VOL 2 N.2 - 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.2 - 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
Elizete Oliveira

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.02 – ANO 2017

Resumos Simples 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

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

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:

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

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

RINITE CRÔNICA GRAVE E SUA ASSOCIAÇÃO COM CARACTERÍSTICAS CLÍNICAS, DEMOGRÁFICAS E DE SAÚDE.

Victor Côrtes Pourchet de Carvalho [1]
José Laerte Junior Boechat Morandi [2]
Beni Olej [3]
José Rodrigo de Moraes [4]

Resumo

Rinite é uma doença inflamatória crônica que afeta indivíduos de todas as idades ao redor do mundo. Apesar disso, é comumente sub-diagnosticada e tratada de forma inadequada. O objetivo do presente estudo é apontar a associação entre as características clínicas, demográficas e de saúde e a gravidade da rinite em pacientes acompanhados no Ambulatório de Alergia e Imunologia Clínica do HUAP/UFF, Niterói, RJ. Trata-se de um estudo transversal com pacientes com sintomas de rinite no período de 2013 a 2015. O desfecho é a gravidade da rinite em dois níveis: leve e grave. As variáveis explicativas são as características de saúde e demográficas e as características clínicas, tipo e sintomas da rinite. Usando modelo de regressão logística foram estimadas as medidas de razão de chance (OR) de o paciente ter rinite grave. Nas análises, apenas três variáveis estão associadas com a chance de o paciente apresentar rinite grave: grupo-etário / não-idosos (OR=1/0, 373=2,7; p-valor=0,037), obstrução nasal (OR=4,478; p-valor=0,006) e tipo de rinite, ou seja, rinite alérgica (OR=2,799; p-valor=0,045). Os pacientes com rinite alérgica, obstrução nasal e idade inferior a 60 anos apresentam maior gravidade da rinite.
Palavras-Chave: Rinite, Obstrução nasal, Idosos, Modelo logístico

Abstract

Rhinitis is an inflammatory and chronic condition that affects all-age subjects around the world. Nevertheless, rhinitis is generally underdiagnosed and misconducted. The aim of this present study is to identify the association between clinical, demographic and health characteristics and the severity of rhinitis in patients seen in the allergic and immunologic clinic of the Hospital Universitário Antonio Pedro/ UFF, Niteroi, Brazil. It is a cross-sectional study whereon patients with symptoms of rhinitis were included between 2013 and 2015. The end-point is the severity of rhinitis in two levels: mild and severe. The explanatory variables were the demographic and health characteristics as the clinical characteristics: the type and the presence of symptoms of rhinitis. It was estimated the odds-ratio (OR) of an individual to have severe rhinitis by using logistic regression model. In the analysis, only three variables were associated with the odds for each individual to have severe rhinitis: age group / on-elderly (OR=1/0, 373=2,7; p-value=0,037), nasal obstruction (OR=4,478; p-valor=0,006) and the type of rhinitis, that is, allergic rhinitis (OR=2,799; p-valor=0,045). Patients with allergic rhinitis, nasal obstruction and younger than 60 years have more severe forms of rhinitis.
Keywords: Rhinitis, Nasal obstruction, Elderly, Logistic model

[1] Universidade Federal Fluminense, torcortes@yahoo.com.br
[2] Universidade Federal Fluminense, jboechat.alergo@gmail.com
[3] Universidade Federal Fluminense, beni@huap.uff.br
[4] Universidade Federal Fluminense, jrodrigo@id.uff.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 rinite é uma doença inflamatória da mucosa do nariz que afeta crianças, adolescentes, adultos e idosos. Em termos clínicos, caracteriza-se pela presença de sintomas como coriza, prurido nasal, espirros e/ou obstrução nasal e etiologicamente pode ser dividida em rinite alérgica ou rinite não alérgica (ORBAN et al., 2009). A rinite frequentemente é sub-diagnosticada ou inadequadamente tratada, e que pode reduzir a qualidade de vida (CAMELO-NUNES & SOLÉ, 2010).

Objetivo

Identificar a associação entre as características clínicas, demográficas e de saúde e a gravidade de rinite dos pacientes do Ambulatório de Alergia e Imunologia Clínica do HUAP-UFF.

Material e Métodos

Trata-se de um estudo transversal, realizado com pacientes com sintomas de rinite acompanhados no Ambulatório de Alergia e Imunologia Clínica do HUAP, em Niterói, Estado do Rio de Janeiro, no período de 2013 a 2015.

O desfecho de estudo refere-se à gravidade da rinite com dois níveis: leve e grave. Com relação às variáveis explicativas, considerou-se, como características clínicas, o tipo de rinite e a presença de sintomas. Além disso, foram consideradas as características de saúde dos pacientes, como a presença de comorbidades, tais como asma e dermatite atópica, e características demográficas, como sexo, grupo etário e cidade de residência.

Usando modelo de regressão logística (KUTNER et al., 2004), foram estimadas as medidas de razão de chance (OR) do paciente ter rinite grave. Na análise ajustada somente foram incluídas as variáveis que na análise bruta apresentaram uma associação significativa com o desfecho, considerando o nível de significância de 20%. Foram mantidas no modelo multivariado apenas as variáveis cuja associação foi significativa a um nível de significância de 5%. As análises estatísticas foram desenvolvidas com o uso do software RStudio, e o modelo logístico foi ajustado usando o comando “glm”.

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

Tabela 1: Associação entre as características clínicas, demográficas e de saúde e a chance do paciente ter rinite grave.

Características	% Pacientes (n=90)	Análise bruta		Análise	
		OR	p-valor*	OR	p-valor*
Sexo					
Masculino	15,6	0,474	0,216		
Feminino	84,4	1	-		
Grupo-etário					
Idoso	46,7	0,369	0,022	0,373	0,037
Não idoso	53,3	1	-	1	-
Cidade					
Niterói	38,9	1,487	0,362		
Fora de Niterói	61,1	1	-		
Coriza					
Sim	70,0	1,185	0,713		
Não	30,0	1	-		
Espirro					
Sim	71,1	0,689	0,427		
Não	28,9	1	-		
Prurido					
Sim	65,6	1,757	0,209		
Não	34,4	1	-		
Obstrução					
Sim	71,1	4,234	0,005	4,478	0,006
Não	28,9	1	-	1	-
Asma					
Sim	16,7	0,808	0,706		
Não	83,3	1	-		
Dermatite					
Sim	7,8	0,698	0,651		
Não	92,2	1	-		
Tipo de rinite					
Alérgica	32,2	2,991	0,022	2,799	0,045
Não alérgica	67,8	1	-	1	-

**Teste de Wald*

Na análise bruta, observou-se que apenas três variáveis estão associadas com a chance de o paciente apresentar rinite grave, considerando o nível de 20%: grupo-etário, obstrução nasal e tipo de rinite. Na análise multivariada (ajustada), essas três variáveis continuaram apresentando significância estatística (p-valor≤ 0,05).

Desse modo, observou-se que pacientes não idosos apresentam uma chance de rinite grave 2,7 vezes maior que a dos idosos (OR=1/0, 373=2,7; p-valor=0,037).

Além disso, verificou-se que a chance dos pacientes com obstrução nasal terem rinite grave é aproximadamente 4,5 vezes maior que a dos pacientes sem obstrução (OR=4,478; p-valor=0,006). A obstrução nasal, o mais proeminente dos sintomas, está associada a eventos respiratórios relacionados aos distúrbios do sono, uma condição que tem profundo efeito sobre a saúde mental, o aprendizado, o comportamento e a atenção (CAMELO-NUNES & SOLÉ, 2010).

Observou-se, ainda, que pacientes com rinite alérgica têm uma chance de gravidade dos sintomas aproximadamente 2,8 vezes maior que a dos pacientes com rinite não alérgica (OR=2,799; p-valor=0,045). Rinite alérgica é uma doença capaz de alterar de forma marcante a qualidade de vida dos pacientes, assim como seu desempenho, aprendizado e produtividade (CAMELO-NUNES & SOLÉ, 2010).

Com relação às medidas de qualidade do ajuste, obteve-se uma taxa global de classificações corretas de 67,8% e medidas de sensibilidade e especificidade iguais a 71,7% e 63,6%, respectivamente; indicando que o modelo tem uma capacidade preditiva razoável.

Conclusão

Através do estudo foi possível concluir que pacientes com rinite do tipo alérgica, com sintoma de obstrução nasal e não idosos tendem apresentar maior gravidade de rinite.

Referências

CAMELO-NUNES, I.C.; SOLÉ, D. Rinite alérgica: indicadores de qualidade de vida. J Bras Pneumol, v. 36, n.1, p. 124-133, 2010.

KUTNER, M.H.; NACHTSHEIM, C.J.; NETER, J. Applied Linear Regression Models. McGraw-Hill, 4th ed., 2004.

ORBAN, N.T.; SALEH, H.; DURHAM, S.R. Allergic and non-allergic rhinitis. In: Adkinson NF Jr, Bochner BS, Busse WW, Holgate ST, Lemanske RF Jr, Simons FER, editors. Middleton's allergy principles & practice, 7th ed. Philadelphia, PA: Mosby Elsevier, 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

MODELAGEM DO ESCORE DE QUALIDADE DE VIDA DE USUÁRIAS DO SISTEMA ÚNICO DE SAÚDE NO MUNICÍPIO DE NITERÓI-RJ

Carlos Augusto Faria[5]
Patrícia Costa de Almeida[6]
Sandra Costa Fonseca[7]
José Rodrigo de Moraes[8]

Resumo

O prolapso genital (PG) é uma disfunção do assoalho pélvico feminino que pode ter impacto negativo sobre a qualidade de vida (QV). O objetivo do estudo foi estabelecer a associação entre as características clínicas e demográficas e o escore de qualidade de vida (QV) de pacientes que atendidas no ambulatório de Uroginecologia do Hospital Universitário Antônio Pedro (HUAP), em Niterói, Estado do Rio de Janeiro. Foi realizado um estudo observacional transversal. No modelo de regressão linear múltipla normal, foram consideradas as seguintes variáveis explicativas: grupo (prolapso, controle), idade (idosas ou não), escolaridade, tipo de parto, índice de massa corporal (IMC), número de gestações e de comorbidades. Para avaliar a significância da associação entre estas características das pacientes e os seus escores de QV, utilizou-se o teste T-Student de significância individual dos parâmetros, considerando o nível de significância de 5%. Observou-se que as mulheres com prolapso têm menores escores de QV quando comparadas com as mulheres sem prolapso, demonstrando que a presença de prolapso genital tem impacto negativo na qualidade de vida das mulheres atendidas no ambulatório do HUAP. As pacientes não idosas, com prolapso genital e com maior quantidade de comorbidades reportam pior qualidade de vida.
Palavras-chave: prolapso genital, qualidade de vida, modelo de regressão linear normal

Abstract

Pelvic organ prolapse is a female pelvic floor disfunction that may have negative impact on the quality of life. The objective of the study was to establish the association between the quality of life scores and clinical and demographic characteristics of patients attended in the Urogynecology ambulatory of the Antônio Pedro University Hospital (HUAP), in Niterói, Rio de Janeiro State. It is an observational, transversal study. In the normal linear regression model were considered the following explanatory variables: group (prolapsed, control), age (elderly, non-elderly) schooling, mode of delivery, body mass index, number of pregnancies and comorbidities. To assess the significance of the association between these characteristics and the scores of quality of life it was used T-student test of parameters individual significance, considering a significance level of 5%. It was observed that women with prolapse have lower scores of quality of life in comparison with women without prolapse, demonstrating that the presence of the pelvic organ prolapse has a negative impact on the quality of life of women attended in the HUAP's ambulatory. The non-elderly patients, with pelvic organ prolapse and with more comorbidities present the worst scores of quality of life.
Key words: pelvic organ prolapse, quality of life, normal linear regression model

[5] Departamento Materno-Infantil, Universidade Federal Fluminense, carlosfaria1965@gmail.com
[6] Universidade Federal Fluminense, patriciaalmeida131@gmail.com
[7] Departamento de Epidemiologia, Universidade Federal Fluminense, sandracfonseca@yahoo.com.br
[8] Departamento de Estatística, Universidade Federal Fluminense, jrodrigo@id.uff.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 prolapso genital (PG) é a disfunção do assoalho pélvico feminino definido clinicamente como o deslocamento caudal das paredes vaginais anterior e/ou posterior e/ou do colo uterino. Leva a sintomas como sensação de abaulamento genital, dor lombar e/ou em baixo ventre e a disfunção sexual (Haylen et al, 2010). Pode trazer impacto negativo sobre a qualidade de vida (QV), ao afetar o contato social, a atividade laborativa, a higiene e a vida sexual das mulheres.

Objetivo

Estabelecer a associação entre as características clínicas, demográficas e de saúde e o escore de qualidade de vida (QV) das pacientes que procuraram o ambulatório de Uroginecologia do Hospital Universitário Antônio Pedro (HUAP), em Niterói, Estado do Rio de Janeiro.

Material e Métodos

Foi realizado um estudo transversal, com pacientes do ambulatório de Uroginecologia do Hospital Universitário Antônio Pedro (HUAP). No modelo de regressão linear múltipla normal, foram consideradas as seguintes variáveis explicativas: grupo (prolapso, controle), faixa etária, escolaridade, tipo de parto, índice de massa corporal (IMC), número de gestações e de comorbidades. Para avaliar a significância da associação entre estas características das pacientes e os seus escores de QV, utilizou-se o teste T-Student de significância individual dos parâmetros, considerando o nível de significância de 5% (Kutner et al., 2004). Quanto a avaliação da normalidade dos resíduos estudentizados do modelo, empregou-se o teste de Shapiro-Wilk, fixando o nível de 5%.

Com relação à estratégia de modelagem, ajustou-se um modelo multivariado incluindo todas as variáveis, mas foram excluídas, uma a uma, considerando a magnitude do p-valor. Este procedimento foi repetido até obter um modelo em que todas as variáveis explicativas tivessem pelo menos uma categoria estatisticamente significante ao nível de 5%. Todas as análises estatísticas foram realizadas com o uso do software R, e o modelo foi ajustado pelo método dos mínimos quadrados utilizando o comando “lm”.

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

Observou-se que a maioria das mulheres atendidas no ambulatório são pessoas idosas (60 anos ou mais), 72,1% têm no máximo o ensino fundamental, 86,8% realizaram parto normal, cesáreo ou ambos os tipos de parto; e que 73,6% apresentam sobrepeso ou obesidade. Além disso, 42,6% das mulheres tem prolapso genital (Tabela 1).

No modelo multivariado contendo todas as variáveis do estudo, observou-se apenas que a "faixa etária" e "número de comorbidades" apresentaram p-valores menores do que o nível de significância de 10%, mas não ao nível de 5%. Ao excluir as variáveis que não apresentaram associação significativa com o escore de QV considerando a magnitude do seu p-valor, obteve-se um modelo multivariado em que as variáveis "grupo", "faixa etária" e "número de comorbidades" estão significativamente associadas com o escore de QV (p-valor≤0,05).

Observou-se que as mulheres com prolapso têm menores escores de QV em comparação as mulheres sem prolapso, demonstrando que a presença de prolapso genital tem impacto negativo na qualidade de vida das mulheres atendidas no ambulatório do HUAP. As pacientes não idosas têm menores escores de QV, comparativamente as idosas, indicando que as pacientes não idosas estão reportando pior qualidade de vida, possivelmente por exercerem atividade laborativa e demandarem vida social e sexual plenas, todas prejudicada pela presença do prolapso. Com relação ao grupo, verificou-se uma associação negativa entre o número de comorbidades e o escore de QV, indicando que o escore de QV reduz em cerca de 3 unidades ao aumentar o número de comorbidades em uma unidade (Tabela 1). Com relação ao diagnóstico do modelo, os resíduos estudentizados não apresentaram nenhum comportamento sistemático com relação aos escores estimados de QV; e estes resíduos apresentaram uma distribuição aproximadamente normal (p-valor = 0,133), como requerido para o ajuste deste tipo de 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

Tabela 1: Associação entre as características clínicas, demográficas e de saúde e o escore de QV das pacientes atendidas no ambulatório de Uroginecologia do HUAP.

Características	% Pacientes (n=68)	Modelo multivariado incluindo todas as variáveis		Modelo multivariado com as variáveis selecionadas	
		Estimativa ($\widehat{\beta}$)	p-valor	Estimativa ($\widehat{\beta}$)	p-valor
Intercepto		59,563	<0,001	65,987	<0,001
Faixa etária					
Não idosa	32,4	-7,359	0,088	-8,678	0,030
Idosa	67,6	0		0	
Escolaridade					
Até o fundamental	72,1	-0,600	0,893		
Médio ou superior	27,9	0			
Tipo de parto					
Parto normal	30,9	5,360	0,438		
Parto cesáreo	22,1	10,053	0,113		
Ambos os partos	33,8	2,094	0,769		
Não teve parto	13,2	0			
IMC					
Magro/Normal	26,4	-3,366	0,499		
Sobrepeso	36,8	3,016	0,489		
Obeso	36,8	0			
Grupo					
Teve prolapso	42,6	-5,966	0,180	-7,666	0,038
Não teve prolapso	57,4	0		0	
Nº de gestações (0 a 10)*	-	-0,070	0,949		
Nº de comorbidades (0 a	-	-2,487	0,088	-3,106	0,020

**Variável numérica.*

Conclusão

As pacientes não idosas, com prolapso genital e com maior quantidade de comorbidades reportam pior qualidade de vida.

Referências

Haylen BT, De Ridder, D. Freeman RM, Swift SE, Berghmans B, Lee J et al. An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for female pelvic floor dysfunction. Int Urogyn. 2010;2(1),5-26.

Kutner MH, Nachtsheim CJ, Neter J. Applied Linear Regression Models. McGraw-Hill, 4th ed., 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

DESAFIOS PARA OS DISCENTES NO INGRESSO AO MERCADO DE TRABALHO

Bianca Sampaio MONTEIRO[9]
Carolina Mazzim Obermuller Carvalho da SILVA [10]
Caroline Ponce de MORAES [11]
Rodrigo Tosta PERES [12]

Resumo

Um assunto sempre presente ao longo da vida acadêmica de um estudante de graduação é a entrada no mercado de trabalho. Trata-se de uma fase permeada por muita expectativa e a preparação provida por parte da instituição de ensino a qual o aluno pertence é de extrema importância para ajudar em sua formação. Este trabalho apresenta uma análise estatística feita a partir de um questionário aplicado a alunos do CEFET/RJ, a fim de mostrar os principais desafios ao iniciar sua carreira profissional. Após a aplicação, foi utilizado o software estatístico RStudio juntamente com sua linguagem de programação R. Análises descritivas e testes de hipóteses foram realizados. As conclusões presentes propiciam uma maior compreensão de como os jovens avaliam este momento e permitem que as instituições de ensino melhorem o processo de formação dos alunos. Essa melhora é fundamental para elevar a confiança para enfrentar os desafios iniciais da vida profissional.
Palavras-Chave: educação estatística, mercado de trabalho, análise descritiva

Abstract

An ever-present subject throughout the academic life of a graduate student is entry into the labor market. This phase is permeated by a lot of expectation and the preparation provided by the educational institution to which the student belongs is of extreme importance to help in its formation. This paper presents a statistical analysis made from a questionnaire applied to CEFET/RJ students, in order to show the main challenges when starting their professional career. After the application, the statistical software RStudio was used along with its programming language R. Descriptive analyzes and hypothesis tests were performed. The present conclusions provide a greater understanding of how students evaluate this moment and allow the educational institutions to improve the process of student training. This improvement is fundamental to raising the confidence to face the initial challenges of working life.
Keywords: statistical education, labor market, descriptive analysis

Introdução

Uma das maiores preocupações dos jovens universitários é a entrada no mercado de trabalho. Trata-se de uma etapa que proporcionará crescimento pessoal, intelectual e profissional dos estudantes, além de contribuir para a sua independência pessoal e financeira.

[9] Estudante de Graduação do Curso de Engenharia de Produção do CEFET/RJ, email: biancasampaiomonteiro@gmail.com
[10] Estudante de Graduação do Curso de Engenharia de Produção do CEFET/RJ, email: carolmocs@gmail.com
[11] Orientador do trabalho. Professor CEFET/RJ, email: poncecefet@gmail.com
[12] Orientador do trabalho. Professor CEFET/RJ, email: rodrigo.peres@cefet-rj.br

Com o avanço da tecnologia e a consequente geração de dados, aumentou-se a necessidade de realizar análises estatísticas para agregar conhecimento a toda informação disponível, como pode-se observar em Duda, Hart, Stork, (2001). Assim, preparar profissionais de engenharia, ainda na graduação para que estejam aptos a realizar análises estatísticas é uma atividade relevante, uma vez que, independente de sua área de atuação, este profissional deve possuir conhecimentos técnicos em estatística para realizar inferências.

Objetivo

Este trabalho foi desenvolvido com o objetivo de mostrar os principais desafios dos alunos de engenharia de produção do CEFET/RJ ao ingressar no mercado de trabalho, podendo assim buscar soluções de melhorias no preparo dos alunos para as fases seguintes da vida.

Material e Métodos

Para a pesquisa, foi aplicado um questionário para um total de 151 estudantes do CEFET/RJ. O questionário conta com 11 perguntas, algumas dicotômicas, outras categóricas. Após a aplicação, foi utilizado o software estatístico RStudio juntamente com sua linguagem de programação R (amplamente conhecida na literatura, como em Dalgaard, (2002)), para construir gráficos e analisar os dados obtidos (Montgomery, Runger, (2012)). As conclusões deste trabalho permitem uma avaliação sobre a forma como os jovens entendem a sua preparação, além de permitir uma reflexão sobre possibilidades de melhoras no processo de formação dos futuros profissionais, para que estes se sintam cada vez mais preparados para os desafios do mercado de trabalho.

Resultados e Discussão

As informações iniciais levantadas pelo questionário proposto dizem respeito à caracterização dos alunos de engenharia de produção do CEFET/RJ que participaram da pesquisa, em relação a período e a gênero (Figuras 1 e 2, respectivamente). Pode-se observar que há um equilibro na quantidade de alunos por período. O período que possui mais entrevistados é o sexto (20,53%) e o que possui menos entrevistados é o terceiro (13,91%). Há um equilíbrio também em relação ao gênero dos alunos.

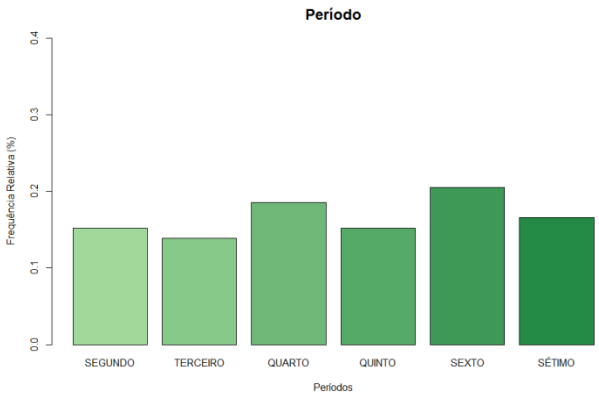


Figura 1 – Distribuição por período

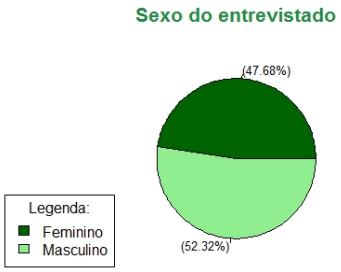


Figura 2 – Distribuição por sexo

Em seguida, foi realizada uma análise cruzada, comparando a sensação de estar preparado por gênero. Pode-se verificar na Figura 3 que, dentre os entrevistados do sexo feminino, 58% se sentem preparados para uma entrevista e dentre os entrevistados do sexo masculino, 63%. Para verificar se a diferença é significativa, executamos um teste de proporção no R, através da função prop.test. A hipótese nula é que há igualdade entre as proporções.

