os fundamentos da Teoria de Valores Extremos, e sua aplicação se deu de forma muito prática. O pacote MCMC4Extremes foi atualizado no decorrer do trabalho, e suas funções se tornaram mais simples, além disso, também foi possível perceber que o pacote ainda pode melhorar.

Após a análise inicial, percebeu-se que o valor do limiar determinado pelo DIP plot foi bem estimado, com isso tivemos uma proporção boa de observações acima do limiar em todas as capitais, isso implica na boa estimação dos parâmetros da GPD, pelo pacote MCMC4Extremes conseguimos então fazer gráficos de predição e de retorno, esses gráficos se mostraram eficientes por meio do comportamento da distribuição preditiva da cauda.

Contudo, os resultados foram bastante satisfatórios, e a modelagem de valores extremos de precipitação das capitais da região Nordeste foi bem estimada.

Referências:

DO NASCIMENTO, F.F. **Modelos Probabilísticos Para Dados Extremos: Teoria e Aplicações:** Teresina, Edufpi, 2012;

MORETTIN, P. A; TOLOI, C.M.C. **Análise de Séries Temporais.**Editora Blucher, 2004;

H. BOLFARINE, H.; SANDOVAL, M.C. **Introdução á Inferência Estatística,** ed. Sociedade Brasileira de Matemática, 2001;

MENDES, B.V.M. Introdução a análise de eventos extremos, Rio de Janeiro, E-papers, 2004;

Wikipédiaª – *Natal* (s.d.), Consultado em 27 de novembro de 2016. Acesso em: https://pt.wikipedia.org/wiki/Natal;

Wikipédiaª – *Fortaleza* (s.d.), Consultado em 27 de novembro de 2016. Acesso em: https://pt.wikipedia.org/wiki/Fortaleza;

Wikipédiaª – *Teresina* (s.d.), Consultado em 27 de novembro de 2016. Acesso em: https://pt.wikipedia.org/wiki/Natal;

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ID 9 - APLICAÇÃO DA COMPOSIÇÃO PROBABILÍSTICA DE PREFERÊNCIAS E DO ÍNDICE DE GINI À ESCOLHA DE JOGADORES DA LIGA INGLESA DE FUTEBOL

Luiz Octávio Gavião[57]
Vitor Ayres Principe[58]
Gilson Brito Alves Lima[59]
Annibal Parracho Sant'Anna[60]

Resumo

Essa pesquisa teve por objetivo aplicar a Composição Probabilística de Preferências Tricotômica (CPP-Tri) com o índice de Gini, para a escolha de jogadores da Liga Inglesa de Futebol, com a finalidade de apoiar a decisão em investimentos no esporte. O CPP é um método probabilístico de apoio à decisão multicritério, sendo o CPP-Tri uma variante utilizada para a classificação de alternativas. O Índice de Gini é uma medida de desigualdade, sendo amplamente utilizada para ordenar países com base na desigualdade de renda das populações. O modelo proposto foi aplicado a um conjunto de 25 jogadores do setor de defesa das equipes, que se destacaram em times emergentes durante a temporada 2015-2016. O procedimento de classificação permitiu selecionar a melhor sub amostra de defensores, composta por oito jogadores. Dentre esses, o procedimento de ordenação priorizou a regularidade ao longo do campeonato, indicando o defensor D.19 como o mais regular nos fundamentos técnicos selecionados. A metodologia proposta permitiu identificar o jogador de maior potencial para a temporada seguinte, por ser o mais regular dentre os jogadores de defesa, pertencentes à classe de melhor desempenho.

Palavras-Chave: CPP; CPP-Tri; Índice de Gini; Futebol.

Abstract

This research aimed to apply the Composition of Probabilistic Preferences Trichotomous (CPP-Tri) with the Gini index for choosing Premier League players, in order to support the investment decision in football. The CPP is a probabilistic method of multi-criteria decision support. The CPP-Tri is a variant used for classifying alternatives. The Gini Index is a measure of inequality, and is widely used to evaluate the income inequality of countries. The proposed model was here applied to a set of 25 defense players of English teams, who have excelled in emerging squads during the 2015-2016 season. The classification procedure allowed select the best sub sample of defenders, comprised of eight players. Among these, the procedure of ordering prioritized regularity throughout the League, indicating the 19th defender as the more regular in selected technical fundamentals. The proposed methodology made it possible to identify the most potential player for the following season, as the more regular among defensive players, belonging to the class of better performance.