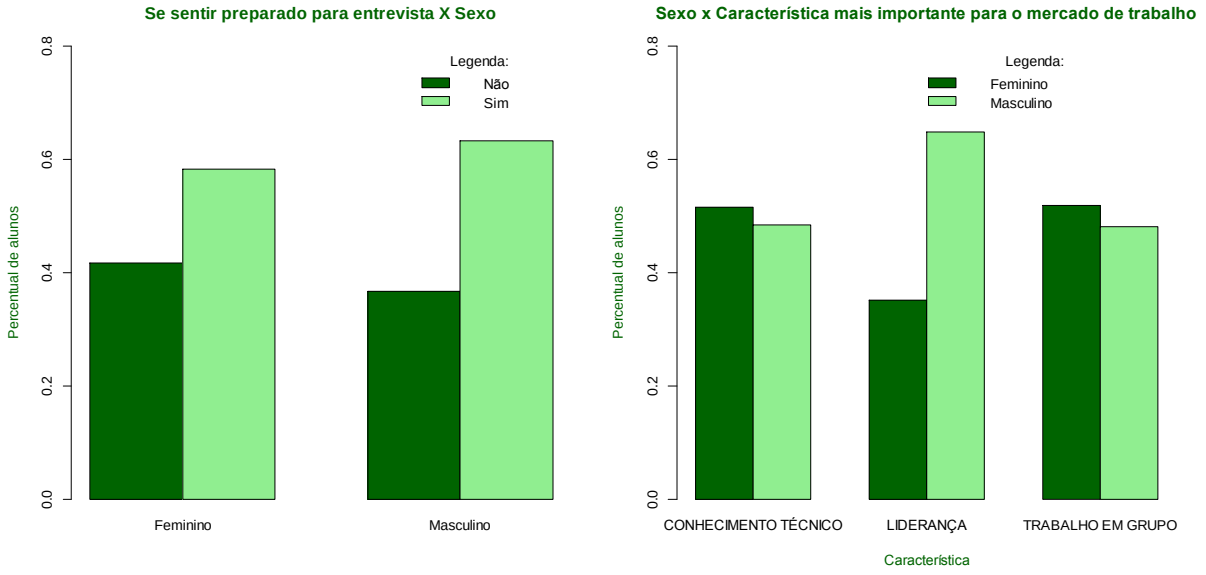


Figura 3 – Distribuição percentual dos alunos por sexo, segundo a sensação de estar preparado ou não para a entrevista.

Conforme a análise descritiva já indicava, concluiu-se que não há evidência para afirmar que estas proporções são diferentes (p valor = 0,65).

Quando analisamos a "característica considerada mais importante para o mercado de trabalho" tivemos que o trabalho em grupo corresponde a mais de 50% dos votos o que nos mostra que realmente houve uma mudança no mercado atual, pois antigamente as empresas e funcionários não se preocupavam com o grupo, cada membro deveria ser bom individualmente. Constatamos também que os homens priorizam mais a característica liderança que as mulheres, embora essa diferença não chegou a ser significante do ponto de vista estatístico (p = 0,12).

Conclusão

Através desse questionário foi possível perceber o tamanho da importância que os estudantes de engenharia de produção dão ao mercado de trabalho, levando-nos assim a uma conclusão de que precisamos investir mais na capacitação dos alunos.

Sendo assim, o CEFET/RJ poderia disponibilizar com uma certa frequência palestras informativas sobre o que esperar e como se portar nestas situações, além de cursos com ferramentas básicas e eletivas relacionadas aos setores empresariais.

Com um investimento e dedicação maior no preparo dos alunos e o comprometimento destes, é possível uma elevação na confiança para enfrentar os desafios do futuro e a vida profissional.

Referências

Duda, R.O., Hart, P.E., Stork, G., Pattern Recognition, 2nd. Ed., Wiley, 2001.

DALGAARD, P. **Introductory Statistics with R**, Statistics and Computing, Springer, 2002.

MONTGOMERY, D.C., RUNGER, G.C., **Estatística Aplicada e Probabilidade para Engenheiros,** LTC, 5ª ed., 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

DIVULGAÇÃO DE NOTÍCIAS NO CEFET/RJ

Christiane Webster CARNEIRO [13]
Mariana Sento Se COSTA [14]
Caroline Ponce de MORAES [15]
Rodrigo Tosta PERES [16]

Resumo

A comunicação no ambiente universitário é de extrema importância no dia a dia dos alunos, docentes e profissionais das instituições. Com o avanço tecnológico, muitas pessoas utilizam apenas a comunicação digital, por acreditar que o alcance é mais rápido e eficiente. Ao emitir um comunicado, é sempre desejado que o maior número de pessoas possível tenha acesso a informação. Este trabalho apresenta uma análise estatística feita a partir de um questionário aplicado a alunos do CEFET/RJ, a fim de mostrar a opinião dos estudantes acerca da comunicação das notícias na instituição. Após a aplicação, foi utilizado o software estatístico RStudio juntamente com sua linguagem de programação R. Análises descritivas e testes de hipóteses foram realizados. A partir desse estudo, espera-se fazer uma reflexão e entender o grau de satisfação dos estudantes em relação a este assunto. As conclusões presentes propiciam um melhor entendimento e possibilitam o aperfeiçoamento dos canais de comunicação existentes.
Palavras-Chave: educação estatística, análise de dados, análise descritiva

Abstract

Communication in the university environment is of extreme importance in the daily life of students, teachers and professionals of the institutions. With the technological advance, many people only use digital communication, believing that the reach is faster and more efficient. When issuing a statement, it is always desired that as many people as possible have access to information. This work presents a statistical analysis made from a questionnaire applied to CEFET/RJ students, in order to show the students' opinion about the news communication in the institution. After the application, the statistical software RStudio was used along with its programming language R. Descriptive analyzes and hypothesis tests were performed. From this study, it is expected to reflect on and understand the degree of students satisfaction in this subject. The present conclusions provides a better understanding and enable the improvement of existing communication channels.
Keywords: statistical education, data analysis, descriptive analysis

Introdução

Um dos assuntos mais citados entre os universitários atualmente é a comunicação, ou quase sempre a falta dela, em seu estabelecimento de ensino seja ele qual for. Muitas vezes, o diretor da instituição lança um comunicado em um determinado veículo e não tem a garantia de que todo o público desejado irá ler o que

[13] Estudante de Graduação do Curso de Engenharia de Produção do CEFET/RJ, email: webster.christiane@gmail.com
[14] Estudante de Graduação do Curso de Engenharia de Produção do CEFET/RJ, email: marisentose@gmail.com
[15] Orientador do trabalho. Professor CEFET/RJ, email: poncecefet@gmail.com
[16] Orientador do trabalho. Professor CEFET/RJ, email: rodrigo.peres@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

foi escrito. Além disso, há alguns casos em que universitários perdem oportunidades como intercâmbio e iniciação científica por não terem tido acesso às informações necessárias no período de tempo pré-estabelecido.

A quantidade de dados disponível atualmente como consequência do avanço computacional, (Duda, Hart, Stork, (2001)) aumentou consideravelmente a demanda por analistas de dados. No presente trabalho, técnicas estatísticas serão usadas a partir do software estatístico RStudio juntamente com sua linguagem de programação R (amplamente conhecida na literatura, como em Dalgaard, (2002)), para construir gráficos e analisar os dados obtidos (Montgomery, Runger, (2012)).

Objetivo

Este trabalho foi desenvolvido com o objetivo de divulgar, através da análise de dados, a opinião dos estudantes do segundo, quarto, quinto e sexto período de engenharia do Campus Maracanã acerca da comunicação das notícias no CEFET/RJ. A partir desse estudo, espera-se fazer uma reflexão e entender o grau de satisfação dos estudantes, de acordo com seu período, em relação a este assunto.

Material e Métodos

Para realizar a pesquisa, foi aplicado um questionário com 10 perguntas para 106 estudantes da graduação, sendo algumas dessas qualitativas e outras quantitativas. Após a aplicação, foi utilizado o software estatístico RStudio para que fosse possível realizar a análise dos dados com rapidez e eficiência. As conclusões deste trabalho permitem uma avaliação sobre a forma de comunicação institucional, além de permitir uma reflexão sobre possibilidades de melhoras neste processo.

Resultados e Discussão

As informações iniciais levantadas pelo questionário proposto dizem respeito à caracterização dos alunos do CEFET/RJ que participaram da pesquisa, em relação a período e a gênero (Figuras 1 e 2, respectivamente). Pode-se observar que há um equilibro na quantidade de alunos por período. Já para a variável gênero, temos que a grande parte, 76,42%, é do gênero masculino.

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

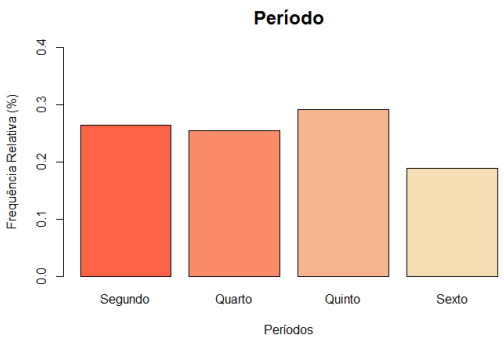


Figura 1 – Análise descritiva do período

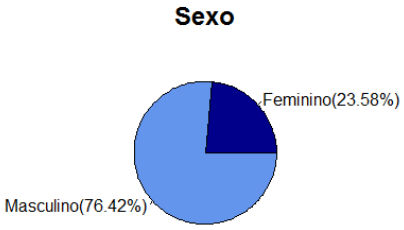


Figura 2 – Análise descritiva por gênero

Em seguida, o objetivo foi avaliar se os entrevistados achavam que havia a necessidade de novos meios oficiais para a divulgação de notícias no CEFET/RJ. Pode-se verificar na Figura 3 que, dentre os entrevistados do sexo feminino, 68% não acham que é necessário criar novos meios, e dentre os entrevistados do sexo masculino, aproximadamente 67% concordam que não há necessidade. Assim podemos concluir que a maior parte dos entrevistados de ambos os sexos não creditam eventuais problemas na comunicação aos canais já existentes. Para verificar se a diferença é significativa, executamos um teste de proporção no R, através da função prop.test. A hipótese nula é que há igualdade entre as proporções. Conforme a análise descritiva já indicava, concluiu-se que não há evidência para afirmar que estas proporções são diferentes (p valor ≈ 1).

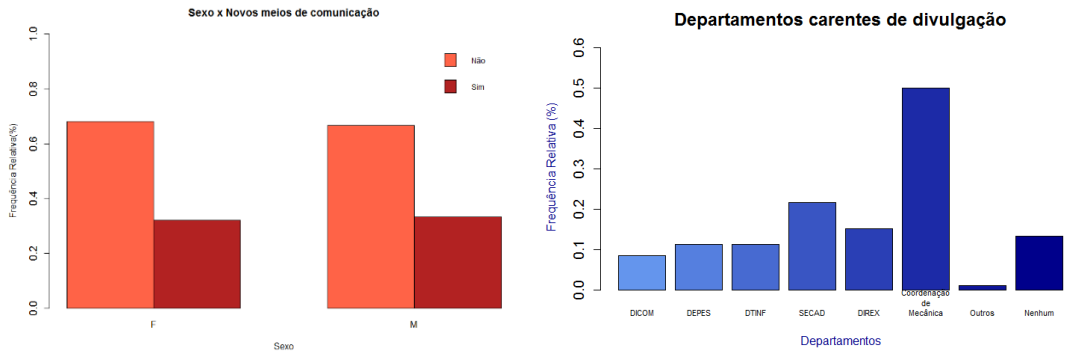


Figura 3 – Necessidade da criação de novos meios de comunicação por gênero e Departamentos carentes de divulgação

O próximo passo foi verificar os departamentos da instituição mais carentes de divulgação. Observe que houve um equilíbrio entre os departamentos, com exceção da coordenação de mecânica. Como boa parte dos alunos que responderam o questionário são alunos de engenharia mecânica, isso é naturalmente explicado.

Outras análises foram feitas. Uma delas foi dividir os alunos entre ciclo básico e profissional e analisar o conhecimento desses dois grupos quanto aos projetos existentes no CEFET. Observamos que não havia diferença significativa entre os que conhecem o projeto de iniciação científica - IC (p valor = 0,55). Mas ao comparar os alunos do segundo período contra os do quarto, o p valor é menor que 0,01. Isso vem de encontro ao que se esperava nesta análise, mostrando que o aluno vai amadurecendo e conhecendo os projetos da instituição a medida em que forem ficando relevantes para o seu momento dentro do curso.

Em relação aos departamentos carentes de divulgação, há diferença de opiniões entre os alunos do ciclo básico e profissional em relação a coordenação de mecânica (p = 0,05).

Conclusão

Cumprindo com o objetivo da pesquisa, há agora uma resposta: as pessoas não estão satisfeitas. Elas também não precisam de novos canais de comunicação. Talvez aperfeiçoar os canais existentes e atualizá-los com maior periodicidade poderia melhorar a avaliação dos alunos.

Referências

Duda, R.O., Hart, P.E., Stork, G., Pattern Recognition, 2nd. Ed., Wiley, 2001.

DALGAARD, P. **Introductory Statistics with R**, Statistics and Computing, Springer, 2002.

MONTGOMERY, D.C., RUNGER, G.C., **Estatística Aplicada e Probabilidade para Engenheiros,** LTC, 5ª ed., 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

UMA PROPOSTA DE ALGORITMO PARA CÁLCULO DA DISTRIBUIÇÃO PREDITIVA DO NÚMERO DE USUÁRIOS NO SISTEMA EM UMA FILA *M/M/*1 EM UMA ABORDAGEM BAYESIANA UTILIZANDO A PRIORI CONJUGADA NATURAL

Nilson Luiz Castelucio Brito[17]
Pedro Henrique Pereira Pinho[18]
Juliana Miranda Corrêa de Guamá[19]

Resumo

Usando-se dados a posteriori para analisar uma fila clássica Markoviana *M/M/1,* sob a perspectiva bayesiana, chega-se ao cálculo proposto por [3]. Este trabalho tem como objetivo implementar um algoritmo em R para realizar o cálculo da distribuição preditiva do número de usuários no sistema em uma fila *M/M/1*. Para o teste, foram usados dados de uma distribuição a priori conforme proposto por [3] e os dados da distribuição a posteriori vêm de uma distribuição geométrica com n=200, combinada com uma distribuição beta de parâmetros 0.6 e 1.7. O referido cálculo envolve a função gama de Euler que, para os valores tratados nos problemas, torna-se impossível o uso de calculadoras científicas ou softwares. Tais valores fazem parte de um trabalho mais amplo que trata da determinação das medidas de desempenho de uma fila *M/M/1* através de uma abordagem Bayesiana. Ao se efetuar algebricamente as referidas contas, deparou-se com a existência de uma relação de recorrência, o que despertou o interesse em tentar obter um algoritmo que fizesse tais cálculos.
Palavras-Chave: Filas Markovianas, inferência Bayesiana, distribuição preditiva.

Abstract

Using a posteriori data to analyze a classic queue Markovian *M/M/1*, from the Bayesian perspective, one arrives at the calculation proposed by [3]. This work aims to implement an algorithm in R to perform the calculation of the predictive distribution of the number of users in the system in an *M/M/1* queue. For the test, we used data from an a priori distribution as proposed by [3] and the posterior distribution data comes from a geometric distribution with n = 200, combined with a beta distribution of parameters 0.6 and 1.7. This calculation involves the Euler gamma function which, for the values treated in the problems, makes it impossible to use scientific calculators or software. These values are part of a larger work that deals with the determination of the performance measures of an *M/M/1* queue through a Bayesian approach. When the accounts were algebraically performed, they encountered the existence of a recurrence relation, which aroused the interest in trying to obtain an algorithm that would make such calculations.
Keywords: Markovian queues, Bayesian inference, predictive distribution.

[17] Unimontes – casteluciobrito@gmail.com
[18] Unimontes – pedro-henrique-pereira-pinho@hotmail.com
[19] Unimontes – juliana.guama@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

Quando se analisa um caso real envolvendo filas, a primeira tarefa consiste em identificar um modelo adequado para descrever suas características. O próximo passo usualmente visa a determinação das medidas de desempenho. Intensidade de tráfego, número de usuários no sistema são alguns exemplos. Tais medidas são expressas em função dos parâmetros do modelo escolhido. A escassez de conhecimento sobre os valores numéricos dos parâmetros torna-se um grande obstáculo. Como obter tais parâmetros? Quais os dados necessários? Como esses dados devem ser coletados? A sugestão é utilizar uma abordagem bayesiana.

Objetivo

O objetivo deste trabalho é propor um algoritmo para calcular a distribuição preditiva do número de usuários no sistema em uma fila M/M/1, em uma abordagem bayesiana utilizando uma priori $\boldsymbol{beta}(\mathbf{0.6}, \mathbf{1.7})$.

Material e Métodos

Assume-se que a fila esteja em equilíbrio. Nestas condições, a intensidade de tráfego $\boldsymbol{\rho} = {}^{\boldsymbol{\lambda}}/_{\boldsymbol{\mu}}$ deve ser menor do que 1, o que significa que a taxa de chegada $\boldsymbol{\lambda}$ é menor do que o tempo de serviço $\boldsymbol{\mu}$. Dessa forma, a distribuição do número de usuários no sistema é dada por:

$$\boldsymbol{P(N = m) = \rho^n(1-\rho), 0 < \rho < 1, m \geq 0} \quad (1)$$

A função de verossimilhança é $\boldsymbol{(1-\rho)^n.\rho^{\sum_i x_i}}$, que, na abordagem bayesiana, é o núcleo de uma distribuição:

$$\boldsymbol{beta}\left(\boldsymbol{n+1}, \sum_{\boldsymbol{i}} \boldsymbol{x_i} + \mathbf{1}\right) \quad (2)$$

Existem duas possibilidades de escolha para a distribuição a priori do parâmetro $\boldsymbol{\rho}$. A conjugada natural $\boldsymbol{beta(a,b)}$ e a priori não-informativa $\boldsymbol{uniforme(0,1)}$. Como em situações práticas, verifica-se que $\boldsymbol{0 < c < \rho < d < 1}$, pode-se trabalhar com uma distribuição $\boldsymbol{uniforme\ truncada}$. No caso da priori conjugada natural, a distribuição a posteriori é $\boldsymbol{\pi_{NC}(\rho|\tilde{x})}$ tem distr (3) o $\boldsymbol{beta(a+\sum_i x_i; n+b)}$ e, no outro caso, $\boldsymbol{beta\ incompleta(c;d;\sum_i x_i + 1; n+1)}$. A partir daí, a distribuição preditiva do número de usuários no sistema na época da partida, quando se utiliza a priori conjugada, é:

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

$$P_{NC}(N = m|dados) = \frac{B(a + \sum_i x_i + m; n + b + 1)}{B(a + \sum_i x_i ; n + b)}$$

Para se obter a distribuição preditiva, faz-se uso de:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} \quad (4)$$

Obtendo-se:

$$\pi_{NC}(\rho|dados) = \begin{cases} \frac{B(a + y + m; n + b + 1)}{B(a + y; n + b)} \rho^{a+y-1}(1 - \rho)^{n+b-1}, 0 \leq \rho \leq 1 \\ 0, \quad caso\ contrário. \end{cases} \quad (5)$$

Com $y = \sum_{i=1}^{n} x_i$.

A função gama de Euler tem a propriedade $\Gamma(n + 1) = n\Gamma(n)$ e, devido aos valores envolvidos, não há como utilizar uma calculadora científica ou um software estatístico para fazer as contas. Ao se calcular algebricamente, verificou-se a existência de uma relação de recorrência e, a partir daí, foi possível implementar um algoritmo em R para fornecer os resultados.

Resultados e Discussão

Fazendo:

$$y = \sum_i x_i,\ \alpha = a + y, \beta = n + b; B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)} \quad (6)$$

Sabendo que $\Gamma(.)$ da propriedade (4) não assume resultado igual a 0, podemos, então reescrever a distribuição preditiva do número de usuários no sistema como:

$$\frac{\frac{\Gamma(\alpha + m)\Gamma(\beta + 1)}{\Gamma(\alpha + \beta + m + 1)}}{\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}} \quad (7)$$

Para obter:

$$\frac{(\alpha + m - 1)(\alpha + m - 2) \ldots (\alpha + m - m)\beta}{(\alpha + \beta + m + 1 - 1)(\alpha + \beta + m + 1 - 2) \ldots (\alpha + \beta + m + 1 - m - 1)} \quad (8)$$

O que fornece o algoritmo implementado em 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

Algoritmo em R:

```
algorithm <- function(alfa,beta,m){
  up <- down <- result <- 1.0
  i <- 1
  fim <- m
  index <- i
  for (i in 1:(fim+1)){
    index <- i
    if(i != fim+1){
      up <- (alfa + m - index)
    }
    down <- (alfa + beta + m + 1.0 - index)
    result <- result * (up/down)
    up <- 1.0
  }
  result <- result * beta
  return (result)
}
```

A tabela 1 fornece os resultados obtidos com $\boldsymbol{n = 200, \sum_i x_i = 38}$ $\boldsymbol{e\ m = 0, \ldots, 5}$.

Tabela 1: Distribuição preditiva a posteriori do número de clientes no sistema

$\boldsymbol{m}$	$\boldsymbol{\pi_{NC}(N = m \mid dados)}$
0	0.839367
1	0.134271
2	0.021944
3	0.003662
4	0.000624
5	0.000108

Conclusão

O algoritmo mostrou-se eficiente no cálculo da distribuição preditiva do número de clientes no caso de uma priori $beta(0.6,1.7)$. Outro algoritmo está sendo testado para o caso de uma priori não informativa $uniformeme\ truncada$.

Referências

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

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

[3] Choudhury, A. & Borthakur, A.C. *Bayesian inference and prediction in the single server Markovian queue.* Metrika 67, 2008, 371-383.

[4] Cruz, F.R.B. & Almeida, M. *Análise de Desempenho em filas M/M/1 usando uma abordagem bayesiana* Proceeding Series of the Brazilian of Applied and Computational Mathematics, vol.3, N.2, 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

ANÁLISE DE CORRESPONDÊNCIA USANDO O PROGRAMA R: UMA APLICAÇÃO EM PESQUISA COM MAGISTRADOS DO TJRN

Rosimere Lopes Monte [20]
Francisco de Assis Medeiros da Silva [21]
André Luís Santos de Pinho [22]

Resumo

Este trabalho busca avaliar a opinião dos Magistrados nas comarcas do Poder Judiciário do Rio Grande do Norte, a respeito da Justiça Estadual, com foco no acesso à Justiça, na confiança da população e na morosidade. Os dados foram obtidos a partir de um instrumento de coleta de uma pesquisa em conjunto com a Universidade Federal do Rio Grande do Norte e o Tribunal de Justiça do Rio Grande do Norte, no período de abril a maio de 2015. Os resultados evidenciam a partir da técnica da análise de correspondência, que a confiança da população na Justiça Estadual do RN, está associado em relação a velocidade da tramitação processual na resolução dos seus conflitos. Ou seja, quanto mais morosa, menor a confiança e vice-versa.
Palavras-Chave: Magistrados. Correspondência. Justiça.

Abstract

This work tries to evaluate the opinion of the Magistrates in the Counties of the Judicial Power of Rio Grande do Norte, a respect of the State Justice, without focus in the Justice, in the confidence of the population and the slowness. The data were obtained from an instrument of collection of a research in conjunction with the Federal University of Rio Grande do Norte and the Court of Justice of Rio Grande do Norte, from April to May 2015. The results show Correspondence analysis technique, which is a trust of the population in the State Court of the RN, is associated with the Speed of the procedural process in the resolution of their conflicts. That is, the more time consuming, the less confidence and vice versa.
Keywords: Magistrates. Correspondence. Justice.