Keywords: CPP; CPP-Tri; Gini Index; Soccer.

[57] Universidade Federal Fluminense (UFF), luiz.gaviao67@gmail.com
[58] Nova Information Management School (NOVA IMS), vitorprin@gmail.com
[59] Universidade Federal Fluminense (UFF), glima@id.uff.br
[60] Universidade Federal Fluminense (UFF), annibal.parracho@gmail.com

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Introdução

A gestão de equipes de futebol envolve significativos investimentos ao final da temporada de torneios. A formação de uma equipe competitiva pode garantir mais recursos, a partir de vitórias nos diferentes campeonatos disputados, maior prestígio mundial às ligas de clubes e ampliação da parcela estimada em quatro bilhões de fãs que consomem produtos e serviços do esporte mais popular do mundo (CURLEY; ROEDER, 2016; HATZIGEORGIOU, 2016; KUPER; SZYMANSKI, 2014).

O futebol é um esporte complexo e não raramente são obtidos resultados surpreendentes com equipes modestas. Os times e jogadores que se destacam na temporada se tornam visados por clubes de maior investimento. Um bom investimento no mercado do futebol pode ser associado à formação de equipes campeãs com recursos limitados (LOUZADA et al., 2014).

A escolha dos jogadores para compor as equipes nas temporadas seguintes apresenta as características de um problema de apoio à decisão multicritério. O bom desempenho em alguns fundamentos técnicos do esporte são normalmente alternados por aproveitamentos insatisfatórios em outros. Um atacante que marcou gols acima da média pode apresentar desempenho ruim em passes e assistências a gol, por exemplo (CURLEY; ROEDER, 2016; KUPER; SZYMANSKI, 2014). Essa irregularidade no desempenho em diferentes critérios demanda uma solução de compromisso entre as medidas conflitantes (POMEROL; BARBA-ROMERO, 2012).

Dentre as diversas metodologias multicritério adequadas ao contexto, destaca-se o procedimento de classificação ordenada (POMEROL; BARBA-ROMERO, 2012). Inicialmente, essa pesquisa aplicou o método da Composição Probabilística de Preferências Tricotômico (CPP-Tri) para alocar jogadores em diferentes classes de desempenho global (SANT'ANNA; COSTA; PEREIRA, 2012). Posteriormente, ordenou-se o conjunto de jogadores da maior classe com base no índice de Gini, que descreve o grau de desigualdade de um conjunto de dados (GINI, 1921).

O uso do CPP-Tri com o índice de Gini potencializa a capacidade de escolha de jogadores promissores. O procedimento de classificação reúne os jogadores por classes, diferenciadas por perfis pré-determinados, tornando possível a identificação de um conjunto de jogadores com melhor performance. O índice de Gini identifica a desigualdade de um conjunto de valores e, dessa forma, permite ordenar os atletas por regularidade em todos os fundamentos, nesse caso, do menor ao maior índice de

Gini. Assim, o clube contratante pode identificar a categoria dos melhores jogadores em sua posição e, simultaneamente, a regularidade em diferentes fundamentos, o que torna o elenco versátil aos treinadores. Essa versatilidade está associada à possibilidade de escalar um jogador com desempenho técnico elevado e regular em diferentes posições. Em competições longas e de alto desgaste físico dos atletas, por vezes é necessário adaptar ou mesmo improvisar jogadores em posições nas quais não estão familiarizados. Na perspectiva proposta com o CPP-Tri e o índice de Gini, o jogador da equipe emergente com maior potencial em uma determinada posição (i.e. goleiro, defensor, meia e atacante) é aquele que pertencer à classe mais elevada de atributos e possuir o menor índice de Gini.

Esse estudo foi aplicado à primeira divisão da Liga Inglesa de Futebol, para a temporada 2015-2016. Essa liga coleta, registra e publica os dados dos clubes e jogadores nas partidas, em diversos fundamentos do esporte (PREMIER-LEAGUE, 2016). Em paralelo à facilidade de acesso aos dados, o elevado investimento dos clubes impacta na qualidade da Liga Inglesa, ampliando a competitividade dos torneios e o desempenho dos atletas (CURLEY; ROEDER, 2016).