Introdução

Este trabalho apresenta o resultado de uma pesquisa de opinião com Magistrados nas comarcas do Poder Judiciário do Rio Grande do Norte, a respeito da Justiça Estadual. A pesquisa integra um projeto executado pela Universidade Federal do Rio Grande do Norte - UFRN, em conjunto com a Escola da Magistratura do Rio Grande do Norte – ESMARN e o Tribunal de Justiça do Rio Grande do Norte – TJRN. Uma das propostas da pesquisa era buscar a opinião dos Magistrados a respeito da Justiça Estadual, com foco no acesso à justiça, na confiança, na morosidade e na prestação jurisdicional.

[20] Universidade Federal do Rio Grande do Norte, rosimeremonte@hotmail.com
[21] Universidade Federal do Rio Grande do Norte medeiros_assis@hotmail.com
[22] Universidade Federal do Rio Grande do Norte pinho@ccet.ufrn.br

A Análise de Correspondência (CA), através do Programa R, foi a principal ferramenta usada para detectar a associação entre as variáveis de interesse, destacadas no objetivo a seguir.

Objetivo

Detectar a relação entre as seguintes variáveis:

- Busca da Justiça pela população do RN e Nível de confiança que ela deposita na Justiça, sob a ótica do magistrado;
- Busca da Justiça pela população do RN e sua morosidade, medida pelo tempo de tramitação processual, segundo o magistrado;
- Nível de confiança da população na Justiça e sua morosidade medida pelo tempo de tramitação processual, sob a ótica do magistrado.

Material e Métodos

A coleta de dados foi realizada por alunos do ensino a distância da UFRN, nos meses de abril e maio de 2015, nas 64 Comarcas do interior do estado e por alunos do ensino presencial na Comarca da capital. A ambição era um censo com os Magistrados de todas as Comarcas, por serem os principais envolvidos no processo, mas alguns se negaram a participar alegando acúmulo de trabalho, outros estavam de licença/férias no período da pesquisa e algumas Comarcas não contavam com Juiz Titular ou Substituto no momento da pesquisa. Tudo isso associado ao fato de que alguns juízes atuavam em mais de uma Comarca. Mesmo assim, 135 Juízes participaram da pesquisa, equivalente a, aproximadamente, 70% do quadro efetivo.

O instrumento utilizado continha 23 questões, sendo 16 fechadas, 3 abertas e 4 para serem atribuídas notas de zero a dez, além de um espaço para comentários opcionais. O Excel 2013 foi utilizado para a construção do banco de dados e o programa **R** foi utilizado para os cálculos, aplicação dos testes, construção dos gráficos e comparação entre as variáveis pela técnica da Análise de Correspondência.

Aplicou-se algumas técnicas de análise exploratória dos dados, em especial, usando tabelas de dupla entrada, indispensáveis para o teste Qui-Quadrado (χ^2). As categorias dessas variáveis foram representadas no mapa de correspondência.

Na avaliação do processo de busca da população pela Justiça foram utilizadas as alternativas muito fácil, fácil, difícil, e muito difícil, além de não sabe avaliar/não respondeu (NSA/NR). O nível de confiança foi categorizado em muito elevado,

elevado, baixo, muito baixo e não sabe avaliar/não respondeu (NSA/NR).

Quanto ao tempo de tramitação processual, para mensurar a morosidade, a escala utilizada foi: muito lenta, lenta, ágil e muito ágil e, também, não sabe avaliar/não respondeu (NSA/NR).

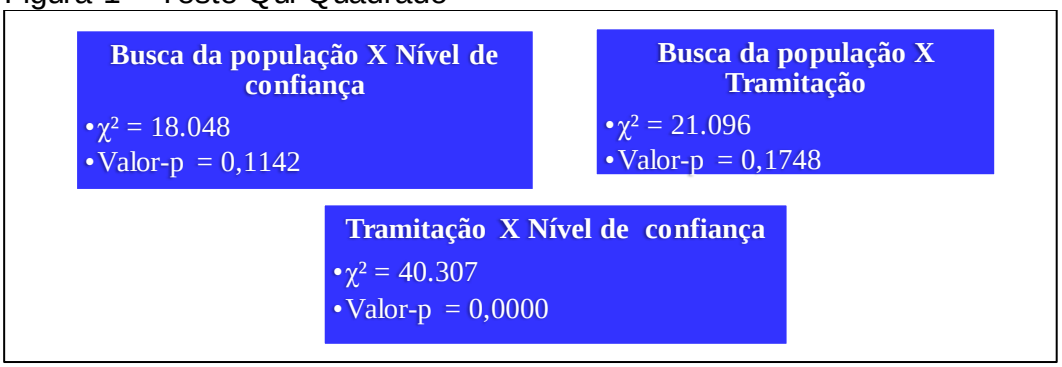
Resultados e Discussão

Após a realização do teste χ^2, a hipótese de que não existe associação entre nível de confiança e tempo de tramitação processual, ao nível de significância de 0,05 foi rejeitada, ensejando a realização da análise de correspondência para identificar as associações (Figura 1).

Assim, por meio da aplicação da função CA, do **R**, verificou-se que os Magistrados que consideram baixo o nível de confiança da população na Justiça, apresentam uma forte tendência a considerar a tramitação processual lenta. Por sua vez, aqueles Magistrados que consideram o nível de confiança das pessoas que procuram a Justiça muito baixo, estão associados a considerar muito lenta a tramitação de processos no Poder Judiciário do RN. Os que disseram que acham elevado o nível de confiança dos jurisdicionados, estão relacionados a considerar como muito ágil ou ágil a tramitação processual (Figura 2).

Por último, na visão dos magistrados a prestação jurisdicional é considerada boa, com médias variando entre 7,0 e 8,5 (Quadro 01).

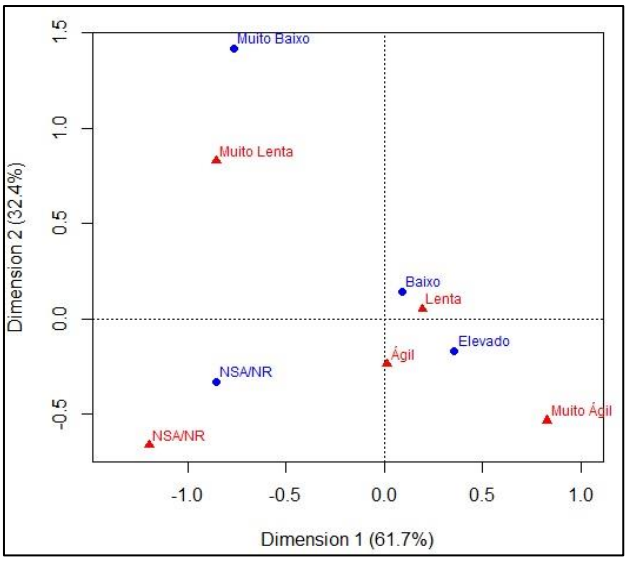
Figura 1 – Teste Qui-Quadrado



Fonte: Pesquisa UFRN/ESMARN/TJRN, abril e maio 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

Figura 2 – Mapa da análise de correspondência



Fonte: Pesquisa UFRN/ESMARN/TJRN, abril e maio de 2015

Quadro 01 - Avaliação da prestação jurisdicional, segundo os magistrados

Itens	Avaliação												
	0	1	2	3	4	5	6	7	8	9	10	NSA / NR	Média
	Ruim/péssima				Regular			Boa		Ótima			
Atendimento dado aos jurisdicionados	-	-	-	1	-	3	7	17	37	32	38	-	**8,5**
Clareza da linguagem utilizada nas decisões	-	-	-	-	-	2	6	14	41	40	31	1	**8,5**
Comprometimento dos servidores	-	-	1	-	1	5	9	20	29	28	38	4	**8,3**
Informações disponibilizadas no site do TJRN	-	-	-	2	2	7	11	31	18	36	24	4	**7,9**
Celeridade da prestação jurisdicional	-	-	2	2	8	17	19	33	23	15	16	-	**7,0**

Fonte: Pesquisa UFRN/ESMARN/TJRN, abril e maio de 2015

Conclusão

Por meio da função CA, do R, verificou-se que os Magistrados que consideram baixo o nível de confiança da população na Justiça, apresentam uma forte tendência a considerar a tramitação processual lenta. Já os Magistrados que consideram o nível de confiança das pessoas que procuram a Justiça muito baixo, estão associados a considerar muito lenta a tramitação de processos no Poder Judiciário do RN. Por sua

vez, os que acham elevado o nível de confiança dos jurisdicionados, estão relacionados a considerar como muito ágil ou ágil a tramitação processual. Portanto, pode-se constatar, sob a ótica dos Magistrados, que a confiança da população na Justiça Estadual do RN, está associada à velocidade da tramitação processual na resolução dos seus conflitos. Por último, na visão dos magistrados a prestação jurisdicional é considerada boa, com médias variando entre 7,0 e 8,5 (Quadro 01).

Referências

1. AGRESTI, A. **Categorical Data Analysis**. *2.* ed. US, 2002.
2. LANDEIRO, V. L. **Introdução ao Uso do Programa R**. Amazônia, 4 mar. 2011.
3. UFRN/ESMARN/TJRN. Projeto de Pesquisa: "**Análise Retrospectiva e Prospectiva da Demanda Judicial e Adequação Organizacional**: **um estudo de caso no Poder Judiciário do Rio Grande do Norte**". UFRN, 2016.
4. 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/.

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 INVESTIMENTO NA EDUCAÇÃO DIMINUI O TRABALHO INFANTIL: MITO OU VERDADE?

Beatriz Gerbase[23]
Vera Correia[24]
Steven Dutt Ross[25]

Resumo

Diante do censo comum de que o investimento em educação teria correlação inversa com a trabalho infantil, o estudo desenvolve a partir dos dados coletados a argumentação necessária para concluir se essa afirmação é verdadeira ou não. Para isso, os dados coletados são provenientes do IBGE e IPEADATA, dentre eles o número de crianças no trabalho infantil e a despesa em educação per capita nos anos 2000 e 2010, onde estudamos a variação em 10 anos, assim a correlação dos dados. A base de dados sobre o trabalho infantil é restrita para a idade de 10 a 13 anos, onde se entende haver maior representatividade para a análise proposta, uma vez que é permitido o trabalho de adolescentes entre 14 e 17 anos, conforme a regulamentação brasileira. A partir da visualização dos dados é possível inferir que apesar do número de crianças que trabalham pouco se alterar nesse período, houve variação por unidades federativas e regiões, entretanto, sem haver influência da despesa em educação, conforme o teste de correlação utilizado nesse estudo.
Palavras-Chave: Trabalho, infantil, despesa, educação, correlação.

Abstract

In face of the common census of that investment in education would have reverse correlation with child labor, the study develops from the collected data the arguments needed to complete if this claim is true or not. For that, the collected data are from IBGE and IPEADATA, among them, the number of children in child labor and education spending per capita in the years 2000 and 2010, where we studied the variation in 10 years, so the correlation of the data. The database on child labor is restricted to the age of 10 to 13 years, where is meant be greater representation to the proposed analysis, since it is allowed the work of adolescents between 14 and 17 years, according to the brazilian legislation. From the visualization of the data it is possible to infer that despite the number of children who work little change during this period, there was variation by states and regions, however, no influence of spending on education, as the test used in this study.
Keywords: Labor, child, investment, education, correlation.

Introdução

Inúmeras são as mazelas sociais que permeiam sociedades modernas, dentre elas, atribuímos ao trabalho infantil a causa de diversas outras que nos cercam diariamente. Comumente, ao tentarmos visualizar soluções para erradicar ou

[23] Aluna da Universidade Federal do Estado do Rio de Janeiro -UNIRIO Email para correspondência: gerbasebeatriz@gmail.com
[24] Aluna da Universidade Federal do Estado do Rio de Janeiro – UNIRIO Email para correspondência: veragcorreia@hotmail.com
[25] Professor de Estatística aplicada às Ciências Sociais da Universidade Federal do Estado do Rio de Janeiro - UNIRIO

minimizar o problema do trabalho infantil, atribuímos o investimento em educação como a melhor forma de detê-lo.

Nesse estudo, desenvolvemos uma análise onde procuramos se há uma correlação entre a taxa de variação do trabalho infantil e a despesa pública em educação, entre os anos 2000 e 2010, tendo em vista que na visão de pessoas leigas, quanto maior o investimento em educação o trabalho infantil nos estados tenderia a diminuir.

Depois de analisarmos os dados coletados pretendemos deixar claro através de números reais e não somente de teoria, o que realmente ocorre ao nosso redor. Para isso, e também para melhor construir nossa análise, diferente dos estudos anteriores sobre o assunto que circulam, não trabalharemos com a faixa de idade entre 10 e 17 anos, tendo em vista que nessa faixa jovens aprendizes também estariam inclusos, o que dificulta uma análise mais real sobre o que realmente acontece em nosso pais no que se diz a respeito de trabalho infantil.

Por isso, utilizamos a faixa etária de 10 a 13 anos, uma faixa que atribuímos um valor maior para esse tipo de análise.

Objetivo

O estudo em epígrafe busca analisar se ao longo de dez anos o aumento da despesa em educação por Unidades Federativas conseguiu contribuir com a diminuição deste problema. O objetivo era verificar se o investimento surtiu efeito, e para isso, somente consideramos a faixa de 10 a 13 anos na análise do trabalho infantil, uma vez que adolescentes entre 14 e 17 anos possuem permissão para trabalhar em condições de jovem aprendiz. Considerando que o IBGE (2015) já possuía um estudo sobre o tema, a comparação se torna inevitável, contrapondo a faixa de idade representada na proposta.

Material e Métodos

Coube analisar dados oriundos do IBGE e IPEADATA, onde foram estudados e relacionados, são eles: número de crianças no trabalho infantil, despesa em educação (per capita), PIB e IDH educação. Ressaltamos que a despesa em educação e o PIB foram deflacionadas tendo como base o IPCA como métrica da inflação. Com o uso do software R (R CORE, 2016), buscou-se analisar de maneira detalhada o comportamento e mudanças por estados e regiões do Brasil.

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 fazer essa análise observou-se a correlação entre a variação do investimento em educação e o trabalho infantil utilizando-se o teste de *Shapiro-Wilk* para verificação da normalidade dos dados referentes a taxa de variação do trabalho infantil e o crescimento da despesa em educação entre 2000 e 2010. Esse teste foi utilizado pois diferentemente do crescimento da despesa em educação, a taxa de crescimento do trabalho infantil não tinha distribuição normal, confirmando-se através do teste, onde o p-valor (4.242e-05) se deu menor do que α estabelecido anteriormente (0,05). Como um dos dados não possuía distribuição normal, a verificação da associação entre a variação do trabalho infantil e a despesa em educação ocorreu através do método Spearman, onde o p-valor (0.4641), maior do que α, não rejeitou a hipótese nula de correlação entre o investimento em educação e a taxa de variação do trabalho infantil no período de 2000 a 2010.

Pesquisamos, então, se havia correlação entre o IDH educação e o crescimento da despesa em educação utilizando o método produto-momento de Pearson, onde observamos o p-valor (0,1991) maior que α e concluímos que também não há correlação entre essas duas variáveis.

Resultados e Discussão

Utilizando uma visualização dos dados de trabalho infantil em tabela, podemos inferir que o trabalho infantil aumentou na maioria das regiões, diminuindo somente no Nordeste. Se tratando do Brasil como um todo temos quase que uma igualdade se compararmos os anos 2000 e 2010, mostrando que não obtivemos mudanças em 10 anos, conforme a seguir:

Tabela 1: Trabalho Infantil no Brasil nos anos 2000 e 2010

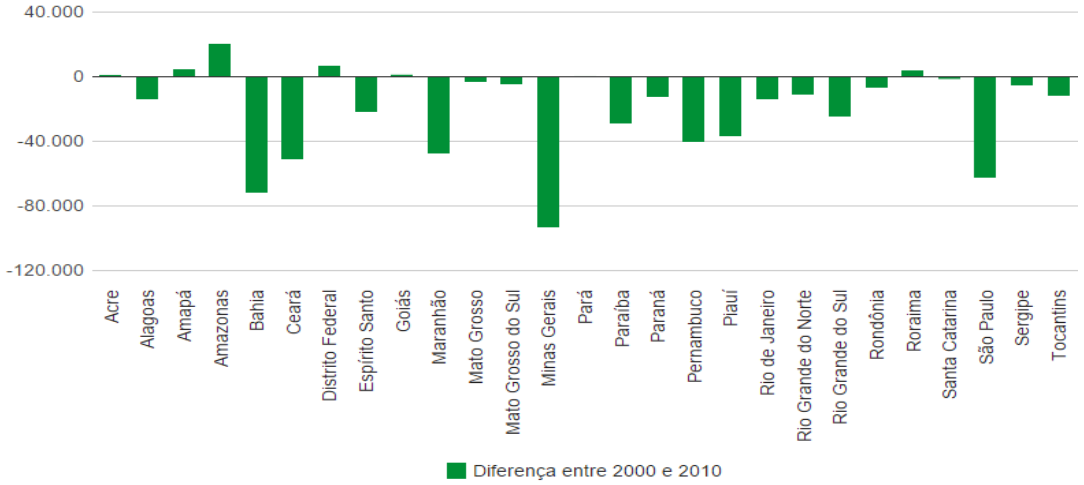
REGIAO	TRABALHO INFANTIL 2000	TRABALHO INFANTIL 2010
NORTE	**88.944**	113.615
NORDESTE	320.100	**273.795**
SUDESTE	**144.924**	156.601
SUL	**105.787**	107.078
CENTRO OESTE	**39.439**	49.743
BRASIL	**699.194**	**700.832**

A partir dessa análise iremos fazer uma comparação mais detalhada dos estados brasileiros e perceber o quanto restringir a idade estudada ajuda a realmente ver onde está instaurada a maior parte do trabalho infantil do Brasil, desmascarando essa análise que inclui uma população maior que desvia o foco da realidade. Para melhor comparação dessa diferença, trouxemos novamente um gráfico do IBGE e logo 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

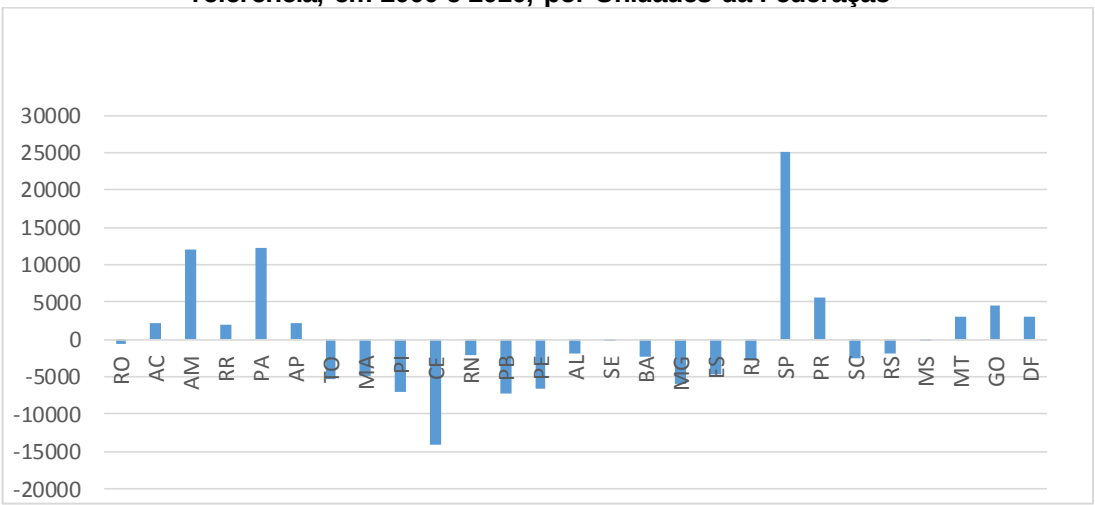
seguida mostramos o gráfico que inclui nosso banco de dados com idade restrita de 10 a 13 anos, que mostra logo de imediato que há uma diferença impactante entre eles.

Gráfico 1: Diferença entre as pessoas de 10 a 17 anos de idade, ocupadas na semana de referência, em 2000 e 2010, por Unidades da Federação



Fonte: IBGE

Gráfico 2: Diferença entre as pessoas de 10 a 13 anos de idade, ocupadas na semana de referência, em 2000 e 2010, por Unidades da Federação



Fonte: Processamento dos autores

A diferença na análise dos dados é bastante evidenciada quando olhamos o estado de São Paulo, onde claramente a discrepância de análise é maior. No Brasil como um todo, adicionalmente, também se nota que com nossa análise é possível visualizar os dados de forma mais compatível com a realidade.

Após essa análise, podemos concluir que o trabalho infantil no Brasil não apresenta drásticas mudanças em seu total, e principalmente na faixa etária do nosso estudo, onde a mudança foi praticamente nula. As características regionais, por outro lado, nem sempre equivalem a de todos seus estados, e que às vezes apenas um

deles é determinante para mudar a característica da região. Também podemos observar que o estado com maior variação foi São Paulo, e que talvez a característica de aumento do trabalho infantil esteja ligada à atividade rural, pelos estados que obtiveram grande aumento, porém não é objeto de estudo desta pesquisa.

Conclusão

Embora em alguns estados específicos a diminuição do trabalho infantil tenha ocorrido, em outros aumentou em escalas desproporcionais. O Brasil se mostrou de maneira semelhante ao longo do tempo, sem apresentar mudança significativa, diferentemente do estudo proposto pelo IBGE. A distribuição do IDH educação pelo Brasil nos mostra que apesar do crescimento do investimento ao longo do tempo (2000 a 2010), a desigualdade entre o norte e o sul do país ainda é acentuada, e principalmente, inegável. Concluímos que o trabalho infantil está atribuído a outros fatores de agravamento, sendo muito mais complexo do que imaginávamos. Focar somente em investimento em educação parece não ser suficiente para retirar as crianças do trabalho infantil. Esses resultados sugerem que em um país de dimensões continentais, as políticas públicas com focos reduzidos garantem uma melhor análise das necessidades e melhor monitoramento de resultados.

Referências

IBGE, 2015. Disponível em: http://www.sidra.ibge.gov.br Acesso em maio/2015.
_____IBGE, 2015. Disponível em: http://censo2010.ibge.gov.br/apps/trabalhoinfantil/outros/graficos.html Acesso em maio/2015.
_____IPEADATA, 2015. Disponível em: http://www.ipeadata.gov.br/. Acesso em maio/2015.
R DEVELOPMENT CORE TEAM, 2016. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/
TOMAZI, N.D., 2013 *Sociologia para o ensino médio*. Editora Saraiva. 3ed. São Paulo.

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

APLICAÇÃO DO PACOTE RQDA NA ANÁLISE DE DADOS PROVENIENTES DE UMA PESQUISA QUALITATIVA

Adriana Oliveira Andrade[26]

Resumo

Um dos desafios da metodologia de análise de conteúdo, utilizada frequentemente nas pesquisas com uma abordagem qualitativa, está relacionado com o tratamento do grande volume de informação textual. Nesse âmbito, o pacote R Qualitative Data Analisys – RQDA (HUANG, 2011) se coloca como um valioso instrumento na medida em que oferece recursos para armazenagem e classificação de segmentos de texto de acordo com categorias criadas pelo pesquisador. O presente trabalho apresenta a implementação da técnica de análise de conteúdo pelo método indutivo com dados de uma pesquisa sobre gestão de pessoas e desse modo ilustra os recursos do pacote RQDA. Através desse pacote foi possível fazer a categorização da fala dos entrevistados e identificar o conteúdo dominante em seu discurso Apesar dessas vantagens, é escassa a produção de trabalhos que utilizem o pacote RQDA. Dessa maneira, pretende-se com esse estudo promover a divulgação desse recurso, desenvolver uma metodologia de trabalho e incentivar o debate sobre o uso do R nas Ciências Humanas.
Palavras-Chave: análise de conteúdo, abordagem qualitativa, categorização

Abstract

One of methodology challenges of content analysis, frequently used in researches with a qualitative approach, is related with the huge quantity of textual information. In this sense, the R Qualitative Data Analysis, RQDA package (HUANG, 2011), is a valuable instrument while offers text mechanisms for storage and classification according to categories created by the researcher. This paper presents the implementation of the technique of content analysis by inductive method using data from a research about people management and illustrates the resources of the RQDA package. Trough this R package was possible making the categorization of interviewer's opinion and identifying the prevalent content in their speech. Although of this advantages, is unusual the production with the RQDA package. This way, it was intended with this study to promoting the RQDA package, developing a methodology of work and encouraging the debate about R in the Human Sciences.
Keywords: content analysis, qualitative approach, categorization.

Introdução

No campo das Ciências Humanas, frequentemente, a produção de conhecimento está vinculada à realização de pesquisas qualitativas. É recorrente o uso de estudos de casos, quando o objetivo é conhecer de maneira aprofundada uma realidade ou ainda compreender os mecanismos envolvidos nos processos sociais (YIN, 2010). Um dos métodos de análise desse tipo de informação é a análise de conteúdo, na qual se pretende explicitar e sistematizar o conteúdo das mensagens

[26] UFRJ/IE/PPED adriana.andrade@pped.ie.ufrj.br

extraídas das entrevistas de modo a identificar sentidos subjacentes ao discurso apresentado, detectar relações e redes não formalizadas e caracterizar o repertório semântico de um grupo (Godoy, 1995b).

Um dos desafios da análise de conteúdo está relacionado com o tratamento do grande volume de informação textual. Nesse sentido, o pacote R *Qualitative Data Analisys* – RQDA (HUANG, 2011) se coloca como uma alternativa viável para a análise das informações provenientes de estudos que trabalham com uma abordagem qualitativa na medida em que oferece recursos para armazenagem e classificação de segmentos de texto de acordo com categorias criadas pelo pesquisador. Ademais, o RQDA é um recurso gratuito e que se apresenta como uma *GUI Aplication* (*Graphical User Interface*) que pode ser manipulado com o teclado e o mouse dispensando, em alguma medida, o conhecimento profundo da programação em R.

Objetivo

Propor uma aplicação do pacote RQDA desenvolvido no R como opção de tratamento e de análise de informações provenientes de pesquisas qualitativas a partir do método de análise de conteúdo.

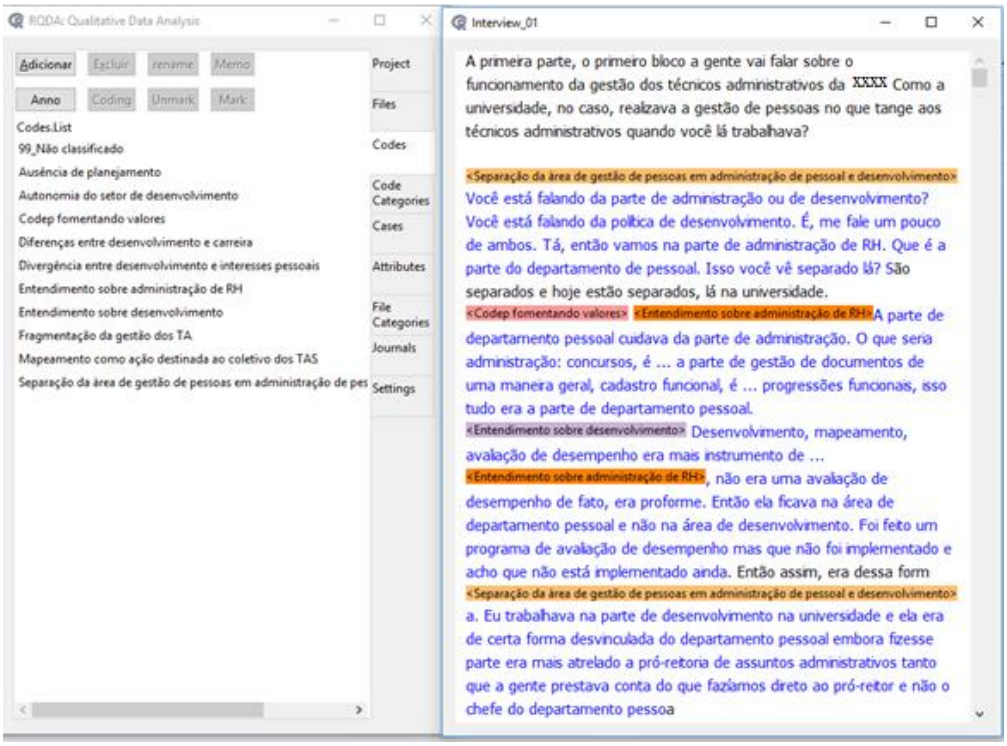
Material e Métodos

O material analisado é proveniente de uma pesquisa realizada com gestores de uma instituição pública federal de ensino superior sobre a política de gestão de pessoas da organização. A pesquisa utiliza uma abordagem qualitativa com a realização de um estudo de caso cujas informações foram produzidas por meio de um roteiro. Os dados foram analisados no pacote RQDA e no R Studio. O presente estudo foi desenvolvido a partir da discussão realizada por Davis Thomas (2006) sobre análise indutiva na qual o autor apresenta um conjunto de etapas para o tratamento das informações pelo método de análise de conteúdo.

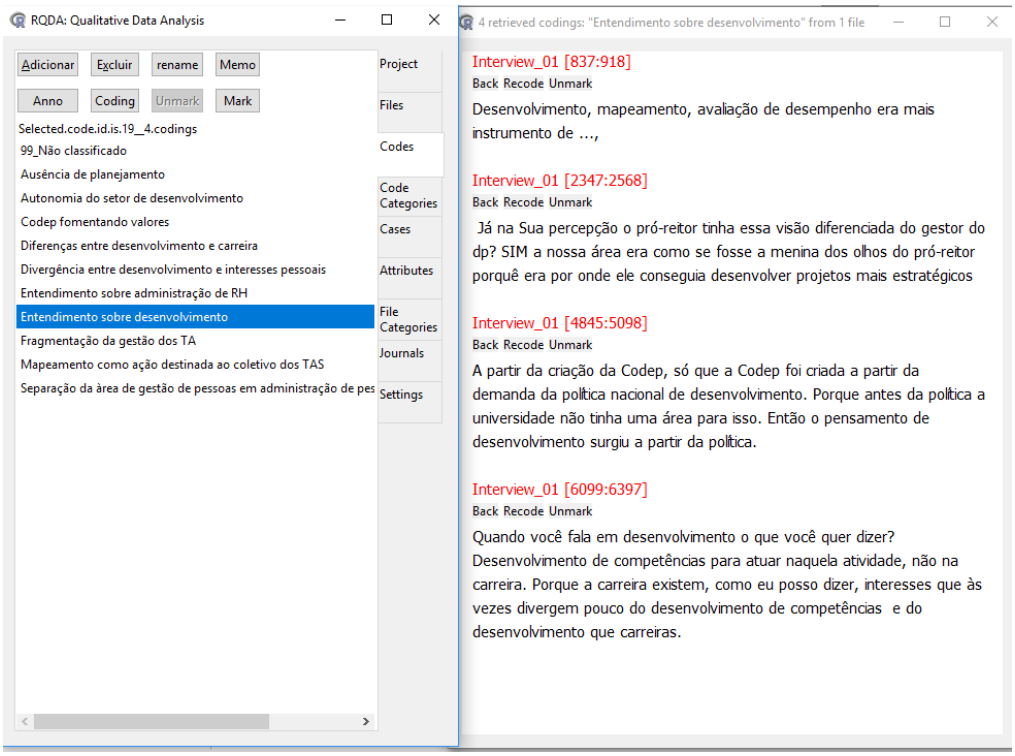
Resultados e Discussão

As etapas propostas por Thomas (2006) foram implementadas no RQDA. A primeira consistiu em importar a informação de cada entrevista realizada como um *file (*arquivo em .txt*)* de um projeto RQDA. Em seguida, para cada *file*, foi realizada a codificação do texto com códigos nomimais (*codes*) criados de acordo com a interpretação realizada pelo pesquisador de cada segmento da fala do sujeito

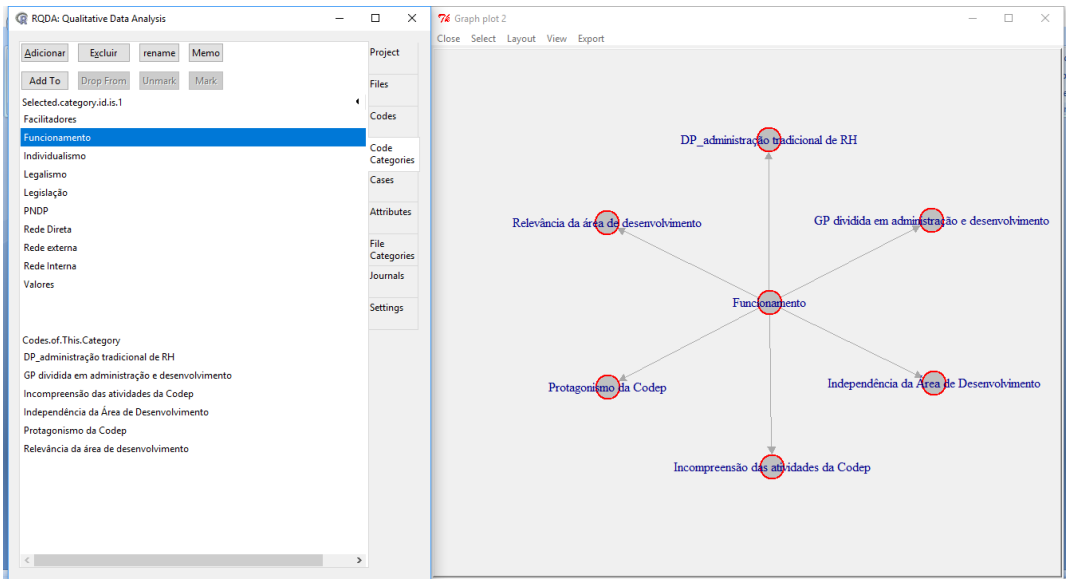
entrevistado, buscando nomear o sentido impresso pelo sujeito da pesquisa. A vantagem desse recurso é que o texto classificado assume uma cor distinta juntamente com o *code* atribuído, destacando sua presença no discurso analisado.



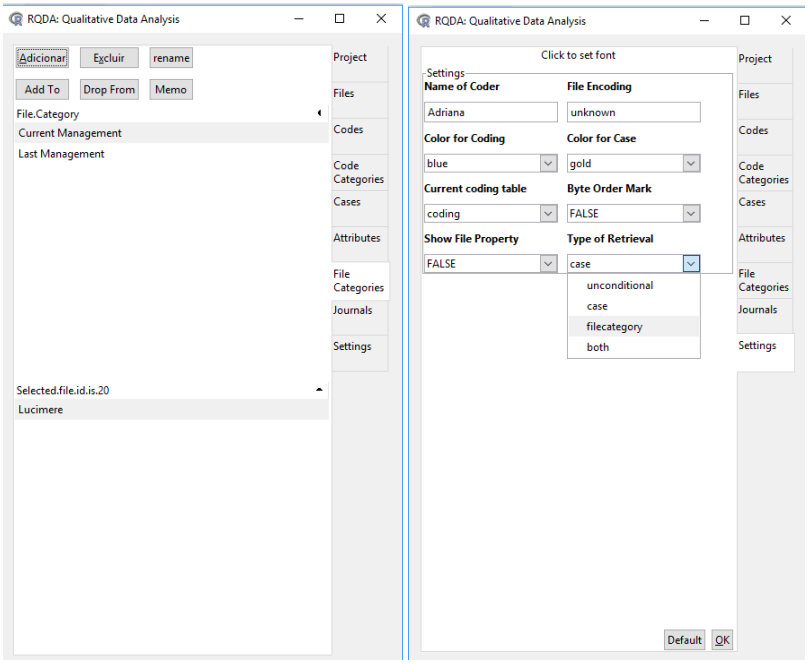
O RQDA possibilita a visualização de todos os fragmentos classificados com o mesmo *code*, o que favorece uma melhor compreensão do discurso do sujeito e a validação das categorias, e ainda facilita a escolha de fragmentos da fala que melhor expressem o significado das categorias criadas.



Na próxima etapa, esses *codes* foram ainda agrupados em categorias de ordem mais geral a partir do recurso de *code.category*. Essa funcionalidade possibilita agregar determinados *codes* que tratam de uma mesma dimensão latente no discurso do sujeito de modo a permitir a construção de categorias mais abstratas que poderão ser relacionadas com o enquadramento teórico adotado. O RQDA possibilita a visualização gráfica do *code.category* com seus respectivos *codes*.



Há a posssibilidade ainda de separar e de visualiar os resultados da codificação por categorias temáticas que agrupam os entrevistados por meio dos seus *files* a partir dos recursos de *File.Categories* e *Type of Retrievel*.



Conclusão

O uso do RQDA possibilitou a análise de dados qualitativos provenientes de um estudo de caso a partir de entrevistas em profundidade. Através desse pacote foi possível fazer a categorização da fala dos entrevistados e identificar o conteúdo dominante em seu discurso. Além da facilidade da sua manipulação do pacote, o RQDA simplificou a realização da análise de conteúdo pela abordagem indutiva. O pacote apresenta ainda o recurso de vincular atributos (variáveis) aos casos analisados o que aumenta as possibilidades de análise. Com relação às limitações do pacote estão alguns *bugs* que tornam mais lenta a análise das informações, mas que de modo algum se constitui como um impedititvo para o seu uso.

Referências

GODOY, Arlida Schmidt. Introdução à pesquisa qualitativa e suas possibilidades. **Revista de administração de empresas**, v. 35, n. 2, p. 57-63, 1995.

HUANG, R. **RQDA: R-based Qualitative Data Analysis**. R package version 0.2-1. URL http://rqda.r-forge.r-project.org/.,

THOMAS, David R. A general **inductive** approach for analyzing qualitative evaluation data. **American journal of evaluation**, v. 27, n. 2, p. 237-246, 2006.

YIN, Robert. K. **Estudo de Caso: planejamento e métodos**. 4. ed. Porto Alegre: Bookman, 2010.

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 DRIVERS OF BREAK-EVEN INFLATION IN BRAZIL: A LASSO APPROACH

Daniel Karp[27*]
Luciano Vereda[28]
Renato Lerípio[29]

Resumo

A break-even inflation – diferença entre a estrutura a termo da taxa de juros nominal e a real – é considerada uma boa medida de expectativa de mercado para inflação. O banco central precisa gerenciar estas expectativas para manter a inflação na meta. Portanto, é importante saber quais são os determinantes da break-even inflation (BEI) para que o banco central entenda precisamente como as expectativas de inflação são formadas pelos agentes. Contudo, esta tarefa não é simples, pois o universo de possíveis variáveis explicativas é grande. Neste artigo nós utilizamos um método de seleção por encolhimento, o least absolute shrinkage selection operator (LASSO), para reduzir o conjunto de possíveis variáveis que explicam a BEI. As estimações foram feitas utilizando o pacote 'glmnet' no software R, um dos poucos programas estatísticos/econométricos a oferecer a possibilidade de estimação via LASSO com a devida flexibilidade. Os resultados são similares aos encontrados na literatura internacional: preços e variáveis financeiras são os determinantes mais importantes da BEI. Particularmente, índices relacionados a custos, como índice de matérias-primas e o salário mínimo, destacam-se como importantes para explicar a BEI.
Palavras-Chave: expectativas de inflação, break-even inflation, LASSO, métodos de seleção por encolhimento.

Abstract

Break-even inflation -- the difference between nominal and real term structure of interest rates -- is a well-recognized measure of the market's expectation about future inflation. The central bank needs to manage the market's expectations in order to keep inflation on the target. Therefore, it is important to know the drivers of break-even inflation (BEI), so the central bank can understand more thoroughly how inflation expectations are formed. However, this is not a simple task, because the universe of possible explanatory variables is big. In this paper we use a shrinkage selection method, the least absolute shrinkage selection operator (LASSO), to reduce the set of possible drivers of the BEI. Estimation was done with 'glmnet' package in R, one of the few statistical/econometric softwares that offers LASSO estimation with flexibility. The results are similar to the ones found in the international literature, with prices and financial variables being the most important drivers of the BEI. Particularly, cost-related variables such as raw materials index and minimum wage stand-out as important explanatory variables.
Keywords: inflation expectations, break-even inflation, LASSO, shrinkage methods.

[27] UFF, lvereda@gmail.com
[28] DIMAC-IPEA, danielkarp@id.uff.br
[29] PUC-Rio, leripiorenato@gmail.com
[*] The authors would like to thank Lucas Maynard for helpful comments.

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

Introduction

Break-even inflation is a well-recognized measure of the market's expectation about future inflation. The central bank needs to manage the market's expectations in order to keep inflation on the target. Therefore, it is important to know the drivers of break-even inflation (BEI), so the central bank can understand more thoroughly how inflation expectations are formed.

It is difficult, however, to build a model to understand the break-even inflation, because there are too many possible explanatory variables. Cicarelli and García (2005), for example, took 27 potential explanatory variables and used bayesian techniques to select the ones which model the European BEI better. In this paper we follow a similar approach, but we use LASSO method to select the variables to model the BEI in Brazil.

Goa

The goal of this paper is to determine, from a great amount of potential explanatory variables, which ones are important to model the break-even inflation in Brazil. This can lead to a better understanding of how inflation expectations are formed in Brazil, an important issue to the central bank.

Data and Methodology

The break-even inflation is the difference between the nominal and real term structure of interest rates (from swap DIxPré and DIxIPCA). It is calculated for 12, 24 and 36 months ahead.

The potential explanatory variables totals 62, they are: IPCA and its lag, its volatility, nine desegregations of IPCA, minimum wage, IPP and its lag, M1 and M3 and its lags, exchange rate and its volatility, Ibovespa's and Dow Jones' returns and its volatilities, spread DI 30x360, oil price and its volatility, raw materials prices, external and internal debt of cities, states, federal government and government owned companies, general expectations index, current transactions balance, exports and imports, eleven desegregations of sales, unemployment, its lag and three desegregations, industrial production and its lag, consumer and managers confidence index. The period goes from 2001 to 2016 in a monthly basis. The series were seasonally adjusted when needed. We also performed ADF tests and differenced the series for which we could not reject the null hypothesis of the unit root presence. Volatilities were calculated via GARCH(1,1)-norm models.

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 model used to reduce the universe of potential variables is the LASSO (least absolute shrinkage and selection operator). This method introduces a penalty to the usual OLS estimator, which forces irrelevant parameters to zero. Following Medeiros, Vasconcellos e Freitas (2016), the LASSO estimator is:

$$\hat{\beta} = argmin_{\hat{\beta}} ||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where β is the vector of parameters, Y is independent variable (in this study the BEI's), X is the matrix of potential explanatory variables and λ is the shrinkage parameter. In order to perform the LASSO estimation we used the 'glmnet' package from R, because standard time-series softwares do not offer this kind of estimation.

The LASSO calculates $\hat{\beta}$'s for different values of λ (in our case from 0 to 100) and for each value it selects a number of variables. In order to choose the best λ we use the standard cross-validation methods (selecting the lambda that minimizes the minimum mean cross-validated error and the lambda that yields the most regularized model) and the method presented in Medeiros, Vasconcellos e Freitas (2016), which uses the Bayesian Information Criterion to select the best λ.

Discussion and Results

Table 1 shows the selected variables for each of the six models for the 12-months ahead BEI.

Table 1. Selected variables for LASSO – 12-months BEI

BIC	MIN	REG
	ibov	
spread DI	spread DI	spread DI
	oil	
raw mat	raw mat	raw mat
	rc balance	
	ind	
	m1	
ipca	ipca	ipca
ipca_res	ipca_res	ipca_res
ipca_sau	ipca_sau	ipca_sau
sm	sm	sm
	ipca vol	
	oil vol	
ipca lag	ipca lag	ipca lag
m1 lag	m1 lag	m1 lag

The results are robust: the same eight variables appeared in all lambda's selection models. They are: IPCA and its lag, the IPCA of residential goods, IPCA of health and personal care, M1's lag, minimum wage, the spread between swap DIxPré for 30 and 360 days and the index of raw materials' prices.

In fact, both BIC and regularized lambda's selection methods chose those same eight variables. Also, the method that minimizes the minimum mean cross-validated error selected the same variables but with some additional ones.

The fact that the DI spread enters the models is interesting, because there is a vast literature that shows this variable is closely related to future inflation. Moreover, the selection of raw materials index and minimum wage are also interesting, because both are proxies for costs, which suggests the BEI has a strong component of inflation of costs.

For 24- and 36-months-ahead (not shown on the table) there are little changes in the selected variables. The main differences are: for 24-months IPCA's volatility enters the models and for 36-months M1 exits the models. Also, for both horizons, other components of inflation are selected in the models. Our results are, in general, in line with Cicarelli and García (2005), as inflation, financial related variables and wages are present in all models. It is important to highlight that sales indices and debt variables do not enter the models.

Conclusion

The break-even inflation seems to be mainly driven by prices (IPCA and desegregations, raw materials prices, and wages) and financial variables (spread DI, Ibovespa returns, oil prices). It is interesting that production, sales and debt related variables are consistently not selected by the models. On the other hand, raw materials prices and the minimum wage are selected in the three methods, suggesting that cost-related variables tend to drive the BEI.

References

Cicarelli, M.; García, J. (2005). "What Drives Euro Area Break-Even Inflation Rates?" *European Central Bank Working Paper Series*, n. 996.

Medeiros, M.; Vasconcellos, G.; Freitas, E. (2016). "Forecasting Brazilian Inflation with High-Dimensional Models". Brazilian Review of Econometrics, 36 (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

PROJEÇÕES DE LONGO PRAZO PARA GASTOS DA PREVIDÊNCIA (RGPS)

Marco Cavalcanti[30]
José Ronaldo Souza[31]
Johann Soares[32]
Daniel Karp[33]

Resumo

A Previdência é uma das partes mais relevantes da Seguridade Social, pois garante, principalmente, os benefícios de idosos ao se aposentarem. Contudo, a Previdência vem sofrendo problemas de sustentabilidade financeira por conta da mudança demográfica em curso no país. O fato de o tamanho do problema não ficar claro para a sociedade como um todo agrava ainda mais o problema. Neste artigo nosso objetivo é mostrar, através de um modelo simples, que a trajetória de gastos da previdência não é sustentável no médio/longo-prazo. Para tanto, um exercício de projeção para os gastos da Previdência (RGPS) é feito. O modelo é "estatístico-contábil" e extrapola as tendências das séries temporais, levando em conta a transição demográfica (e a taxa de mortalidade implícita) projetada até 2050. Como a base de dados é muito grande (o que gera dificuldades em alguns softwares), utilizamos o R (e suas funções básicas) para fazer as projeções. Os resultados indicam uma trajetória explosiva para os gastos – não condizente com a estrutura demográfica do Brasil para os anos vindouros –, o que sugere a necessidade de uma reforma no sistema previdenciário.
Palavras-Chave: projeções de longo-prazo, previdência, estabilidade macroeconômica.