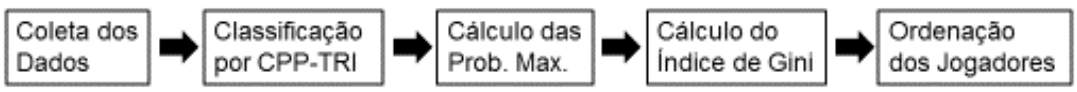
Objetivo

A pesquisa tem por objetivo aplicar uma combinação da metodologia multicritério de classificação ordenada (CPP-Tri) com o índice de Gini, para o processo de escolha de jogadores promissores em diferentes fundamentos, das equipes emergentes da primeira divisão da Liga Inglesa de futebol, com a finalidade de apoiar a decisão em investimentos.

Material e Métodos

Essa pesquisa se desenvolveu em cinco etapas, conforme a Figura 1.

Figura 1 - Etapas do método



A pesquisa foi iniciada com a coleta dos dados publicados na central de dados da Liga Inglesa (PREMIER-LEAGUE, 2016). Na aplicação do método proposto, foram selecionados 25 jogadores do setor de defesa (i.e. laterais e zagueiros), que se destacaram ao longo da temporada, avaliados em 12 critérios. Por não pertencerem

aos cinco clubes de maior investimento (i.e. Manchester United, Manchester City, Arsenal, Liverpool e Chelsea), considerou-se que esses jogadores apresentaram potencial para a transferência a esses clubes. Os seguintes fundamentos técnicos individuais foram explorados: gols sofridos, interceptações, alívio de pressão na zaga, disputas vencidas, disputas perdidas, divididas vencidas, disputas de cabeça vencidas, disputas de cabeça perdidas, passes certos, lançamentos, faltas cometidas, e cartões. Os dados foram padronizados conforme o número de partidas de cada jogador. Os critérios "gols sofridos", "disputas perdidas", "faltas cometidas" e "cartões" apresentam impacto negativo aos resultados, sendo então multiplicados por (-1) para a modelagem. A base de dados encontra-se no Apêndice I.

Na segunda etapa foi aplicada a metodologia de classificação probabilística dos dados. O método CPP-Tri foi proposto por Sant'Anna, Costa e Pereira (2012), com base no método CPP (SANT'ANNA; SANT'ANNA, 2001), com diferentes aplicações (SANT'ANNA, 2014; SANT'ANNA; COSTA; PEREIRA, 2015). O CPP-Tri apresenta finalidade similar ao método ELECTRE Tri (YU, 1992), porém com abordagem probabilística dos dados, tornando-o relevante a contextos de elevada incerteza e imprecisão dos dados. A programação em linguagem "R" encontra-se no Apêndice II, sendo codificado em imagem "QR" por necessidade de constrição do texto.

O CPP utiliza a abordagem probabilística em um problema multicritério, em que as avaliações numéricas iniciais de cada alternativa (e.g. jogadores) segundo cada critério (e.g. fundamentos técnicos) são tratadas como parâmetros de locação de distribuições de probabilidade. No caso aqui explorado, assumiu-se, além de distribuições normais, idêntica distribuição e independência entre as perturbações que provocam a imprecisão nas medidas, e variâncias determinadas pelos dados disponíveis. (SANT'ANNA; COSTA; PEREIRA, 2012).

O cálculo do CPP-Tri requer a definição antecipada das classes a categorizar as alternativas. As classes são identificadas por perfis representativos, que podem ser construídos com base em valores numéricos que, segundo cada critério, caracterizariam uma alternativa em diferentes níveis de qualidade (i.e. excelente, ótima, boa, dentre outras). No problema em questão foram atribuídas cinco classes, sendo representadas por um único perfil. Essas cinco classes foram definidas por quantis de 10%, 30%, 50%, 70% e 90%, conforme proposto por Sant'Anna (2015).

Para efetuar a classificação, são calculadas as probabilidades de as alternativas apresentarem valores (X_j) acima (A^+) e abaixo (A^-) dos perfis das classes

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

(Y_{ihj}), conforme as Equações (1) e (2), para a i-ésima alternativa, j-ésimo critério e h-ésimo perfil. O menor valor absoluto da diferença entre essas probabilidades (i.e. $|A^{+}-A^{-}|$) determina a classificação da alternativa (SANT'ANNA; COSTA; PEREIRA, 2012). A programação do CPP-Tri em "R" encontra-se no Apêndice II.

$$A_{ij}^{+} = \prod_{h} P\left[X_j > Y_{ihj} \right] \tag{1}$$

$$A_{ij}^{-} = \prod_{h} P\left[X_j < Y_{ihj} \right] \tag{2}$$

Na terceira etapa, as avaliações dos jogadores classificados na mais elevada categoria de desempenho (i.e. Classe 4) foram transformadas em probabilidades de maior preferência em cada critério. Esse procedimento de cálculo do CPP assume por premissa a imprecisão dos dados e a natureza de incerteza do problema em análise. As probabilidades de maximizar (M_{ij}) a preferência da i-ésima alternativa do j-ésimo critério e a de minimizar (m_{ij}) essa preferência são definidas com as equações (3) e (4), em que f_X representa a função de densidade de probabilidade (pdf) da i-ésima alternativa, D_{Xi} o seu suporte e F_X a função de distribuição cumulativa das demais alternativas (cdf), indicadas com a notação (-*i*) (SANT'ANNA et al., 2012).

$$M_{ij} = \int_{D_{X_i}} \left[\prod F_{X_{-i}}(x_{-i}) \right] f_{X_i}(x_i) dx_i \tag{3}$$

$$m_{ij} = \int_{D_{X_i}} \left[\prod \left(1 - F_{X_{-i}}(x_{-i})\right) \right] f_{X_i}(x_i) dx_i \tag{4}$$

Na quarta etapa, o cálculo dos índices de Gini foi efetuado a partir do pacote "ineq" do software "R" (R-CORE-TEAM, 2016). O índice de Gini foi aplicado às probabilidades obtidas na etapa anterior, conforme os Apêndice III e IV. Para uma amostra de dados, esse índice pode ser calculado a partir da Equação (5) (DIXON et al., 1987), em que y_i representa os valores da amostra, para $i=1$ a n e dados dispostos de forma crescente ($y_i \leq y_{i+1}$),

$$G = \frac{1}{n}\left(n+1-2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right) \tag{5}$$

Por fim, os índices de Gini foram ordenados do menor ao maior, indicando os jogadores mais regulares da maior classe, conforme descritos na Seção seguinte.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

Resultados e Discussão

A Tabela 1 apresenta os resultados do CPP-Tri aos 25 defensores. Os valores correspondem ao produto das probabilidades de cada jogador estar posicionado acima ou abaixo do perfil de cada classe, por critério. Dessa forma, o jogador D.3 apresenta 1,6E-10 de probabilidade de estar acima e 0,044 de probabilidade de estar abaixo do perfil da 5ª classe, que representa a categoria mais elevada em qualidade de desempenho dos atletas.