Abstract

The social security is an important part of welfare-state policies, because it ensures the retirement benefits for the elderly. However, in Brazil, the social security financial sustainability is endangered due to the fast demographic change in course. The fact that the problem is not clear to the general public deepens the problem. In this paper we aim at showing, through a simple model, that the spending trajectory is not sustainable in the mid/long-run. In order to achieve this goal we do a forecasting exercise for the social security (RGPS) expenses. The model is 'statistical-accounting" and extrapolates the trend of the time-series, taking into account the demographic transition (and the implied mortality rate) projected until 2050. As the database is too big (leading to some difficulties in standard softwares), we use R (and its basic functions) to do the projectios. The results suggest an explosive trajectory to the RGPS expenses – not in line with the demographic structure in Brazil for the next years --, which suggests the need of a system reform.
Keywords: long-run projections, social security, retirement, macroeconomic stability.

[30] DIMAC-IPEA. Emails: marco.cavalcanti@ipea.gov.br,
[31] ronaldo.souza@ipea.gov.br,
[32] johann.soares@ipea.gov.br,
[33] daniel.vasquez@ipea.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

Introdução

A Previdência é uma das partes mais relevantes da Seguridade Social, pois garante, principalmente, os benefícios de idosos ao se aposentarem. Contudo, a Previdência vem sofrendo problemas de sustentabilidade financeira por conta da mudança demográfica em curso no país. Um agravante do caso brasileiro é o fato de que a pirâmide etária inverterá antes de o país ter alcançado um nível razoável de desenvolvimento. Este cenário indica a necessidade premente de uma reforma da Previdência, a fim de garantir a sustentabilidade do sistema e das contas públicas em geral. Porém, existe o desafio de convencer o grande público desta necessidade.

Objetivo

O objetivo deste estudo é prover, através de um exercício simples*, informações que ajudem a esclarecer o problema da Previdência para o grande público. Para tanto, projetamos os gastos da Previdência (RGPS) até o ano de 2050 e mostramos que os gastos alcançarão níveis insustentáveis, um resultado que o grande público pode entender claramente.

Material e Métodos

Os dados utilizados foram: a) o fluxo e estoque de benefícios (base do Regime Geral de Previdências Social – RGPS); b) projeções demográficas (IPEA e IBGE) e; c) projeções de PIB, inflação e produtividade (IPEA).

Trabalha-se com 448 desagregações de benefícios, conforme tabela a seguir:

Tabela 1: Desagregações Utilizadas no Modelo

Gênero	Clientela	Faixa Etária		Espécies
Feminino	Rural	-19	55-59	Ap. por Idade
Masculino	Urbano	20-24	60-64	Ap. Tempo de Cont.
		25-29	65-69	Ap. Invalidez
		30-34	70-74	Pensões por Morte
		35-39	75-79	Auxílio Doença
		40-44	80-84	Assistenciais
		45-49	85-89	Outros
		50-54	90+	

Fonte: elaboração própria com base nos dados do RGPS.

O grande número de desagregações e dados a serem manipulados sugere o uso de uma plataforma que facilite a realização de simulações e reduza a

* Este trabalho é parte de projeto mais amplo da DIMAC-IPEA, "Modelo de Projeções Macroeconômicas e Previdenciárias de Médio e Longo Prazo". As projeções aqui feitas são preliminares e servirão de insumo e benchmark para um modelo de equilíbrio geral com gerações sobrepostas (OLG), que é o objetivo final do projeto.

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

probabilidade de erro operacional, motivo pelo qual utilizou-se o software R e suas rotinas básicas.

Modelagem: A projeção da despesa total da Previdência requer projeções das quantidades de benefícios e do valor médio de cada benefício, para cada uma das desagregações consideradas. Para a projeção das quantidades, classificam-se os benefícios como temporários ou permanentes. Para os *benefícios temporários,* projetam-se as probabilidades de concessão para cada desagregação com base na extrapolação das "probabilidades de estoque" observadas entre 2004 e 2014[1]: $Prob\ Estoque_{g,c,i,e,t} = \frac{Estoque_{g,c,i,e,t}}{População_{g,c,i,t}}$. Dadas estas probabilidades e as projeções demográficas, obtêm-se os estoques de benefícios para cada desagregação até 2050:

$$Estoque_{g,c,i,e,t}^{2019-2049,\ T} = (Prob\ Estoque_{g,c,i,e,t}^{2014,\ T})(População_{g,c,i,t}^{2019-2049})$$

No caso dos *benefícios permanentes,* as probabilidades de concessão de novos benefícios são calculadas a partir da razão entre benefícios concedidos e estoques. Uma vez concedidos, a probabilidade de sobrevivência dos benefícios é dada pela probabilidade de sobrevivência implícita da população brasileira, para cada desagregação ($Prob\ Sobr_{g,c,i,e,t} = \frac{População_{g,c,i+1,e,t+1}}{População_{g,c,i,t}}$). Dadas estas probabilidades e as projeções demográficas, obtêm-se os estoques de benefícios:

$$Estoque_{g,c,i+1,e,t+1}^{P} = (Prob\ Sobr_{g,c,i,e,t})(Estoque_{g,c,i,e,t}^{P}) + Concedidos\ liq._{g,c,i+1,e,t+1}$$

O valor médio dos benefícios é projetado em função de cenários para o IPCA, PIB e produtividade da economia. A partir das projeções das quantidades e valores dos benefícios, calcula-se a despesa total da Previdência:

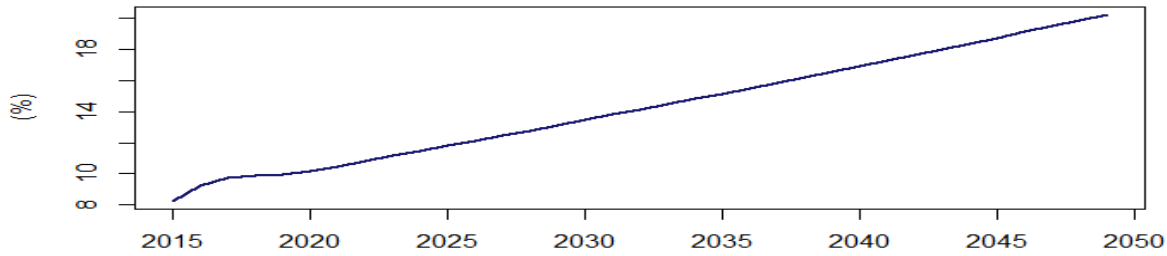
$$Despesa\ Total_{g,c,e,t} = (Valor\ Médio\ Anual_{g,c,e,t})(Estoque\ Médio\ Anual_{g,c,e,t})$$

Resultados e Discussão

As projeções agregadas mostram um aumento de gastos em proporção do PIB de 8,2% para 20,24%, como pode ser visto no gráfico abaixo.

[1] Os subscritos referem-se a: g – Gênero, c – Clientela, i – Idade, e – Espécie de Benefício, t – tempo.

Figura 1: Despesa Previdenciária em relação ao PIB

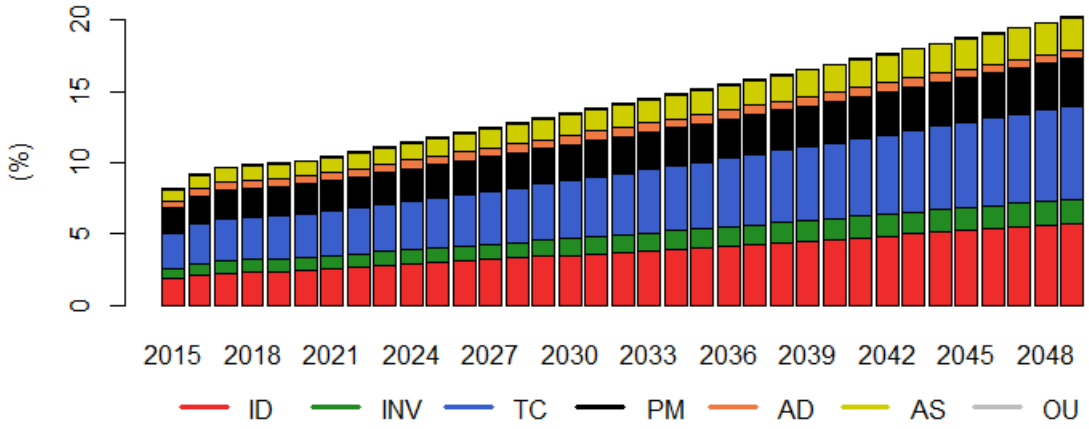


Fonte: elaboração própria com base nos resultados do modelo.

Estimativas para a receita de contribuições previdenciárias ainda estão em construção, mas as projeções demográficas mostram que a quantidade de pessoas na População em Idade Ativa cairá em torno de 3% até 2050, resultando em queda na receita potencial.[2] Fica claro que, na ausência de reformas profundas, o sistema entrará em trajetória insustentável.

Uma vantagem do estudo é o nível de desagregação dos dados. No gráfico abaixo é possível ver a evolução de cada tipo de benefício até 2050.

Figura 2: Despesa Previdenciária por Tipo de Benefício em Relação ao PIB[3]



Fonte: elaboração própria com base nos resultados do modelo.

As aposentadorias por tempo de contribuição, por exemplo, são as que mais aumentarão. Neste sentido, as propostas da PEC 287, como a instituição de idade

[2] A não ser que a produtividade média da economia aumente fortemente, um cenário pouco razoável dado nosso histórico recente e perspectivas.
[3] Onde ID é Aposentadoria por Idade; INV é Aposentadoria por Invalidez; TC é Ap. por Tempo de Contribuição; PM é Pensão por Morte; AD é Auxílio Doença; AS é Benefício Assistencial; e OU é Outros.

mínima para todos, tem o potencial de reduzir estes gastos, aumentar o tempo de geração de receita previdenciária e contribuir para a sustentabilidade do sistema.

Pode-se argumentar que o foco deste exercício na simplicidade e clareza de resultados gera uma perda de precisão das projeções. Contudo, nossos resultados são robustos se comparados a modelos mais complexos.[4]

Conclusão

O modelo aqui proposto é simples, mas traz resultados importantes. O estudo mostra que a trajetória de gastos da Previdência é explosiva e insustentável. Em 2050, 20% do PIB terá que ser gasto apenas com o Regime Geral de Previdência Social. Isto é, sem contar com o Regime Próprio de Previdência Privada, o que agravaria mais ainda a situação. Espera-se que este modelo simples e seus resultados sejam mais uma fonte de informações para convencer o grande público da necessidade de reforma da Previdência.

Referências

MEIRELLES, H. “Carta anexada à Proposta de Emenda Constitucional 287”. Brasília, 2016.

[4] Comparamos nossos resultados com modelos construídos por especialistas na área de Previdência e os resultados são 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

THE TROUBLE WITH THE LINEAR TAYLOR RULE FOR BRAZIL: DEVELOPING AN ALGORITHM TO ESTIMATE A THRESHOLD AUTOREGRESSIVE MODEL WITH EXOGENOUS VARIABLES.

Johann Soares[34]
Matheus Rabelo[35]

Resumo

Nesse paper, estimamos uma regra de Taylor (1993) linear para o Brasil com dados mensais de 2002:01 até 2016:11 e encontramos evidências de não linearidade. Desta maneira, procedemos de acordo com Salgado, Garcia e Medeiros (2005) estimando um modelo auto-regressivo com limiar admitindo variáveis exógenas (TARX). Como não existe um pacote no R para estimar o modelo desejado, o desafio foi desenvolver um algoritmo para: (i) estimar consistentemente o valor do limiar usando o método de Chan (1993); (ii) utilizar a informação obtida para estimar o modelo TARX. Os resultados trazem revelações em termos da função de reação do Banco Central do Brasil (BCB) no ambiente do regime de metas para a inflação. O valor estimado do limiar (8.5% a.a.) sugere um nível não desprezível de conivência com a inflação por parte do BCB no período considerado. O benefício relativo de usar o R para realizar esse exercício, é a capacidade de trabalhar facilmente durante todas as etapas necessárias em um mesmo ambiente, desde o download dos dados e suas transformações, até a obtenção dos resultados da estimação.
Palavras-Chave: Regra de Taylor, política monetária, threshold, mudança de regime.

Abstract

In this paper, we estimate the so-called linear Taylor rule (1993) for Brazil with monthly data from 2002:01 to 2016:11 and we found evidences of non-linearity. Thus, we proceed as well as Salgado, Garcia and Medeiros (2005) estimating a Threshold Autoregressive Model With Exogenous Variables (TARX). As there is no R package available to estimate the desired model, the challenge was develop an algorithm to: (i) estimate consistently the threshold value using the Chan's method (1993); (ii) use the obtained information to estimate a TARX model. The results bring some insights in terms of the Central Bank of Brazil's (CBB) reaction function in the environment of the Inflation Targeting Regime for monetary policy. The threshold value estimated (8.5%YoY) suggest a not despicable level of connivance with inflation by the CBB in the considered period. The relative benefit in using R to perform this exercise, is the capacity to work easily during all of the necessary steps in the same environment, since data download and its necessary transformations until the obtaining the results of the estimation.
Keywords: Taylor rule, monetary policy, threshold, regime switching.

[34] Faculdade de Economia-UFF, johann.soares@hotmail.com
[35] Faculdade de Economia-UFF, matheusrabelo@id.uff.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

Introduction

Widely diffused in monetary policy literature, the so-called Taylor rule (1993) characterizes the central banks reaction function. Since June 1999, the Central Bank of Brazil (CBB) adopted the inflation targeting regime, i.e., CBB manages the monetary policy having an inflation target as a focus. Besides that, are set lower and higher bounds around the central target to accommodate shocks.

It's reasonable suppose that CBB reacts to economic conditions in a distinct way — distinct Taylor rules — depending on the inflation level. Therefore seems a better alternative to model a *nonlinear* reaction function for CBB instead a linear. Thus, in this paper we estimate a Threshold Autoregressive Model with Exogenous Variables (TARX) as well as Salgado, Garcia and Medeiros (2005).

Goal

Investigate if a *nonlinear* Taylor rule fit the data better than a linear one.

Material and methods

The linear version of the Taylor Rule adopted is:

$$i_t = \alpha + \rho_i i_{t-1} + \rho_\pi \pi_t + \rho_y y_t + \varepsilon_t$$

Where i_t is the Selic rate accumulated in the month; π_t is the 12-month accumulated (obtained by '*PerformanceAnalytics*' package) of the consumer price index measured by IPCA; y_t is the cycle component of the Hodrick-Prescott decomposition (obtained by '*mFilter*' package) of the industrial production seasonally adjusted (obtained by '*seasonal*' package). We used monthly data from 2002:01 to 2016:11 that was extracted directly from CBB website using the '*rbcb*' package.

After the estimation of the above model, we found evidences of non-linearity from RESET and McLeod-Li tests. Therefore, we followed Salgado, Garcia and Medeiros (2005) estimating the TARX model below:

$$i_t = \begin{cases} \alpha_1 + \rho_{i,1} i_{t-1} + \rho_{\pi,1} \pi_t + \rho_{y,1} y_t + \varepsilon_{t,1}, & if\ x_{t-d} \geq \tau \\ \alpha_2 + \rho_{i,2} i_{t-1} + \rho_{\pi,2} \pi_t + \rho_{y,2} y_t + \varepsilon_{t,2}, & if\ x_{t-d} < \tau \end{cases}$$

Where τ is the threshold value; x is the threshold variable; and d is the delay parameter. We followed Enders (2015) estimating the above model by OLS assuming the assumption that $Var(\varepsilon_{t,1}) = Var(\varepsilon_{t,2})$. It's important to say that there is no R package to estimate a TARX model. Thus, we developed a function that works as follows: (i) we used the Chan's (1993) method to estimate consistently the threshold

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

value[1]; (ii) after the knowledge of the threshold value and its delay, we estimate the TAR model allowing for exogenous variables.

Results and Discussion

The linear version of the Taylor rule for the sample that we are using brings the following results[2]:

$$\hat{\imath}_t = 0.05^{*} + \ 0.89^{***}\, i_{t-1} + \ 0.01^{**}\, \pi_t - 2.20^{***}\, y_t$$

All the coefficients are statically significant at 10% and signs are in line with the applied work for Brazil. However, despite those initial desired characteristics, the RESET and McLeod-Li tests, performed by using "*lmtest*" and "*TSA*" packages respectively, suggest *non-linearity*. In this way, we proceed estimating a TAR model in which a threshold variable is the IPCA delayed in 5 months[3], which seems reasonable in light of the inflation targeting regime for monetary policy in Brazil. The results of the estimated TARX are the following:

(i) The estimated threshold value was 8.5% year over year (YoY). Considering the actual higher bound of tolerance of 6.5% YoY set by the CBB. Besides that, the results above suggest that the CBB in the period that was considered reacts more strongly when the IPCA lies above the threshold value, which means a not despicable level of connivance with inflation. The results of the estimation are the following:

$$\hat{\imath}_t = \begin{cases} 0.19^{**} + 0.72^{***}\, i_{t-1} + 0.01\, \pi_t - 2.50^{*}\, y_t, & if\ \pi_{t-5} \geq 8.5 \\ 0.00 + 0.91^{***}\, i_{t-1} + 0.01^{***}\, \pi_t - 2.07^{***}\, y_t, & if\ \pi_{t-5} < 8.5 \end{cases}$$

(ii) The coefficient associate to lagged interest rate take into account interest rate smoothing by CBB; above results suggests that monetary policy in Brazil is more smooth in the lower regime, i.e., where the IPCA is below the threshold value of 8.5%YoY; which indicate that in the highest regime, the CBB could overreact.

(iii) Not coincidently, the coefficient associated to output gap has a bigger value in the lower regime. This means that in lower inflation scenario, the CBB

[1] We use only the middle 70% of the values of i_t.
[2] Where * denotes significant at 10% of significance level; ** denotes significant at 5% of significance level and *** denotes significant at 1% of significance level.
[3] The delay parameter was those that minimize the residuals sum of squares.

could manage the monetary policy giving greater weight to economic activity.

(iv) Finally, comparing the Akaike Information Criteria and Bayesian Information Criteria, the *nonlinear* model is able to fit better the data in the sample considered.

Conclusion

In this paper, we found evidences against the usage of the linear Taylor rule in the considered period (from 2002:01 to 2016:11). In this way, we proceed the modelling of the reaction function of the CBB through a TARX specification following Salgado, Garcia and Medeiros (2005). In fact, as there is no R package to estimate a TARX model, the challenge was develop an algorithm that estimate a threshold value consistently and after that, using the obtained information, estimate the required model.

We use the IPCA as the threshold value its estimate (8.5%YoY) suggest a not despicable level of connivance with inflation by the CBB. Besides that, the monetary policy is softer and CBB could manage the policy instrument to achieve goals in terms of economic activity in the lower regime.

Finally, the greater advantage to use R for proceed the desired exercise was the possibility of work easily in a same environment, since data download and its necessary transformations until the obtaining the results of the estimation.

References

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. The annals of statistics, 520-533.

Enders, W. (2004). Applied Econometric Time Series, by Walter. Technometrics, 46(2), 264.

Salgado, M. J. S., Garcia, M. G., & Medeiros, M. C. (2005). Monetary policy during Brazil's Real Plan: estimating the Central Bank's reaction function. *Revista Brasileira de Economia*, *59*(1), 61-79.

Taylor, J. B. (1993, December). Discretion versus policy rules in practice. In Carnegie-Rochester conference series on public policy (Vol. 39, pp. 195-214). North-Holland.

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

MINERAÇÃO DE TEXTOS: UM ESTUDO DE CASO COM DADOS DO *TWITTER*

Carla Cristina Passos Cruz[36]
Jessica Quintanilha Kubrusly[37]

Resumo

O trabalho realizou um estudo de caso em Mineração de Textos com textos do *Twitter*. Foram extraídos 750 textos (documentos) de 15 contas, vinculadas a 5 temas: “esportes”, “política”, “moda”, “viagem e turismo” e “ciência e tecnologia”. De cada tema foram escolhidas 3 contas e de cada conta foram retirados, ao acaso, 50 documentos. Após o pré-processamento restaram 648 documentos que juntos continham 697 diferentes termos (palavras). Esse material foi estruturado em uma matriz termo-documento e então submetido à uma Análise de Conglomerados, com o objetivo de agrupar documentos que tratam de assuntos semelhantes. Como resultado, foram obtidos 15 conglomerados, sendo a maioria deles formado por documentos oriundos predominantemente de em um único tema. Porém 4 conglomerados apresentaram documentos oriundos dos 5 diferentes temas. Além disso, 1 desses 4 conglomerados concentrou mais de 60% dos 697 documentos, o que mostra uma não identificação de um assunto comum entre os documentos desse conglomerado. Toda a análise foi realizada utilizando o *software R*.
Palavras-Chave: Mineração de Textos, Análise de Conglomerados, *Twitter.*

Abstract

In this work, we present a case study which applies text mining to Twitter data. For this, 750 texts (documents) of 15 accounts were extracted, linked to 5 themes: "sports", "politics", "fashion", "travel and tourism" and "science and technology". For each theme, 3 accounts were chosen and 50 documents were taken at random from each account. After the preprocessing, there were 648 documents that together contained 697 terms (words). This material was structured in a term-document matrix and submitted to a Conglomerate Analysis, with the objective of grouping documents that deal with similar subjects. As a result, we have 15 groups, most of them consisting of documents originating predominantly from a single theme. However, 4 groups presented documents from the 5 different themes. In addition, 1 of these 4 groups concentrated more than 60% of the 697 documents, which shows a non-identification of a common subject among the documents of that conglomerate. The analysis was performed using software R.
Keywords: Text Mining, Conglomerate Analysis, Twitter.

Introdução

Atualmente, observa-se o aumento do registro do fluxo de dados gerados pela interação homem e máquina através do avanço da tecnologia. Tarefas comuns do dia-a-dia, como postagens em redes sociais, cadastro em algum *site* para a realização de uma compra, são exemplos em que os dados estão envolvidos. Hoje sabe-se que a

[36] Departamento de Estatística - UFF, e-mail: carlapassos2889@gmail.com
[37] Departamento de Estatística - UFF, e-mail: jessica@est.uff.br

grande maioria dos dados gerados são provenientes de dados não estruturados, ou seja, estão em forma de textos, o que dificulta a extração e exploração das informações. Logo surge a necessidade de se aplicar técnicas que possam ajudar no processamento e análise desses dados. A área de estudos dedicada a isso é chamada de Mineração de Textos.

Objetivo

O objetivo deste trabalho consiste em aplicar técnicas de Mineração de Textos em dados obtidos através da rede social *Twitter* e, a partir do método de Análise de Conglomerados, agrupar os documentos (textos) do *Twitter* por assunto.

Material e Métodos

O material bruto usado para este trabalho é composto por um conjunto de textos do *Twitter.* Foram escolhidas 15 contas, sendo 3 contas de cada um dos 5 temas pré-definidos: “esportes”, “política”, “moda”, “viagem e turismo”, “ciência e tecnologia”. Para cada conta foram extraídos 1.000 textos (documentos) e realizado um sorteio aleatório de 50 deles. Dessa forma foram coletados 750 documentos. A data dos *tweets* variou de 01/06/2016 à 04/10/2016. A extração dos textos foi feita utilizando os pacotes *ROAuth* (Gentry & Lang, ROAuth: R Interface For OAuth, 2015), *twitteR* (Gentry, twitteR: R Based Twitter Client, 2015) e *RCurl* (Lang, 2016) do *software R*.