Tabela 1 - Resultados do CPP-Tri

JOG	Classe 1		Classe 2		Classe 3		Classe 4		Classe 5		Res.
	Acima	Abaixo	Acima	Abaixo	Acima	Abaixo	Acima	Abaixo	Acima	Abaixo	
D.1	3,9E-02	6,8E-12	4,4E-03	4,5E-09	1,5E-04	6,9E-07	2,4E-06	3,9E-05	1,6E-09	7,7E-03	4
D.2	6,7E-02	6,5E-12	1,1E-02	4,5E-09	6,9E-04	1,2E-06	2,2E-05	1,1E-04	7,3E-09	9,6E-03	4
D.3	2,7E-02	9,5E-10	3,4E-03	4,1E-07	1,4E-04	5,8E-05	2,1E-06	1,9E-03	1,6E-10	4,4E-02	3
D.4	3,8E-02	2,0E-11	6,1E-03	1,6E-08	2,5E-04	2,6E-06	4,0E-06	1,4E-04	1,8E-09	1,5E-02	4
D.5	8,1E-03	5,1E-09	6,4E-04	1,3E-06	1,0E-05	7,2E-05	4,5E-08	1,2E-03	5,4E-12	6,1E-02	3
D.6	5,6E-02	5,5E-11	1,1E-02	5,0E-08	6,9E-04	1,1E-05	1,6E-05	5,7E-04	3,0E-09	2,2E-02	4
D.7	6,1E-04	6,1E-07	1,8E-05	5,5E-05	1,0E-07	1,3E-03	1,6E-10	1,0E-02	1,9E-15	1,4E-01	2
D.8	3,0E-02	3,9E-10	3,8E-03	1,7E-07	1,2E-04	1,9E-05	1,4E-06	6,4E-04	3,3E-10	3,7E-02	3
D.9	7,2E-03	1,1E-09	5,6E-04	2,8E-07	1,0E-05	2,5E-05	1,2E-07	1,1E-03	1,2E-11	3,7E-02	3
D.10	1,4E-02	4,2E-10	9,6E-04	1,1E-07	2,7E-05	1,4E-05	3,4E-07	6,0E-04	4,1E-11	2,7E-02	3
D.11	1,4E-02	5,4E-10	1,2E-03	1,9E-07	4,6E-05	3,1E-05	6,1E-07	1,1E-03	2,4E-11	2,6E-02	3
D.12	5,7E-02	4,0E-12	1,1E-02	4,5E-09	8,8E-04	1,6E-06	2,7E-05	1,4E-04	6,4E-09	8,2E-03	4
D.13	9,3E-03	9,3E-08	6,5E-04	1,5E-05	1,2E-05	7,2E-04	6,1E-08	9,7E-03	2,5E-12	1,5E-01	2
D.14	8,6E-03	2,0E-09	7,9E-04	7,3E-07	1,7E-05	5,9E-05	1,2E-07	1,5E-03	9,1E-12	4,5E-02	3
D.15	1,2E-01	6,6E-12	2,6E-02	6,3E-09	2,0E-03	1,7E-06	5,8E-05	1,1E-04	4,2E-08	1,3E-02	4
D.16	4,1E-02	8,0E-14	3,7E-03	5,4E-11	2,1E-04	2,1E-08	7,0E-06	3,1E-06	3,7E-09	1,1E-03	4
D.17	5,1E-03	8,6E-08	3,4E-04	1,6E-05	4,4E-06	6,6E-04	1,5E-08	7,9E-03	5,8E-13	1,3E-01	2
D.18	1,9E-02	1,8E-09	1,8E-03	5,3E-07	4,0E-05	3,9E-05	3,1E-07	9,4E-04	6,1E-11	5,4E-02	3
D.19	9,8E-02	1,2E-11	1,3E-02	6,3E-09	9,5E-04	1,8E-06	3,4E-05	1,6E-04	1,4E-08	1,3E-02	4
D.20	1,9E-03	2,8E-08	7,1E-05	3,8E-06	1,3E-06	2,8E-04	5,5E-09	4,7E-03	4,3E-14	5,4E-02	2
D.21	7,2E-03	1,5E-07	4,6E-04	2,3E-05	9,6E-06	1,2E-03	5,1E-08	1,6E-02	8,2E-13	1,5E-01	2
D.22	3,2E-02	9,0E-10	3,8E-03	3,5E-07	2,1E-04	6,2E-05	3,3E-06	2,0E-03	2,0E-10	4,0E-02	3
D.23	1,4E-02	3,9E-09	1,1E-03	8,4E-07	3,7E-05	1,0E-04	3,9E-07	2,9E-03	1,3E-11	4,8E-02	3
D.24	1,6E-02	3,4E-09	1,1E-03	6,2E-07	3,2E-05	7,7E-05	2,8E-07	1,9E-03	1,1E-11	3,2E-02	3
D.25	3,7E-02	1,2E-10	4,8E-03	5,8E-08	2,1E-04	1,1E-05	2,8E-06	4,0E-04	4,3E-10	1,6E-02	3

Entretanto, a definição das classes depende da diferença absoluta entre as probabilidades de estar acima e abaixo de cada classe. Por exemplo, para o jogador D.6, os módulos dessas diferenças equivalem a 5,55E-02 (Classe 1); 1,07E-02 (Classe 2); 6,81E-04 (Classe 3); 5,56E-04 (Classe 4) e 2,17E-02 (Classe 5). Assim, verifica-se que o menor módulo determina a localização do jogador D.6 na Classe 4, conforme descrito na coluna “Res”. Cabe também ressaltar que não houve jogadores distribuídos nas classes extremas 1 e 5.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

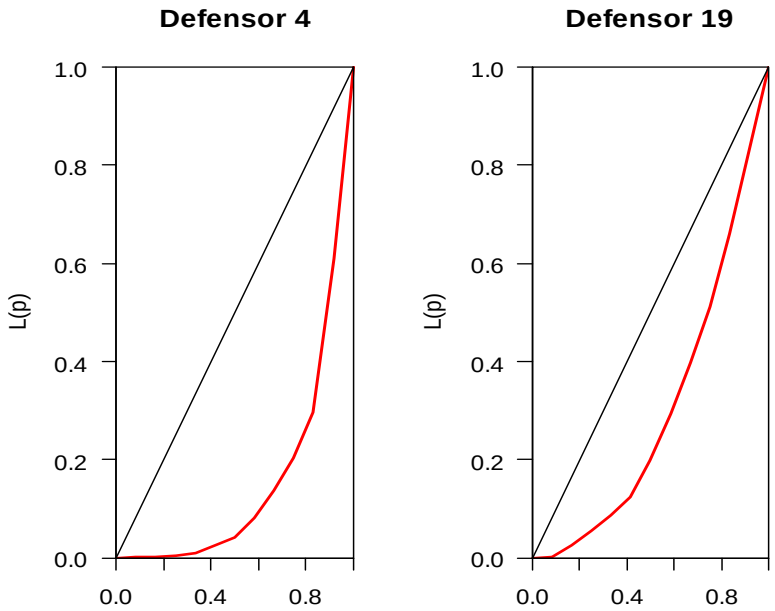
A Tabela 2 apresenta os resultados do índice de Gini para os jogadores posicionados na Classe 4.

Tabela 2 - Resultados do índice de Gini

Jogador	Time	Classe	Índice de Gini	Rank
D.1	Time E	4	0,63345	7
D.2	Time E	4	0,47695	3
D.4	Time E	4	0,68168	8
D.6	Time E	4	0,55045	6
D.12	Time F	4	0,50725	4
D.15	Time A	4	0,45462	2
D.16	Time A	4	0,52923	5
D.19	Time A	4	0,38611	1

A coluna “Rank” indica a ordenação dos jogadores a partir dos resultados do índice de Gini. O menor índice obtido pelo jogador D.19 representa a maior regularidade nos diferentes critérios. Dessa maneira, verifica-se que na classe dos melhores defensores, esse jogador foi o mais homogêneo tecnicamente, representando o melhor potencial de investimento para a temporada seguinte.

Figura 2 - Curvas de Lorenz



A representação gráfica na Figura 2 traduz a diferença entre o jogador mais regular e o mais irregular da Classe 4. As curvas de Lorenz são frequentemente utilizadas para descrever o índice de Gini. As diferenças de regularidade entre os jogadores D.4 (menor) e D.19 (maior) podem ser observadas ao se comparar a linha diagonal, que indica a máxima regularidade, com as curvas. A diagonal representa a

igualdade nas avaliações em todos os critérios, que equivale à desigualdade nula. Quanto maior a distensão da curva de Lorenz, indicando uma "folga" acentuada entre os extremos, maior a desigualdade dos valores da amostra.

Conclusão

Essa pesquisa teve por finalidade aplicar o CPP-Tri com o índice de Gini, para a escolha de atletas promissores em diferentes fundamentos de uma modalidade esportiva, com a finalidade de apoiar a decisão em investimentos. Essa metodologia foi aplicada a um conjunto de 25 jogadores de defesa da primeira divisão da Liga Inglesa de Futebol, que se destacaram em times emergentes durante a temporada 2015-2016.

Os resultados demonstraram a adequação da abordagem multicritério com método de classificação ordenada. O procedimento de classificação permitiu selecionar a melhor sub amostra de defensores, que compuseram a Classe 4. Dentre esses, o procedimento de ordenação priorizou a regularidade ao longo do campeonato, indicando o defensor 19 como o jogador de maior potencial de sucesso na temporada seguinte.

O modelo proposto permitiu identificar a categoria dos melhores jogadores de defesa e, simultaneamente, os mais regulares em diferentes fundamentos técnicos. Dessa forma, as comissões técnicas podem identificar a versatilidade de seus elencos. Essa característica está associada à possibilidade de escalar um jogador com desempenho técnico elevado e regular em diferentes posições. Em competições longas e de alto desgaste físico dos atletas, por vezes é necessário adaptar ou mesmo improvisar jogadores em posições nas quais não estão familiarizados.