O processo de Mineração de Textos possui 4 etapas: coleta; pré-processamento; extração da informação; e análise de resultados. Entre elas, o pré-processamento é a etapa mais importante, dividida em: *tokenização*; eliminação das *stop words;* conversão das letras maiúsculas em minúsculas; remoção de figuras e *emoticons*; remoção de acentos; remoção de caracteres especiais; remoção de números; remoção de *links*; remoção de espaços extras; normalização, na qual são utilizados processos de lematização e do “dicionário de termos” (*thesaurus)*; criação da matriz termo-documento; e seleção de termos. Veja mais em Johnson et al. (Johnson & Wichern, 2007).

Ao final do pré-processamento entende-se que cada documento é um conjunto de termos (palavras). Então é criada a matriz termo-documento, que guarda a informação sobre os termos em cada documento. A posição *(i,j)* dessa matriz é igual a 1 se o termo *j* aparece no documento *i* e 0 caso contrário. Logo, cada documento é representado por um vetor de 0’s e 1’s, vetores linhas da matriz termo-documento.

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

Uma vez finalizada a matriz termo-documento é realizado o agrupamento dos documentos, a partir do Método Hierárquico Aglomerativo da Análise de Conglomerados (Johnson & Wichern, 2007). Entre textos foi adotada a similaridade cosseno (Choi, Cha, & Tappert, 2010), que é adequada para dados binários e definida pelo cosseno do ângulo entre os vetores que representam cada documento. Para a distância entre conglomerados foi adotado o método de Ward, que busca minimizar o desvio padrão dentro de cada grupo (Johnson & Wichern, 2007).

Todas as análises foram feitas no *software R* (R Core Team, 2017).

Resultados e Discussão

O estudo foi iniciado com 750 textos (documentos) do *Twitter*. Para as etapas de pré-processamento foram usados os pacotes *tm* (Feinerer & Hornik, 2015), *NLP* (Hornik, 2016) e *bitops* (Maechler & Dutky, 2013). Para algumas etapas não contempladas por tais pacotes, como a remoção de acentos, imagens, *emoticons* e caracteres especiais, foram criadas rotinas e funções no *R*. Em alguns casos ainda foi necessária a inspeção manual, como a união de termos compostos, palavras que pertencem a um mesmo radical e palavras sinônimas.

Ao final do pré-processamento observou-se que entre os 2.197 termos, 1.498 deles aparecem em um único documento. Por entender que os termos que aparecem em um único documento não agregam informação sobre a similaridade entre documentos, optou-se por usar apenas os termos que aparecem em dois ou mais documentos: 697 termos. Além disso, entre os 750 documentos, 102 deles ficaram com um ou nenhum termo após o pré-processamento. Por entender que textos muito pequenos não apresentam bons resultados na Análise de Conglomerados, optou-se por usar apenas documentos com 2 ou mais termos: 648 documentos.

A Análise de Conglomerados foi realizada com o pacote *tm* (Feinerer & Hornik, 2015). Para o cálculo das distâncias entre os elementos amostrais e conglomerados utilizou-se os pacotes *proxy* (Meyer & Buchta, 2017) e *stats* (R Core Team, 2017), respectivamente. Quanto à escolha dos grupos, foi aplicada a inspeção visual a partir do dendrograma, que resultou em 15 grupos.

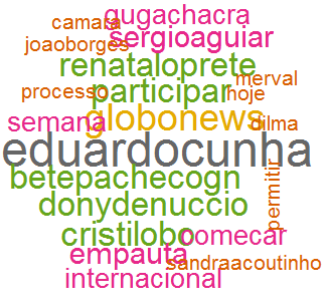
Fez-se então um levantamento das contas e termos dos documentos de cada grupo, a fim de avaliar se o agrupamento foi capaz de caracterizar um assunto. Também foi feita, com o uso do pacote *wordcloud* (Fellows, 2014), a nuvem de palavras por grupo, que contribuiu para essa avaliação. A nuvem de palavras mostra

as 100 palavras mais frequentes em cada grupo, e o tamanho das palavras é proporcional a sua frequência no grupo. Algumas nuvens apresentadas a seguir possuem menos de 100 palavras, isso ocorre quando o grupo tem menos de 100 palavras.

Observou-se que os grupos 4, 5 e 6 tiveram uma ótima caracterização, pois todos possuem textos pertence a um único tema. Como exemplo veja na Figura 1 a nuvem de palavras do grupo 4, onde todas as palavras são relacionadas ao tema "moda". Já os grupos 2, 8, 9 e 12 apresentam a maioria dos textos em um único tema, o que indica uma boa caracterização. Veja na Figura 2 a nuvem de palavras do grupo 12, com todas as palavras relacionadas ao tema "política", mais especificamente, ao jornalismo da Globo News, provavelmente oriundas de publicações do jornalista Gerson Camarotti, que teve sua conta entre as 15 escolhidas.

Figura 1: Nuvem Grupo 4.

Figura 2: Nuvem Grupo 12.



Os grupos 1, 10, 11 e 13 apresentaram a maioria dos textos divididos em 2 temas e por isso dizemos que a sua caracterização é bipartida. Apesar disso, por exemplo para o grupo 11, cuja nuvem de palavras está apresentada na Figura 3, podemos perceber que, embora dividido entre os temas "moda" e "viagem e turismo", o grupo é caracterizado por um assunto comum.

Figura 3: Nuvem Grupo 11.

Figura 4: Nuvem Grupo 15.

Já os outros 4 grupos não tiveram uma identificação de tema, seus documentos se dividiram em proporções semelhantes entre os 5 temas e a caracterização desses grupos foi considerada ruim. Vale o destaque para o grupo 15, cuja nuvem de palavras está apresentada na Figura 4, que é o grupo com maior quantidade de textos, 406, e apresenta palavras referentes a assuntos bem distintos.

Conclusão

Constatou-se que 7 dos 15 conglomerados criados pela Análise de Conglomerados apresentaram documentos que em sua maioria pertencem à um único tema, o que mostra um bom resultado do método de agrupamento, que foi capaz de identificar documentos com assuntos em comum. Por outro lado, 4 conglomerados não tiveram um tema em destaque, sendo que um deles ainda concentrou uma quantidade muito grande de documentos: 406 documentos, o que representam mais de 60% de todos os 648 documentos.

Acredita-se que a existência de conglomerados com caracterização ruim seja consequência da baixa quantidade de termos por documento, uma vez que os documentos foram retirados do *twitter* e por isso contém no máximo 140 caracteres. Uma alternativa para melhorar os resultados seria extrair documentos maiores, com mais palavras, mas esses teriam que ser retirados de outras redes sociais, *blogs* ou *sites*.

Referências

Choi, S., Cha, S., & Tappert, C. (2010). A survey of binary similarity and distance measures. *J. syst. cybern. informatics, 8*, pp. 43--48.

Feinerer, I., & Hornik, K. (2015). *tm: Text Mining Package.* Fonte: https://CRAN.R-project.org/package=tm

Fellows, I. (2014). *wordcloud: Word Clouds.* Fonte: https://CRAN.R-project.org/package=wordcloud

Gentry, J. (2015). *twitteR: R Based Twitter Client.* Fonte: https://CRAN.R-project.org/package=twitteR

Gentry, J., & Lang, D. (2015). *ROAuth: R Interface For OAuth.* Fonte: https://CRAN.R-project.org/package=ROAuth

Hornik, K. (2016). *NLP: Natural Language Processing Infrastructure.* Fonte: https://CRAN.R-project.org/package=NLP

Johnson, R., & Wichern, D. (2007). *Applied multivariate statistical analysis.*

Lang, D. (2016). *RCurl: General Network (HTTP/FTP/...) Client Interface for R.* Fonte: https://CRAN.R-project.org/package=RCurl

Maechler, M., & Dutky, S. (2013). *bitops: Bitwise Operations.* Fonte: https://CRAN.R-project.org/package=bitops

Meyer, D., & Buchta, C. (2017). *proxy: Distance and Similarity Measures.* Fonte: https://CRAN.R-project.org/package=proxy

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Fonte: https://www.R-project.org/

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

INTEGRAÇÃO POR APROXIMAÇÃO: SIMULAÇÃO VIA MONTE CARLO

Janaina Fabiana de Lima Dantas[38]
Jorge Alves de Sousa[39]

Resumo

Para utilizarmos o método de integração por aproximação se faz necessário utilizar simulação. Esse termo simulação refere-se ao tratamento controlado pelo pesquisador em um ambiente, especificamente o computador, em meio a existência de tantos métodos numéricos, o Método de Monte Carlos (MMC) tem tido um destaque especial, tendo em vista que se aplica tanto a problemas com teor não probabilístico quanto àqueles com estrutura probabilística. Neste contexto o objetivo geral deste trabalho foi utilizar um método alternativo de simulação para obtermos aproximações numéricas de integrais complexas por meio do software estatístico R. Acrescentando-se o fato de que a capacidade computacional de processadores aumentou muito nestes últimos anos. Portanto, na aplicação prática deste trabalho, consta-se a eficiência do método, onde as estimativas simuladas foram muito próximas dos valores calculados de forma exata.
Palavras-Chave: Estrutura probabilística, integrais complexas, métodos numéricos

Abstract

To use the integration method by approximation, it is necessary to use simulation. The term simulation refers to controlled treatment by the researcher in an environment, specifically the computer. Among so many numerical methods, Monte Carlo Method has been a special highlight, since it applies both to probabilistic and non-probabilistic problems. In this context, this work's goal is to use an alternative method to simulation to obtain complex numerical integration approximation using as a tool the statistical software R. Another important fact is that the processors' computation capacity increased a lot the last years. Therefore, in this paper's practical application, the method efficiency was calculated, showing that the simulated estimates were very close to the values exactly calculated.
Keywords: Probabilistic structure, complex integrals, numerical methods

Introdução

A integração numérica ainda tem sido um processo lento e oneroso em análises paramétricas, diferentemente dos outros métodos aproximados, que são extremamente rápidos (Paquetti,2008). Para utilizarmos o método de integração por aproximação se faz necessário utilizar simulação. Esse termo simulação refere-se ao tratamento controlado pelo pesquisador em um ambiente, especificamente o computador, onde os problemas reais são analisados através de reproduções. Na utilização de simulação podemos observar que em alguns casos ela descreverá um sistema no qual todos os componentes são conhecidos e o seu comportamento é dito determinístico. Em outros, os componentes são aleatórios, não havendo regra

[38] Universidade Federal de Campina Grande, fabianalimapl@hotmail.com
[39] Universidade Federal de Campina Grande, jorgeas@ufcg.edu.br

matemática que o descreva e sim o uso de estruturas probabilísticas, sendo estocástico o processo de simulação, ou seja, baseado em distribuições de probabilidade. Segundo Press 2007, no estudo de vários problemas matemáticos, recorrer ao Método de Monte Carlo (MMC) tem sido notadamente interessante, pois através de experimentos de amostragem estatísticas realizadas em um computador apresentam soluções aproximadas. Diante de tantos métodos numéricos, o MMC tem tido um destaque especial, tendo em vista que se aplica tanto a problemas com teor não probabilístico quanto àqueles com estrutura probabilística.

Objetivo

Diante do exposto, o objetivo geral deste trabalho é utilizar um método alternativo de simulação para obtermos aproximações numéricas de integrais complexas por meio do software estatístico R.

Material e Métodos

Inicialmente se implementou o algoritmo no software R pelo MMC, onde, a integral utilizada para validação foi a Função Densidade de Probabilidade do Modelo exponencial (Bussab e Morettin, 2003).

```
#metodo monte carlo
f<-function(x) exp(-x)
n=100 # n de 100 a 10000000
theta<-mean(f(runif(n,1,3)))*(3-1)
print(theta)
integrate(f,1,3) #integracao exata
#grafico da funcao
curve(f(x),1,3, xlab="x", ylab="f(x)")
```

Para se verificar a eficácia do método, utilizou-se o boxplot e executando a função 50 vezes para as simulações, de acordo com o seguinte algoritmo:

```
#metodo monte carlo
int.exp = function(n, a, b) {
x = runif(n, a, b)
y = exp(-x)
int.exp = (b - a) * mean(y)
return(int.exp)
}
n = c(20, 50, 100, 200, 500)
```

```
y = matrix(0, ncol = length(n), nrow = 50)
for (j in 1:length(n)) {
m = NULL
for (i in 1:50) m = c(m, int.exp(n[j], 1, 3))
y[, j] = m
}
boxplot(data.frame(y), names = n)
```

Resultados e Discussão

Na Tabela 1. Podemos observar que existe uma variação considerável na estimativa da integral (erro de Monte Carlo), que decresce conforme aumento do número de simulações.

Na figura 1, se observa o boxplot executado em 50 estimativas para as diversas situações simuladas, o mesmo permitiu avaliar a simetria dos dados, a dispersão e a existência ou não de outliers, sendo adequado para compararmos os conjuntos de simulações e a quantidade de erros correspondentes para cada categoria.

Tabela 1. Resultado numérico da integral aproximada para o modelo exponencial usando o Método Monte Carlo

n	**Resultado**
100	0.3565906
1000	0.3196586
10000	0.3169927
100000	0.3186645
1000000	0.3179756
10000000	0.3181011
100000000	0.3180888

No algoritmo, a saída no terminal para o resultado exato 0.3180888 com ε(erro) $< 3.5e^{-15}$.

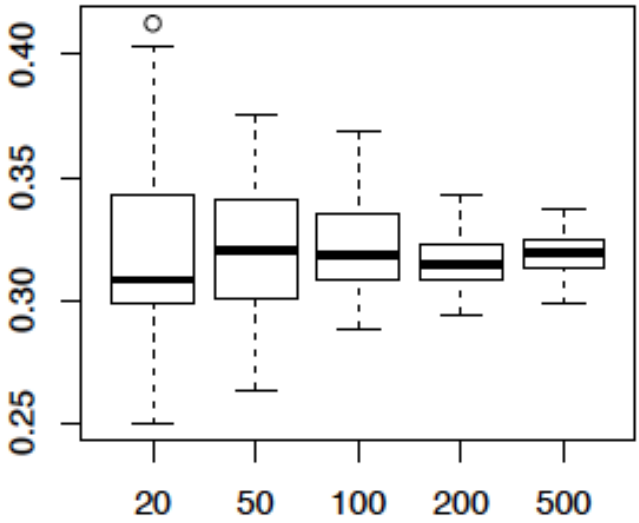


Figura 1. Boxplot para 50 estimativas da integral para distintos valores de n = 20; 50; 100; 200 e 500.

Conclusão

É reconhecido que os MMCs atualmente, são uma das ferramentas mais precisas para a obtenção desses resultados. Acrescentando-se o fato de que a capacidade computacional de processadores aumentou muito nestes últimos anos. Portanto, na aplicação prática deste trabalho, consta-se a eficiência do método, onde as estimativas simuladas foram muito próximas dos valores calculados de forma exata.

Referências

BUSSAB, W.O.; MORETTIN, P.A. Estatística Básica. 8ª edição, Editora Saraiva. 20013.

PASQUETTI, E. Métodos Aproximados de Solução de Sistemas Dinâmicos Não-Lineares. Tese (Doutorado) — PUC-Rio, 2008.

PRESS, W. H. Numerical Recipes 3rd edition: The art of scientific computing. Cambridge University press, 2007.

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

IMPACTO DA REDUÇÃO DA QUANTIDADE DE ALTERNATIVAS DE UM ITEM DO ENEM NA ESTIMAÇÃO DA PROFICIÊNCIA DO PARTICIPANTE

Alexandre Jaloto[40]
Natália Caixeta Barroso[41]

Resumo

Este trabalho tem como objetivo verificar o impacto da redução do número de alternativas nos itens do Enem na estimação das habilidades dos(as) participantes. Foi realizada uma simulação que consistiu na aplicação de três instrumentos diferentes a uma mesma população de 1000 estudantes produzida aleatoriamente. Um dos instrumentos (A) era composto por 45 itens de cinco alternativas; o segundo (B), por 45 itens de quatro alternativas; e o terceiro (C), por 45 itens de três alternativas. Os resultados mostram que as diferenças entre as proficiências estimadas pelo instrumento A e o instrumento B foram pequenas, se comparadas às diferenças entre as estimadas pelo instrumento A e o instrumento C. Portanto, este trabalho aponta para a possibilidade do uso de um instrumento de 45 itens de quatro alternativas, em vez de cinco, para estimar a proficiência de estudantes.
Palavras-Chave: Teoria de Resposta ao Item, Pseudo-chute, ENEM.

Abstract

This work aims to verify the impact of the reduction of an item alternatives number in National Secondary Education Examination (Enem) on the estimation of participant abilities. One performed a simulation that consisted of an application of three different instruments to the same randomly produced population of 1000 students. One of the instruments (A) was consisted by 45 items from five alternatives; the second (B), by 45 items from four alternatives; and the third (C), by 45 items from three alternatives. The results show that the differences between proficiencies estimated by instrument A and instrument B were small when compared to the differences between those estimated by instrument A and instrument C. Therefore, this work points to the possibility of using an instrument of 45 Items of four alternatives, rather than five, to estimate student proficiency.
Keywords: Item Response Theory, Pseudo-guessing, ENEM.

Introdução

O Exame Nacional do Ensino Médio (Enem) surgiu como um modelo de avaliação que tem como referência principal a articulação entre a educação básica e a cidadania (BRASIL, 2009). O Enem teve sua primeira versão aplicada em 1998, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), autarquia do então Ministério da Educação e do Desporto. Desde a sua criação até 2008, o participante tinha sua nota calculada por meio de um instrumento de 63 itens objetivos de cinco alternativas com o uso da Teoria Clássica dos Testes (TCT). A partir

[40] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP. alexandre.jaloto@inep.gov.br
[41] INEP. natalia.caixeta@inep.gov.br

de 2009 esse cálculo passou a se dar por meio da Teoria de Resposta ao Item (TRI), por meio de quatro instrumentos de 45 itens objetivos de cinco alternativas.

O INEP adota atualmente uma metodologia de elaboração de itens que indica a necessidade da existência de distratores plausíveis. Ou seja, em um item, as alternativas incorretas devem retratar possíveis hipóteses de raciocínio utilizadas por aqueles que não dominam a habilidade testada para buscar a solução da situação-problema apresentada (BRASIL, 2010). No entanto, por vezes há uma dificuldade em construir quatro distratores plausíveis. Devido a essa dificuldade, mostra-se necessário refletir sobre a possibilidade de redução da quantidade de distratores em cada item.

Objetivo

Este trabalho tem como objetivo verificar o impacto da redução do número de alternativas nos itens do Enem na estimação das habilidades dos(as) participantes.

Material e Métodos

Para verificar o impacto da redução do número de alternativas, foi realizada uma simulação que consistiu na aplicação de três instrumentos diferentes a uma mesma população de 1000 estudantes. Um dos instrumentos (A) era composto por 45 itens de cinco alternativas; o segundo (B), por 45 itens de quatro alternativas; e o terceiro (C), por 45 itens de três alternativas. As proficiências da população foram geradas aleatoriamente, utilizando a semente 1.000, por meio da função *rnorm*, indicando média 0 e desvio padrão 1. Doravante, tais proficiências serão denominadas "proficiências originais". Para os três instrumentos, os parâmetros *a* e *b* dos itens foram fixados; a única diferença entre eles era o valor do parâmetro *c* de seus itens: os do A eram 0,20, os do B eram 0,25 e os do C eram 0,33. O parâmetro *a* de todos os itens valia 1,2. Os parâmetros *b* variavam de -1,381 a 3,271, amplitude equivalente à observada para a área de Ciências da Natureza e sua Tecnologias no mapa de itens do Enem publicado pelo Inep, quando transformada para a escala (500,100). Na simulação, cada participante respondeu a cada um dos testes apenas uma vez.

Para os cálculos e as simulações foram utilizados os pacotes *catIrt, mirt, Metrics* e *mirtCAT* do programa *R*. As respostas dos participantes foram simuladas a partir da função *simIrt* do pacote *catIrt*; e as proficiências foram estimadas com uso da função *fscores* do pacote *mirt*.

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

As análises consideram as proficiências originais e as proficiências estimadas por meio dos três instrumentos (PA, PB e PC). Foram verificados o erro quadrático médio (EQM) e a correlação linear de Pearson entre as proficiências originais e as proficiências estimadas. Tanto as proficiências originais quanto as estimadas pelos três instrumentos foram padronizadas.

Resultados e Discussão

Os resultados indicam que há uma diferença entre os EQM das medidas das proficiências dos três testes e a proficiência original; sendo que o erro gerado pelo instrumento A (0,2577) foi o menor e o do C (0,3697), o maior. A diferença entre os erros gerados pelos três testes foi menor entre o teste A e B (0,0154) do que entre o teste A e C (0,112). A maior correlação observada foi a do teste A (0,8710), seguida bem de perto pela do B (0,8633). A Tabela 1 sintetiza os resultados obtidos com as simulações.

Tabela 1. Correlações e erros quadráticos médios entre as proficiências estimadas e as originais.

	PA	PB	PC	Dif. |PA-PB|	Dif. |PA-PC|
Correlação	0,8710	0,8633	0,8149	0,0077	0,0561
Erros quadráticos médios	0,2577	0,2731	0,3697	0,0154	0,1120

Legenda: dif.=diferença

O Gráfico 2 mostra as curvas de informação dos três instrumentos. É possível notar que a curva do instrumento A é a que possui o maior pico de informação, enquanto a do instrumento C é a que possui o menor. O Gráfico 2 apresenta as curvas de densidade das estimadas com os três instrumentos. Quando se compara a curva de PA com as outras duas, nota-se que PA e PB estão mais próximas do que PA e PC.

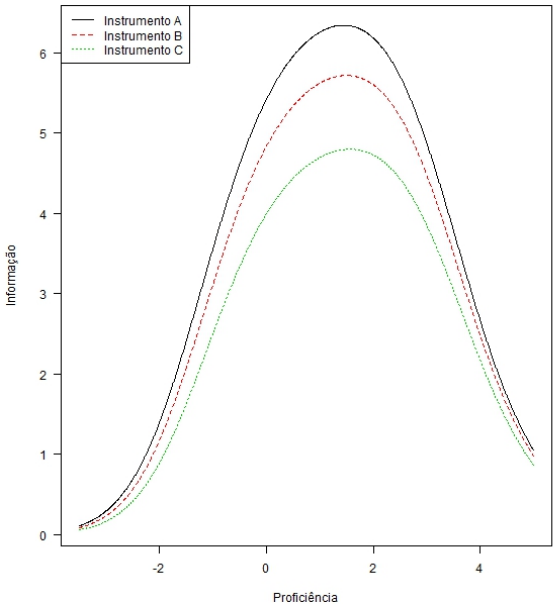


Gráfico 2. Curva de informação dos três instrumentos.

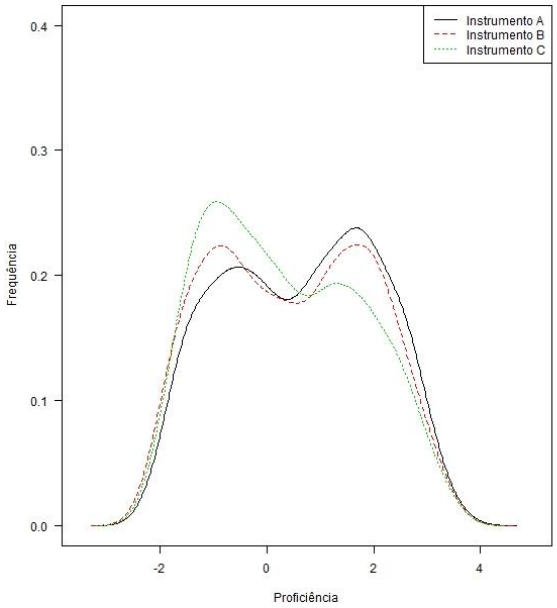


Gráfico 2. Curvas das densidades das proficiências estimadas por meio dos três instrumentos.