Para estudos futuros, visualiza-se a aplicação com jogadores de outras posições e uso do método em outras modalidades esportivas. Também é possível variar os critérios de desempenho técnico conforme as diferentes posições em campo, pois certas características dos atacantes, por exemplo, não se aplicam aos goleiros. Outro aperfeiçoamento à base de dados consiste em padronizar os desempenhos por minutos jogados, ao invés do número de partidas, contribuindo para a maior precisão dos dados iniciais.

Referências

CURLEY, J. P.; ROEDER, O. English Soccer's Mysterious Worldwide Popularity. **Contexts**, v. 15, n. 1, p. 78–81, 2016.

DIXON, P. M. et al. Bootstrapping the Gini coefficient of inequality. **Ecology**, v. 68, n. 5, p. 1548–1551, 1987.

GINI, C. Measurement of inequality of incomes. **The Economic Journal**, v. 31, n. 121, p. 124–126, 1921.

HATZIGEORGIOU, A. Can Sports Promote Exports? The Role of Soccer Matches in International Trade. **Global Economy Journal**, v. 16, n. 1, p. 1–32, 2016.

KUPER, S.; SZYMANSKI, S. **Soccernomics: Why England Loses, Why Germany and Brazil Win, and Why the U.S., Japan, Australia, Turkey - and Even Iraq - Are Destined to Become the Kings of the World's Most Popular Sport**. New York: Nation Books, 2014.

LOUZADA, L. M. et al. Return on Investment in Sports Marketing Initiatives : A Study Focusing on a Brazilian Soccer Team. **International Journal of Business Administration**, v. 5, n. 5, p. 71–83, 2014.

POMEROL, J.-C.; BARBA-ROMERO, S. **Multicriterion decision in management: principles and practice**. New York: Springer, 2012. v. 25

PREMIER-LEAGUE. **Stats Centre - Football Association Premier League**. Disponível em: <https://www.premierleague.com/stats>. Acesso em: 11 nov. 2016.

R-CORE-TEAM. **R: A language and environment for statistical computing.http://www. R-project. org**Vienna, Austria, 2016.

SANT'ANNA, A. P. et al. Análise multicritério baseada em probabilidades de preferência. In: OLIVEIRA, V. F. DE; CAVENAGHI, V.; MÁSCULO, F. S. (Eds.). . **Tópicos emergentes e desafios metodológicos em Engenharia de Produção: casos, experiências e proposições - Volume V**. Rio de Janeiro: ABEPRO, 2012. p. 258.

SANT'ANNA, A. P. **Aplicação do CPP-Tri à classificação dos países pelos critérios do IDH**. Encontro Nacional de Engenharia de Produção - XXXVI ENEGEP. **Anais**...Curitiba: 2014

SANT'ANNA, A. P. **Probabilistic Composition of Preferences, Theory and Applications**. New York: Springer, 2015.

SANT'ANNA, A. P.; COSTA, H. G.; PEREIRA, V. CPP-Tri: um método de classificação ordenada baseado em composição probabilística. **Relatórios de Pesquisa em Engenharia de Produção (UFF)**, v. 12, n. 8, p. 104–117, 2012.

SANT'ANNA, A. P.; COSTA, H. G.; PEREIRA, V. CPP-Tri: a sorting method based on the probabilistic composition of preferences. **International Journal of Information and Decision Sciences**, v. 7, n. 3, p. 193–212, 2015.

SANT'ANNA, A. P.; SANT'ANNA, L. A. F. P. **Randomization as a stage in criteria**

combining. International Conference on Industrial Engineering and Operations Management - VII ICIEOM. **Anais**...Salvador: 2001

YU, W. **ELECTRE Tri (aspects méthodologiques et manuel d'utilisation) Document- Université de Paris-Dauphine, LAMSADE,** 1992.

Apêndices

I - Base de dados

II - Script CPP-Tri

III - Prob. Max Classe 4

IV - Script Gini

[1] *p*-valores menores de 0,05 indicam a rejeição da hipótese nula para os testes estatísticos associados ao nível de 5%.