Conclusão

Os resultados mostram que as diferenças entre as PA e PB foram pequenas, se comparadas às diferenças entre PA e PC. Portanto, este trabalho aponta para a possibilidade do uso de um instrumento de 45 itens de quatro alternativas, em vez de cinco, para estimar a proficiência de estudantes. Vale ressaltar, no entanto, que este é um estudo preliminar em que foi feita apenas uma simulação, sem contemplar a adoção de repetições do procedimento. Portanto, é necessário que haja novos esforços para verificar a estabilidade dos diferentes estimadores dado que esses resultados podem refletir características da simulação das respostas ou da população sorteada e não apenas da quantidade de alternativas dos itens.

Referências

BRASIL. INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Exame Nacional do Ensino Médio (ENEM):** textos teóricos e metodológicos. Brasília: MEC/INEP, 2009.

_____. INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Guia de elaboração e revisão de itens**. Brasília: MEC/INEP, 2010. Disponível em: <http://download.inep.gov.br/outras_acoes/bni/guia/guia_elaboracao_revisao_itens_2012.pdf>. Acesso em: 19 jun. 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

MODELAGEM PREDITIVA DO TIPO DE ACIDENTE VEICULAR NA BR-101 NO ESTADO BAHIA UTILIZANDO O MODELO DE GRADIENT BOOSTED REGRESSION TREES

Adelmo Menezes de Aguiar Filho[42]
Eduardo Sampaio Soares[43]
Karla Patrícia Santos Oliveira Rodrigues Esquerre[44]
Tarssio Barreto Brito[45]

Resumo

O presente trabalho objetiva compreender a relação entre os tipos de acidentes veiculares e algumas causas possíveis na rodovia BR-101 localizada no Estado da Bahia. Para realizar as análises, uma base de dados do Departamento Nacional de Infraestrutura de Transporte (DNIT) (2005 a 2011) foi utilizada. O modelo de predição foi construído utilizando-se as características espaço-temporais da rodovia aplicada à técnica de gradient boosted regression trees, GBRT, através do pacote LightGBM. A qualidade do modelo foi avaliada com base em um conjunto de dados de teste que representa 30% do total de dados, através da matriz de confusão, a razão de não-informação, bem como o multi-class logarithmic loss. O modelo mostra um bom desempenho qualitativo. A variável mais influente na identificação do tipo de acidente corresponde à localização do trecho da rodovia (km). As condições climáticas, a qualidade e características dos trechos da rodovia, entre outros, são sugeridas para serem utilizadas em estudos futuros como variáveis preditivas como forma de melhorar o desempenho preditivo do modelo.
Palavras-Chave: Acidente veicular, Gradient boosted regression trees, Bahia

Abstract

The present work aims at understanding a relation between vehicular accident types and some possíble causes at BR-101 highway located at Bahia State. In order to carry out the analyzes, a database from the National Department of Transport Infrastructure (DNIT) for seven-year period (2005 to 2011) was utilized. The prediction model was built by using the space-time characteristics of the highway applied to the technique of gradient boosted Regression trees, GBRT, through the LightGBM package. The quality of the model was evaluated based on a test data set which represents 30% of the total data, through the confusion matrix, the non-information ratio, as well as the multi-class logarithmic loss. Model shows a good qualitative performance. The most influential variable in the accident type identification corresponds to the location of the stretch of the highway (km). Climatic conditions, the quality and characteristics of the stretches of the highway, among others, are suggested to be used in future studies as predictive variables as a way to improve model predictive performance.
Keywords: Vehicular accident, Gradient boosted regression trees, Bahia

[42] Universidade Federal da Bahia, adelmo.aguiar.filho@gmail.com
[43] Universidade Federal da Bahia, soares.eduardo.sampaio@gmail.com
[44] Universidade Federal da Bahia, karla.esquerre@gmail.com
[45] Universidade Federal da Bahia, tarssio.disap@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

A rodovia BR – 101 possui um dos maiores trechos rodoviários do Brasil, contando com 4615 km de extensão (DNIT, 2017). Seu traçado liga o extremo do Rio Grande do Norte ao Rio Grande do Sul, atravessando 12 estados. No entanto, por conta do grande tráfego em toda a sua extensão, principalmente devido ao transporte rodoviário de cargas, a quantidade de acidentes é elevada, sendo listada entre as rodovias mais perigosas da malha rodoviária brasileira (Ministério da Justiça e Segurança Pública, 2013). A predição de acidentes em rodovias e sua severidade vem sendo objeto de estudo intenso nos últimos anos (Podofilini et. al., 2015). Assim, o desafio se encontra na identificação de padrões do sistema, possibilitando o melhor entendimento do fenômeno, bem como o seu potencial controle.

Objetivo

Desenvolvimento de modelo preditivo para identificação do tipo de acidente veicular ocorrido no trecho baiano da rodovia BR-101, bem como a compreensão das variáveis de maior influência no fenômeno modelado.

Material e Métodos

Dados provenientes do Departamento Nacional de Infraestrutura de Transportes (DNIT), informando a listagem dos acidentes ocorridos por quilômetro da rodovia e suas características (horário, tipo, severidade etc.), foram obtidos para o período entre os anos de 2005 a 2011 (DNIT, 2017). As informações de hora, mês, dia da semana e o quilômetro da rodovia (em intervalos de 5 km) de ocorrência dos acidentes foram extraídos para compor as variáveis preditoras do modelo. Os tipos de acidentes, variável resposta, informados pelo DNIT foram condensados em quatro grandes categorias: Colisão, Saída, Abalroamento e Capotagem.

A predição do tipo de acidente mais provável dada as características espaço-temporais da rodovia foi realizada através da técnica de *gradient boosted regression trees*, GBRT, através do pacote de funções LightGBM (Guolin, 2016). GBRT constitui uma variação do algoritmo *Random Forest*, na qual as árvores de decisão são construídas sequencialmente, corrigindo o erro associado às árvores de decisão já construídas (Chung, 2013). A avaliação da qualidade do modelo foi realizada com um

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

conjunto de teste (30% do dados) por meio da matriz de confusão e da razão de não-informação e do *multi-class logarithmic loss.*

Resultados e Discussão

Análise exploratória dos dados

A proporção dos tipos de acidentes encontrados na rodovia em questão concorda com o publicado pelo IPEA (2015). Neste relatório sobre acidentes de trânsito nas rodovias federais brasileiras destaca-se como tipos de acidentes mais recorrentes: Colisão Traseira (29%), Colisão Lateral e Transversal (27%) e Saída de Pista (15%). Na rodovia em questão estes valores são, respectivamente: 17%, 19% e 21%.

Por outro lado, as diferenças entre acidentes com atropelamento e capotagem distinguem-se bastante do panorama nacional, nesta rodovia cerca de 15% dos acidentes foram capotamentos e 8% foram atropelamentos, enquanto que no panorama nacional estes valores são, respectivamente, 4,4% e 2,5% (IPEA, 2015). Quanto aos dias de ocorrência dos acidentes, as estatísticas do DPRF (2011) apontam a concentração destes acidentes entre a sexta-feira e o domingo, sendo que na Bahia a ocorrência destes eventos é mais notável aos sábados. No trecho da Rodovia analisada, observa-se um comportamento diferente. Independentemente do tipo de acidente ocorrido, têm-se observado a concentração destes nas segundas-feiras, a explicação desta diferença quanto ao panorama nacional deve ser buscada a partir do entendimento das dinâmicas de transportes inerentes ao trecho. Uma das possíveis interpretações diz respeito ao fato da rodovia cortar polos de desenvolvimentos regionais como cidades de médio porte e regiões próximas a refinarias, indústrias e outros espaços regionais de geração de empregos (Figura 1).

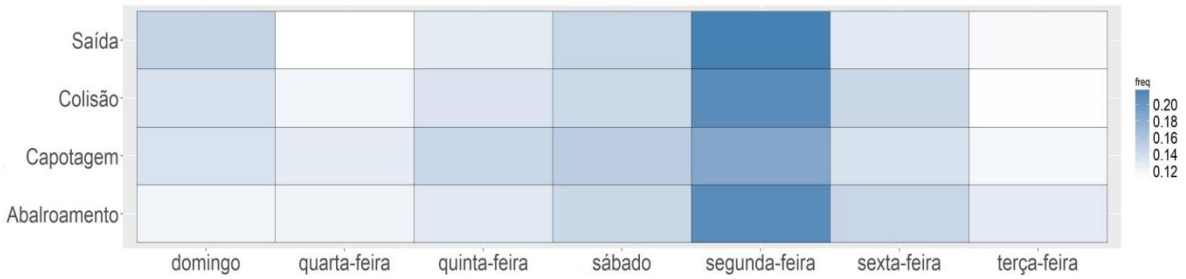


Figura 1. Frequência de acidentes por dia da semana.

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

Corrobora com o exposto acima, a Figura 2, nesta representação gráfica dos acidentes por hora ocorrida. No panorama dos acidentes em rodovias federais, elencados pelo horário de ocorrência, destaca-se como horário de concentração de acidente o período entre as 17 às 20 horas, porém para o trecho analisado este valor é fortemente concentrado próximo às 7 horas. Esta concentração observada no turno da manhã e principalmente às segundas-feiras sugerem que existe um elevado tráfego decorrente de migrações semanais entre as pequenas cidades próximas a BR-101 e os polos empregatícios existentes.

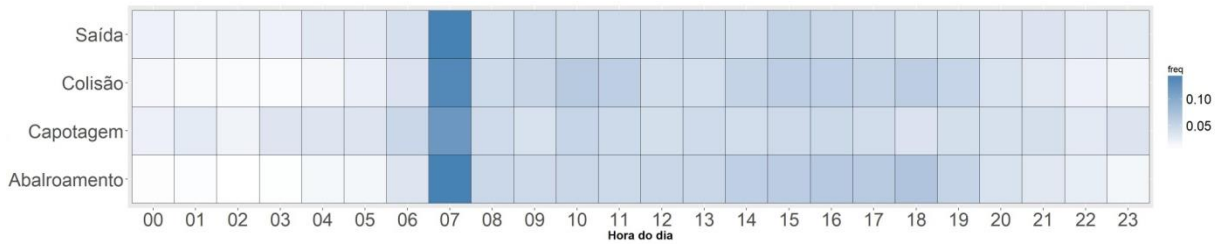


Figura 2. Frequência de acidentes por hora do dia.

Modelo GBRT

Conforme apresentado na Figura 3, a variável de maior influência na decisão do tipo de acidente corresponde a localização do trecho da rodovia (km). Este resultado já era esperado, uma vez que o quilômetro da rodovia resume uma série de informações críticas, como a iluminação e qualidade da pista; limite de velocidade, presença de obstáculos e acostamento etc. As demais variáveis, que compõem as chamadas *características da viagem* (Chung, 2013), mostraram-se ainda relevantes.

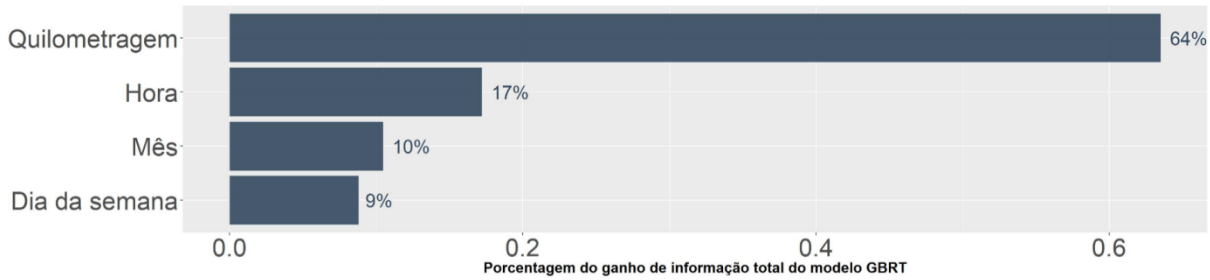


Figura 3. Importância das variáveis preditoras do modelo.

O melhor modelo obteve um valor de 1,26 para *multi-class logarithmic loss*, indicando uma capacidade classificativa equivalente ao dos modelos ganhadores 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

competição similar ocorrida no portal KAGGLE para classificação do tipo de crime na cidade de San Francisco (KAGGLE, 2017). Com uma precisão de 0,4 em relação ao uma razão de não informação de 0,34, verifica-se que o modelo não gera suas predições aleatoriamente, fato corroborado pela matriz de confusão (Figura 4), em que, em geral, dada uma categoria, o modelo prediz com maior frequência a categoria correta.

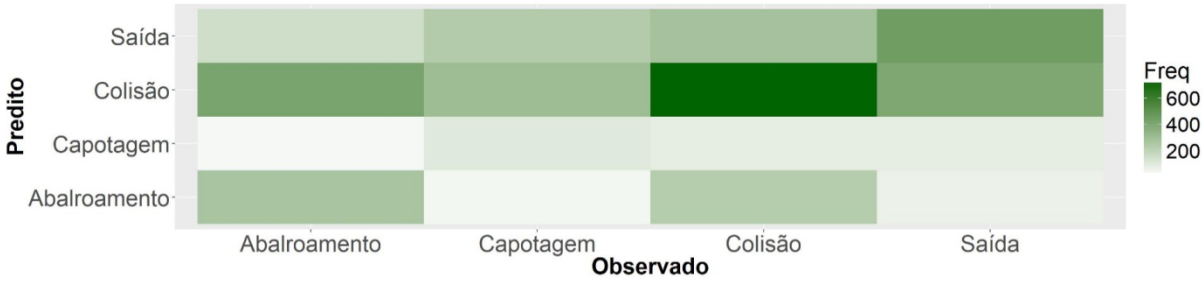


Figura 4. Matriz de confusão do modelo GBRT.

Conclusão

Dentro do contexto do uso de bancos de dados abertos, observa-se que a diversidade de aplicações do aprendizado de máquina exibe potencialidades no que tange ao estudo de acidentes de trânsito em rodovias. Aperfeiçoamentos do modelo ainda são necessários. Para isto, sugere-se a adição de variáveis preditoras que descrevam as condições climáticas, a qualidade e a características dos trechos da rodovia, além das características do motorista envolvido no acidente.

Referências

Chung, Y.S., 2013. Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. Accid. Anal. Prev. 61, 107–118.

DNIT. Banco de Dados de Acidentes de Trânsito. Disponível em <http://www.dnit.gov.br/rodovias/operacoes-rodoviarias >. Acesso em 10 de Março de 2017.

Guolin Ke (2016). lightgbm: Light Gradient Boosting Machine. R package version 0.1.

IPEA. Acidentes de Trânsito nas Rodovias Federais Brasileiras. Relatório de Pesquisa, Brasília, p.7-17, 2015. Anual.

KAGGLE. San Francisco Crime Classification - Public Leaderboard. Disponível em <https://www.kaggle.com/c/sf-crime/leaderboard>. Acesso em 10 de março de 2017.

Podofilini et. al., 2015. Safety and Reliability of Complex Engineered Systems: ESREL 2015. 730 Pages. ISBN 9781138028791.

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 USO DO MÉTODO AHP PARA TOMADA DE DECISÃO: PROPOSTA DE AFILIAÇÃO DA UNIRIO ÀS REDES COLABORATIVAS EM PROL DO DESENVOLVIMENTO SUSTENTÁVEL

Amanda Bergh Navarro[46]
Michelle Cristina Sampaio[47]

Resumo

Diante das evidências e comprovações científicas sobre o avanço da insustentabilidade, é preciso criar novas aspirações para enfrentar os desafios do desenvolvimento sustentável. Dessa forma, o padrão de organização em Redes tem alterado a forma de pensar e agir das organizações, desenvolvendo um novo olhar para estes desafios. Com isso, este trabalho tem por objetivo propor e fundamentar à Universidade Federal do Estado do Rio de Janeiro (UNIRIO) associações às Redes Colaborativas elencadas por Navarro e Sampaio. Para tal, utilizou-se o método Analytic Hierarchy Process (AHP) Auxílio Multicritério a Tomada de Decisão no software R. Dessa forma, as Redes Colaborativas para a UNIRIO associar-se foram: Global University Network for Innovation, Health and Environmental Sciencies Institute, GUPES América Latina e Caribe, Red de Indicadores de Sostenibilidad en las Universidades, Ambientalização Curricular em Instituições de Ensino Superior, e The Sustainability Literacy Test. Por fim, espera-se contribuir para a discussão do papel das Redes Colaborativas como espaço de inovação e de repensar praticas coletivas.
Palavras-Chave: Propósito Comum, Coletivo, Sustentabilidade

Abstract

In the face of scientific evidences on the advancement of unsustainability, new aspirations must be created to face the challenges of the transition to sustainable development. In this way, the pattern of organization in Network has changed the way of thinking and acting of individuals and organizations; developing a new approach for these challenges. With this, the objective of this work is to propose and underpin to the Federal University of the State of Rio de Janeiro (UNIRIO) associations to Collaborative Networks listed by Navarro and Sampaio. For that, the Analytic Hierarchy Process (AHP) was used Multicriteria support on the Decision Making in software R. Therefore, the Collaborative Networks for UNIRIO associate with were: Global University Network for Innovation, Health and Environmental Sciencies Institute, GUPES Latin America and the Caribbean, Network of Indicators of Sustainability in Universities, Curricular Ambientalization in Institutions of Higher Education, and The Sustainability Literacy Test. Lastly, we hope to contribute to the discussion about the role of Collaborative Networks as a space of innovation and rethinking collective practices.
Keywords: Common purpose, Collective, Sustainability

[46] Universidade Federal do Estado do Rio de Janeiro, abergh.navarro@gmail.com
[47] Universidade Federal do Estado do Rio de Janeiro, michelle.sampaio@unirio.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 dos recursos naturais pelo homem tornou-se uma questão emergente, assim como as pressões humanas sobre os processos que regulam o planeta; apontando a incompatibilidade entre desenvolvimento sustentável e os padrões de consumo vigentes (Steffen *et al.*, 2011).

A Organização das Nações Unidas (ONU), em seu documento "Transformando Nosso Mundo: A Agenda 2030 para o Desenvolvimento Sustentável" aponta dezessete Objetivos de Desenvolvimento Sustentável (ODS) com 169 metas associadas a estes. Entre os objetivos vê-se a importância do quadripé: meio ambiente, sociedade, economia e governança; e sua indissociação para a sustentabilidade do desenvolvimento da humanidade (Assembly, 2015).

Visto isto, Navarro e Sampaio (2017) abordam Redes Colaborativas como um padrão de organização coerente com os desafios da sustentabilidade; elencam 35 Redes Colaborativas que visam o desenvolvimento sustentável e alinham com os ODS, tipologizadas em Redes Colaborativas entre organizações (RCO), Redes Colaborativas em prol de uma metodologia (RCM) e Redes Colaborativas entre pesquisadores (RCP).

Objetivo

Propor e fundamentar à Universidade Federal do Estado do Rio de Janeiro (UNIRIO) associações às Redes Colaborativas em prol do Desenvolvimento Sustentável elencadas por Navarro e Sampaio (2017), que possuam objetivos comuns às da UNIRIO, segundo o método *Analytic Hierarchy Process* (AHP) Auxílio Multicritério a Tomada de Decisão no *software* R.

Material e Métodos

Mapearam-se as iniciativas institucionais na UNIRIO já instauradas que visem à sustentabilidade – projetos de Iniciação Científica (IC) e Redes em que a UNIRIO já é afiliada – e suas conformidades com os ODS; realizou-se e aplicou-se questionários junto a Coordenação de Relações Internacionais – CRI/ UNIRIO e a Comissão Permanente de Sustentabilidade Institucional – COPESI/ UNIRIO; realizaram-se e aplicaram-se questionários junto às Redes elencadas por Navarro e Sampaio (2017) as quais foram segregadas em dez perfis segundo o alinhamento com os ODS: (i) aos ODS 4, 9 e 17; (ii) aos ODS 4 e 17; (iii) aos ODS 9 e 17; (iv) aos ODS 10, 11 e 17; (v)

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

aos ODS 3, 6, 14 e 17; (vi) aos ODS 4, 8, 16 e 17; (vii) aos ODS 2, 6, 7, 14, 15 e 17; (viii) 3, 7, 8, 14, 15 E 17; (ix) apenas ao ODS 17; e (x) todos os ODS.

Por conseguinte, realizou-se a análise segundo o método AHP, técnica qualitativa de análise de decisão e planejamento de múltiplos critérios (Saaty, 1991), no *software* R para fundamentar as possíveis associações da UNIRIO em Redes Colaborativas. Segundo Saaty (1991), a segurança das tomadas de decisões é dada por meio de dois processos de análise: inconsistência e sensibilidade. Para Saaty (1991), considera-se aceitável o valor de inconsistência < 0,10.

Resultados e Discussão

A partir das entrevistas realizadas às Redes elencadas por Navarro e Sampaio (2017) e dos questionários à CRI e à COPESI, foi possível a identificação dos critérios de decisão: objetivo, tipo de setor e público alvo. E, segundo as informações coletadas nas entrevistas com a COPESI e CRI, dos projetos de IC, e das Redes em que a UNIRIO já é filiada, foi possível realizar as relações hierárquicas das alternativas (Tabela 1) para as atribuições dos pesos.

Tabela 1: Relações hierárquicas dos níveis de importância entre os critérios decisores e entre os perfis das Redes; das RCO e RCM para a fundamentação das paridades.

Grupo	Critérios decisores	Relações dos níveis de importância entre os critérios decisores	Relações dos níveis de importância entre os perfis das redes
RCO	Objetivos	Objetivo > Público Alvo > Tipo de setor	ODS 4, 9 e 17> ODS 4 e17 > ODS 9 e 17> ODS 10, 11 e 17 > Todos os ODS> ODS 3, 6, 14 e 17 > ODS 2, 6, 7, 14, 15 e 17 = ODS 4,8, 16 e 17 > ODS 17.
	Tipo de setor		Público> Privado.
	Público Alvo		Universidades> Universidades e outros.
RCM	Objetivos	Objetivo > Público Alvo	ODS 4 e 9> ODS 9 e 17> Todos os ODS> ODS 3, 7, 8, 14, 15 e 17> ODS 17
	Público Alvo		Universidades> Universidades e outros.

*Todas as RCM são públicas, e, portanto, não foi necessário que o critério "tipo de setor" fosse um critério de escolha. Fonte: Autoria própria.

Nesse sentido, obteve-se que as RCO de maior peso são: Global University Network for Innovation com 11,30%, ILSI Health and Environmental Sciencies Institute com 10,50%, e GUPES América Latina e Caribe com 9,10%; e que as RCM de maior

peso são: Red de Indicadores de Sostenibilidad en las Universidades com 21,90%, Ambientalização Curricular em Instituições de Ensino Superior com 21,90%, e The Sustainability Literacy Test com 16,40%.

Ademais, ambas as análises de sensibilidade deram no critério ”objetivo”, índice de inconsistência de 0,044 e 0,04 para as RCO e RCM, respectivamente; 0,0 para o critério “tipo de setor” na análise das RCM; 0,0 para o critério “público alvo” e índice geral de inconsistência para ambas as análises.

Por fim, considera-se relevante ressaltar que os resultados são inerentes ao método, já que os tomadores de decisão têm grande participação na priorização das alternativas; e que, por mais que inovação seja, por vezes, necessária, o método não consegue propor soluções contrárias à cultura organizacional.

Conclusão

Este trabalho propôs e fundamentou associações da Universidade Federal do Estado do Rio de Janeiro às Redes Colaborativas em prol do Desenvolvimento Sustentável elencadas por Navarro e Sampaio (2017), segundo o método *Analytic Hierarchy Process* (AHP) Auxílio Multicritério a Tomada de Decisão no *software* R, que se destacaram por seu maior peso, frente às demais Redes, colocando-se como Redes prioritárias para a UNIRIO associar-se.