Anais do II Seminário Internacional de Estatística com R
The World of Big Data Analysis
Niterói-RJ-Brasil 23 e 24 de maio de 2017

ENCERRAMENTO – prof. Orlando Celso Longo

Prezados, boa noite,

Hoje chegamos ao encerramento do II Seminário Internacional de Estatística com R – II SER.

Não foi fácil organizar esta edição nas proporções que um evento deste porte demanda, sem recursos das agências de fomento, como já foi mencionado na abertura, não nos atenderam com repasses, apesar termos entrado com pedido de reconsideração, só a FAPERJ se pronunciou favorável, mas me parece que "morremos na praia"

Quero agradecer a Comissão Organizadora e Cientifica pelo enorme empenho para tornar realidade este evento, aos funcionários em particular a Elizete e ao corpo discente pela "garra" em especial a Raquel e Flavia.

Também, não vamos esquecer a nossa Coordenadora Profa. Luciane que foi incansável, que esteve sempre à frente de todos os momentos. Agradeço também aos docentes que ministraram os minicursos, oficinas, os coordenadores das apresentações orais e avaliadores de pôsteres.

Os nossos agradecimentos também, aos parceiros de outras instituições que nos apoiaram, já mencionados na abertura do evento, aos palestrantes nacionais e internacionais que nos abrilhantaram com palestras de alto nível.

O II SER, diferentemente do I SER, veio com inovações sem perder o alto nível e qualidade, foram inseridas apresentações orais de trabalhos com premiação e minicursos, além das sessões pôster com premiação já na primeira edição, palestras, teds e apresentação de blogs.

Desde o final do I SER, que as Comissões iniciaram a labuta com ideias, “brainstorming”, discursões, descartes/aceites, planejamento entre outras.

Por ser a UFF uma Universidade pulverizada na cidade de Niterói, tivemos que alocar as atividades em três Campis, com isto houve alguns desencontros, apesar dos cuidados que tivemos, alguns participantes e docentes de fora da UFF, se confundiram e foram para locais diferentes que não estavam programados.

Faremos o possível para que no próximo evento concentremos em um só Campus.

Comunico a todos que em breve teremos os anais à disposição de todos com ISSN já registrado no IBICT.

Informo ainda que será disponibilizado em forma de e-book os conteúdos das palestras do I SER.

Apesar de todos nós estarmos exaustos, estamos felizes pelo realizado, vejo no olhar de vocês um “gostinho de quero mais”, então, vamos levar para o III SER este gostinho.

Uma boa noite e retornem em paz.

Obrigado

Trabalhos Premiados

De acordo com a avaliação da Comissão Científica do II Seminário Internacional de Estatística com R, foram premiados os seguintes trabalhos:

Sessão de Comunicação Oral – categoria melhor artigo

1º. Colocado: Esquema Operacional de Baixo Custo para Verificação Estatística de Modelos Numéricos de Previsão do Tempo, de autoria de Nilza Barros da Silva e Natália Santos Lopes

2º. Colocado: Aplicação da Composição Probabilística de Preferências e do Índice de Gini à escolha de jogadores da Liga Inglesa de Futebol, de autoria de Luiz Octávio Gavião, Vitor Ayres Principe, Gilson Brito Alves Lima, Annibal Parracho Sant'Anna

3º. Colocado: Risco sistêmico na rede bancária brasileira: uma abordagem com Vine-cópula, de autoria de Andrea Ugolini e Miguel A. Rivera-Castro

Sessão Pôster – categoria melhor pôster

1º. Colocado: Impacto da Redução da Quantidade de Alternativas de um item do Enem na Estimação da Proficiência do Participante, de autoria de Alexandre Jaloto e Natália Caixeta Barroso

2º. Colocado: Shiny em Gráficos de Controle Estatístico de Processo, de autoria de Andréa Cristina Konrath, Rodrigo Gabriel de Miranda, Elisa Henning e Olga Maria Formigoni Carvalho Walter

3º. Colocado: The Drivers of Break-Even Inflation in Brazil: A Lasso Approach, de autoria de Daniel Karp, Luciano Vereda e Renato Lerípio.

Penúltimas revisões em 09/6/2017 por Luciane às 17:47; 16/6/2017 por Luciane às 10:25

Última revisão em 17/6/2017 por Luciane às 12:00