Referências

ASSEMBLY, U. G. Transforming our world: the 2030 Agenda for Sustainable Development. **New York: United Nations**, 2015.

NAVARRO, A. B.; SAMPAIO, M. C. **Redes Colaborativas em prol do Desenvolvimento Sustentável**. Congresso Nacional de Excelência em Gestão. Rio de Janeiro. XII. 2016.

_____. **Collaborative Networks and Their Relation to Sustainable Development Goals**. International Conference on Environmental, Cultural, Economic & Social Sustainability. Niterói: Common Ground. XIII 2017.

SAATY, T. L. **Método de Análise Hierárquica**. São Paulo: Mc Graw-Hill, 1991.

STEFFEN, W. et al. The Anthropocene: From global change to planetary stewardship. **Ambio,** v. 40, n. 7, p. 739-761, 2011. ISSN 0044-7447.

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

UTILIZANDO O PACOTE SHINY / R-PROJECT NA MODELAGEM DE ESTRUTURAS A TERMO DE JUROS

João Dantas de Melo Neto[48]
Marco Aurélio dos Santos Sanfins[49]
Valentin Sisko[50]
Daiane Rodrigues dos Santos[51]

Resumo

No presente artigo, analisa-se o uso do pacote Shiny/R-project na modelagem da Estrutura a Termo de Juros. A Estrutura a Termo de Juros é de suma importância, pois conseguimos analisar a taxa futura de juros a partir do método de Nelson Siegel (1987) e usando-a juntamente com o Shiny podemos observar essa taxa de um modo interativo. Assim, podemos escolher o período e conseguimos fazer uma análise mais rápida e prática da Estrutura a Termo de Juros. A utilização do pacote Shiny na análise da Estrutura a Termo de Juros é um modo mais prático, rápido e fácil, pois com o shiny conseguimos analisar de forma interativa e podemos compartilhar a experiência de modo gratuito e online. O que poderia facilitar o compartilhamento para que todos os interessados pudessem também ter uma experiência sobre a Estrutura a Termo de Juros.
Palavras-Chave: Shiny, Nelson Siegel, Estrutura a Termo de Juros

Abstract

In this article, we analyze the use of the Shiny / R-project package in the modeling of the yield curve. The yield curve is of great importance, since we can analyze the future rate of interest from Nelson Siegel's method (1987) and using it together with Shiny we can observe this rate in an interactive way. Thus, we can choose the period and we were able to make a quicker and more practical analysis of the yield curve. The use of the Shiny package in the Analysis of the yield curve is a more practical, quick and easy way, because with the shiny we can analyze this data in an interactive way and we can share the experience for free and online. This could facilitate the sharing so that all interested parties could also have an experience on the yield curve.
Keywords: Shiny package, Nelson Siegel's method, The Structure of the Term of Interest, yield curve

Introdução

O Shiny é um pacote de R-project que torna fácil a construção de aplicações interativas na web, permitindo o compartilhamento de análises e gráficos feitos pelo software. Essa ferramenta está em amplo desenvolvimento no *R-project*, pode ser largamente utilizada em qualquer lugar no qual a internet esteja presente. Com

[48] UFF – Universidade Federal Fluminense, joaodantas@id.uff.br
[49] UFF – Universidade Federal Fluminense, marcosanfins@automata.uff.br
[50] UFF – Universidade Federal Fluminense, valentin33@gmail.com
[51] UFRRJ – Universidade Federal Rural do Rio de Janeiro, daianesantoseco@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

certeza o pacote *Shiny,* em um futuro próximo, tem um grande potencial para revolucionar as formas de comunicação científicas atualmente empregadas.

Objetivo

No mercado financeiro em geral, como pode ser visto em Chauvert e Potter, (2001), Hamilton (2002), Estrella (2005), Braun (2010), os agentes sempre estão procurando modelar o comportamento das taxas de juros para qualquer intermediação financeira que se faça necessária. Essas taxas de juros são facilmente identificadas e modeladas pelos modelos de Estrutura a Termo das taxas de Juros (ETTJ), inicialmente apresentadas nos estudos de Nelson e Siegel (1981). O objetivo desse trabalho é utilizar os modelos implementados por Santos, Ramalhete e Sanfins (2011) e com seus respectivos *outputs* gerar, via web - através do *Shiny* - resultados que possam ser visualizados por qualquer usuário da web que tenha interesse nesse tema e deseje obter informações mais precisas sobre o comportamento das ETTJ.

Segundo Rocha et. al. (2015), "o *Shiny* combina o poder computacional do *software* R com a interatividade da web moderna, o que o torna uma opção interessante para uso em computação científica".

De acordo com Doi et. al. (2016), "*Shiny* é um framework de aplicação web para R que requer apenas conhecimento na linguagem de programação R. Com *Shiny,* pode-se construir uma ferramenta de ensino que seja interativa, dinâmica, fácil de usar, visualmente atraente e com funcionalidade semelhante ao Java".

Material e Métodos

Basicamente os métodos utilizados neste trabalho, consistem em implementar os *output* que foram desenvolvidos pelo sistema criado por Santos, Ramalhete e Sanfins em 2011. Estes *outputs,* como também outros de interesse desta linha de pesquisa, já se encontram em fase de implementação via *web* através do pacote *Shiny*.

Resultados e Discussão

Os resultados obtidos até aqui são bem promissores, principalmente por se tratarem de inovações no mundo acadêmico e sem fins lucrativos. Essa maneira inovadora de visualização de dados fará com que muitos interessados nessa aplicação específica, que muitas das vezes não possuem conhecimentos avançados

em programação, possam obter informações relevantes. Nas imagens abaixo encontramos o *output* do programa *shinny* com a tabela e o gráfico dos dados coletados para um determinado instante de tempo t (04/04/2017).

Escolha uma Data: aaaa-mm-dd

2017-04-04

Ordem	Codigo	PU	Duration	Taxa	Ajuste
1	A00	99954.62	1	12.12	15.18
2	K17	98994.81	12	23.63	14.56
3	M17	98087.51	34	15.39	13.50
4	N17	97300.33	55	13.36	12.68
5	Q17	96524.31	76	12.45	12.01
6	U17	95731.50	99	11.74	11.43
7	V17	95075.13	119	11.29	11.02
8	X17	94391.91	140	10.95	10.68
9	Z17	93763.26	159	10.75	10.43
10	F18	93133.54	179	10.53	10.22
11	J18	91232.55	239	10.16	9.81
12	N18	89295.90	302	9.91	9.63
13	V18	87325.30	366	9.78	9.58
14	F19	85437.83	428	9.71	9.59
15	J19	83506.76	488	9.75	9.63
16	N19	81584.93	550	9.77	9.69
17	V19	79573.94	616	9.80	9.75
18	F20	77631.09	680	9.84	9.80
19	N20	73939.70	803	9.94	9.90
20	V20	72069.81	868	9.98	9.94
21	F21	70307.19	931	10.01	9.98
22	N21	66936.31	1053	10.08	10.04
23	F22	63640.91	1181	10.12	10.09
24	F23	57581.25	1431	10.21	10.16
25	F25	47195.41	1932	10.29	10.25
26	F26	42712.68	2185	10.31	10.28
27	F27	38611.31	2435	10.35	10.31

Tabela 1: Output do shiny com os dados utilizados na modelagem da curva de juros do dia 04/04/2017.

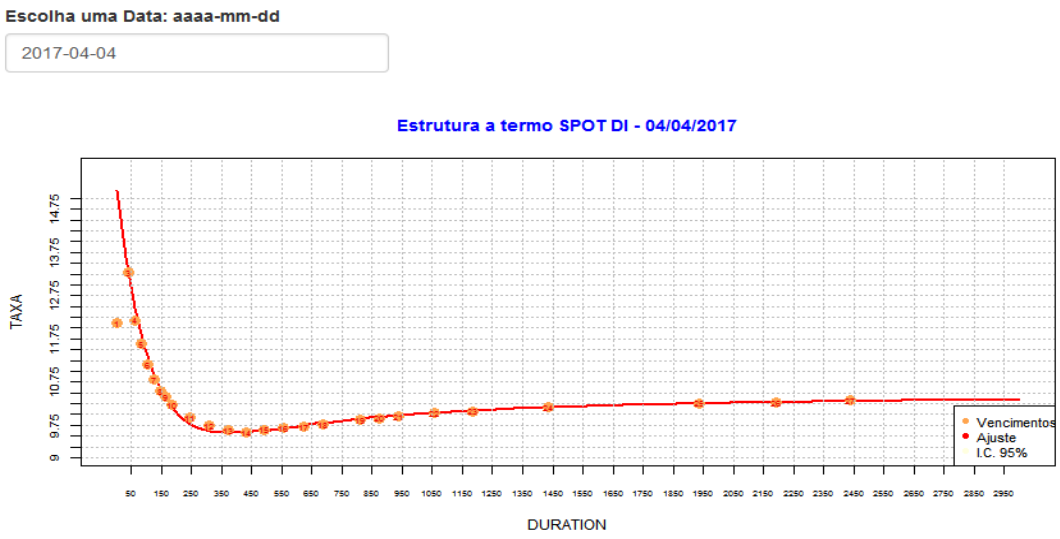


Gráfico 1: Output do shiny com os dados utilizados na modelagem da curva de juros do dia 04/04/2017.

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

É importante salientar que atualmente essas informações, em sua grande maioria, são comercializadas por empresas especializadas com um custo muito alto, impossibilitando o acesso a grande maioria das pessoas.

Conclusão

A gama de possibilidades que o pacote Shiny propicia e as ferramentas atualmente empregadas no mercado financeiro fizeram com que os participantes do projeto refletissem e chegassem à conclusão de que estamos apenas no começo de novo paradigma, acerca de como as informações científicas irão se difundir pelo mundo. Este trabalho foi de extrema importância e relevância para os alunos e para os pesquisadores envolvidos na descoberta de uma fronteira de como os *outputs* podem ser interativos e divulgados via *web*.

Referências

[1] BRAUN L. F.; **Risco de Taxa de Juros em Fundos e Ações – O Impacto de Nível, Inclinação e Curva de Juros da Estrutura a Termo**. Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciências Atuariais do Instituto de Gestão de Riscos Financeiros e Atuariais da PUC-Rio, 2010.

[2] CHAUVET, M. AND POTTER S.; **Forecasting recessions using the yield curve. Riverside**. Journal of Forecasting, 2001.

[3] DOI, J.; POTTER, G.; WONG, J.; ALCARAZ, I. and CHI, P.; **Web Application Teaching Tools for Statistics Using R and Shiny**. Technology Innovations in Statistics Education, 2016.

[4] ESTRELLA, A.; Why **Does the Yield Curve Predict Output and Inflation?** Economic Journal, Vol. 115, No. 505, pp. 722-744, 2005.

[5] LITTERMAN, R. E SCHEINKMAN, J., **Common Factors Affecting Bond Returns**, The Journal of Fixed Income, 1991.

[6] NELSON C. and A. SIEGEL. **Parsimonious Modeling of Yield Curves**. Journal of Business, 60, 4, 473-489, 1987.

[7] HAMILTON, J. D. and D. H. KIM. **A Re-Examination of the Predictability of Economic Activity Using the Yield Spread**. Journal of Money, Vol. 34, 2002.

[8] Pacote SHINY, R-project, https://cran.r-project.org/web/packages/shiny/index.html

[9] R-project, https://cran.r-project.org/

[10] ROCHA, B. F.; TERNES, S.; ROSSI, M. **Disseminando a aplicação do R-Shiny em métodos quantitativos e computação científica na web.** MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 11, 2015.

[11] SANTOS, D. R.; RAMALHETE, R. M. ; SANFINS, M. A. **Modelagem Da Estrutura a Termo da Taxa de Juros - ETTJ**. IV congresso UFSC de Controladoria e Finanças, 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

ESTIMAÇÃO ESTOCÁSTICA DA ESTRUTURA A TERMO DAS TAXAS DE JUROS SOBERANA UTILIZANDO A TÉCNICA DE SIMULATED ANNEALING

Beatriz Jardim Pina Rodrigues[52]
Marco Aurélio dos Santos Sanfins[53]
Valentin Sisko[54]
Daiane Rodrigues dos Santos[55]

Resumo

No presente artigo, analisa-se a Estrutura a Termo de Taxa de Juros (ETTJ) pelo método de Nelson Siegel (1987). A ETTJ é extremamente importante, pois demonstra o comportamento da taxa de juros por período de tempo. Assim, fornece informações sobre a expectativa de inflação futura e de crescimento do produto. Esse trabalho irá modelar a ETTJ, partindo da proposta de Siegel (1987), utilizando o método do *Simulated Annealling*. Tal método tem como principal vantagem o fato de poder otimizar problemas com um grande número de parâmetros. O modelo de Siegel (1987) contém quatro parâmetros a serem estimados que facilitam a identificação dos movimentos na curva de juros. São eles: o nível, a inclinação, a curvatura na curva de juros e o seu ponto de inflexão. Para estimação dos componentes principais usa-se *o Simulated Annealing*, que pode ser considerado um método heurístico melhorativo de busca aleatória na vizinhança. Utilizaremos aplicações do R-project com pacote Shiny para realização dos cálculos do *Simulated Annealing* e a elaboração dos gráficos da ETTJ.
Palavras-Chave: ETTJ, Simulated Annealing, Componentes Principais, Modelo de Nelson Siegel

Abstract

In this article, the yield curve is analyzed by Nelson Siegel's method (1987). The yield curve is extremely important, as it demonstrates the behavior of the interest rate by period of time. Thus, it provides information on the expected future inflation and output growth. This work will model the yield curve, starting from the proposal of Siegel (1987), using the method of *Simulated Annealing*. This method has the main advantage of being able to optimize problems with a large number of parameters. The model proposed by Siegel (1987) contains four parameters to be estimated that facilitate the identification of movements in the interest curve. They are: the level, the slope, the curvature in the curve of interest and its point of inflection. For the estimation of the main components we use *Simulated Annealing*, which can be considered an heuristic method of random search in the neighborhood. We will use applications of R-project with Shiny package to realize the calculations of the *Simulated Annealing* and the elaboration of the graphs of yield curve.
Keywords: TSIR, Simulated Annealing, Main Components, Nelson Siegel's model

[52] UFF – Universidade Federal Fluminense, beatrizjardim@id.uff.br
[53] UFF – Universidade Federal Fluminense, marcosanfins@automata.uff.br
[54] UFF – Universidade Federal Fluminense, valentin33@gmail.com
[55] UFRRJ – Universidade Federal Rural do Rio de Janeiro, daianesantoseco@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 modelagem da Estrutura a Termo das Taxas de Juros (ETTJ) tem recebido várias contribuições na literatura. Em um número importante destes trabalhos busca-se a estimação de componentes, tipicamente aditivas responsáveis por características bem definidas das curvas de juros, como pode ser visto em Diebold e Li (2006), Bressan (2007) e Santos (2014).

O trabalho mais conhecido nesta linha é possivelmente o artigo de Litterman e Scheinkman (1991), no qual é utilizada a técnica de Componentes Principais. Nesse artigo, de uma maneira exploratória, os autores identificam três componentes capazes de explicar algo em torno de 98% da variabilidade das taxas implícitas de papéis de várias maturidades no mercado americano. A maior importância da decomposição obtida foi a possibilidade de interpretação dos movimentos que ocorrem com as taxas de juros. Com efeito, Litterman e Scheinkman (1991) identificam as três componentes como sendo responsáveis, respectivamente e em ordem de importância, por movimentos no nível, na inclinação e na curvatura da curva de juros.

Objetivo

A análise de Componentes Principais é, essencialmente, técnica de análise exploratória de dados visando à redução de dimensionalidade. Seu uso mais difundido visa indicar estruturas passíveis de serem modeladas estatisticamente em processos similares aqueles estudados. Neste sentido Vieira Neto (2001), desenvolve um modelo para a evolução da Estrutura a Termo de Taxas de Juros impondo uma forma analítica para as três componentes identificadas em Litterman e Scheinkman (1991), além de propor formas analíticas para uma quarta componente. O objetivo deste trabalho é estimar as quatro componentes propostas pelo modelo de Nelson Siegel (1987) utilizado no artigo de Vieira Neto (2001) usando o método de simulação estocástica - *Simulated Annealing* - que consiste em um algoritmo iterativo utilizado para problemas de otimização. Seu diferencial é poder otimizar problemas com um grande número de parâmetros.

Material e Métodos

O modelo de Nelson Siegel (1987), conforme expressão abaixo contém quatro parâmetros a serem estimados, entretanto, como forma de facilitar o processo de estimação, a maioria dos métodos utilizados fixa um valor para o parâmetro λ. Esse

parâmetro governa a taxa de decaimento exponencial, pequenos (ou grandes) valores de λ estão associados a um decaimento suave (ou rápido), segundo em Diebold e Li (2006) a inclusão deste parâmetro possibilita um ajuste melhor para às maturidades longas (ou curtas).

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau}\right) + \beta_{3t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau}\right) \qquad (1)$$

A principal justificativa deste trabalho é a possibilidade da estimação de todos os parâmetros envolvidos no modelo, fato este que se conseguiu utilizando-se as técnicas de simulação estocástica.

Resultados e Discussão

Para modelar a ETTJ foi utilizado no processo de otimização a seguinte função perda:

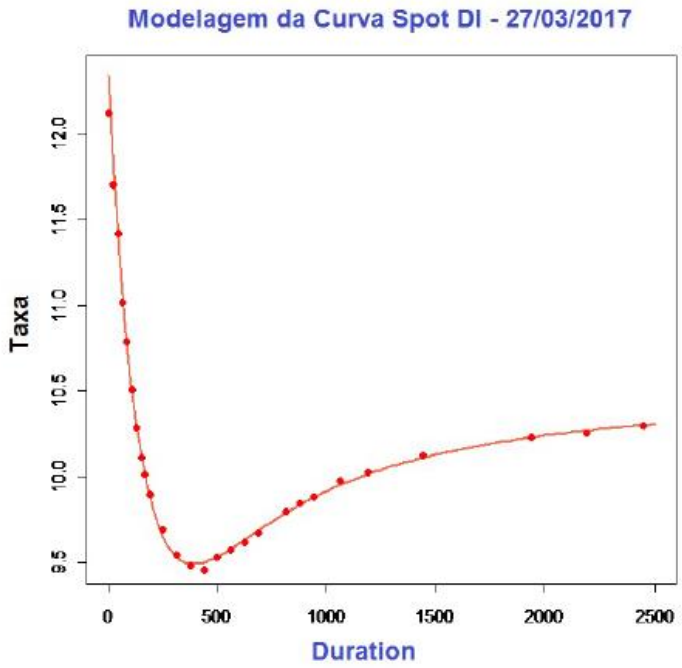
$$\underset{\sim}{\hat{\beta}_t} = \min \sum_{\tau} [Y_t(\tau) - y_t(\tau)]^2 \qquad (2)$$

Onde $Y_t(\tau)$ é a taxa observada no instante t para os níveis de τ (*durations*), e $y_t(\tau)$ é o valor obtido através da curva ajustada pelo modelo de Nelson Siegel. Na tabela abaixo encontramos os dados coletados para um determinado instante de tempo t (27/03/2017).

Código	PU (Preço Unitário)	Duration (Duração)	Taxa	Limite Inferior	Ajuste	Limite Superior
A00	99,959.32	1	10.80	10.74	10.62	10.78
Q14	98,544.19	36	10.81	10.55	10.70	10.86
U14	97,704.19	57	10.81	10.59	10.75	10.90
V14	96,833.33	79	10.81	10.64	10.79	10.95
F15	94,303.17	144	10.81	10.76	10.92	11.07
J15	91,947.64	205	10.88	10.87	11.02	11.17
N15	89,526.71	266	11.05	10.96	11.11	11.27
F16	84,587.30	394	11.30	11.13	11.28	11.43
J16	82,289.83	455	11.44	11.19	11.35	11.50
N16	79,995.23	518	11.52	11.25	11.41	11.56
V16	77,713.20	583	11.55	11.31	11.47	11.62
F17	75,596.29	645	11.66	11.36	11.52	11.67
N17	71,500.83	769	11.71	11.45	11.60	11.75
F18	67,513.05	894	11.74	11.52	11.67	11.83
F19	60,317.36	1144	11.78	11.63	11.78	11.94
F20	53,836.04	1397	11.81	11.71	11.86	12.02
N20	50,836.04	1520	11.87	11.74	11.89	12.05
F21	48,048.76	1648	11.86	11.77	11.92	12.07
F22	42,800.17	1899	11.92	11.81	11.97	12.12
F23	38,026.32	2150	12.00	11.85	12.00	12.15
F25	29,988.00	2653	12.10	11.90	12.05	12.20

Tabela 1: Dados utilizados na modelagem da curva de juros do dia 27/03/2017.

A seguir os valores observados, bem como os valores ajustados da ETTJ, estão plotados.



Todos os *outputs*, bem como a modelagem e estimação dos parâmetros da ETTJ, foram feitos por meio de *scripts* desenvolvidos na linguagem *R-Project*. Um dos maiores problemas no desenvolvimento dos *scripts* foi relacionado ao *input* para a estimação dos parâmetros, isso porque os dados estão disponíveis em planilhas Excel no site da BMF-Bovespa. Uma solução encontrada foi automatizar a leitura dos dados, como forma de possibilitar, em um momento futuro, novas modelagens para outras datas. Para essa tarefa, foi utilizada a função *download.file,* a qual possibilita realizar o download de um determinado arquivo. Tal arquivo, em nosso caso, foi a planilha de dados que contém informações sobre os preços unitários de cada título de DI (depósitos Interbancários) e que, como exemplo, ilustramos os dados coletados para o dia 27/03/2017 e os mesmos se encontram na tabela 1.

Como todos sabem, o método *Simulated Annealing* é uma poderosa ferramenta de otimização e, por consequência, extremamente eficiente no tocante à estimação dos parâmetros envolvidos na curva de Nelson Siegel. Outros métodos computacionalmente menos dispendiosos e tão precisos quanto o *SA* poderiam ser utilizados na estimação. Entretanto, a escolha pelo *SA* foi feita devido a sua característica de sempre encontrar uma solução para a estimação, desde que o número de interações necessárias tenda para infinito para uma determinada precisão desejada. Essa vantagem do *SA* faz com que sempre possamos estimar a ETTJ, o

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 para outros métodos de estimação isso, por determinadas circunstâncias, não possa ser garantido.

Conclusão

Neste trabalho foi proposto como diferencial a modelagem da Estrutura a Termo da Taxa de Juros - ETTJ pelo Método descrito por Nelson Siegel (1987) utilizando a técnica *Simulated Annealing*. Após o ajuste, ficou claro que tanto o modelo como a forma empregada para obter os valores dos parâmetros da curva de Nelson Siegel foram, a princípio, adequados.

Este pôster é a extensão do trabalho apresentado por Santos, Ramalhete e Sanfins em 2011 e, em conjunto com a aplicação do pacote Shiny do R-project (que torna fácil a construção de aplicações interativas na *web* e permite o compartilhamento de análises e gráficos feitos pelo *software),* faz parte de um projeto de extrema importância e relevância para o mercado financeiro. Pois avança no conhecimento acerca da Estrutura a Termo das Taxas de Juros e torna o acesso as informações mais igualitário, visto que o projeto não tem fins lucrativos.

Referências

[1] BRESSAN, A. A.; ALVES, R. A.; CAETANO, R. A. e IGUIAPAZA, R. A.; *"Modelagem Multifatorial da Estrutura a Termo de Juros de LTN's utilizando Análise de Componentes Principais".* Encontro ANPEC, 2007.

[2] DIEBOLD, F. X. and LI. C.; *"Forecasting The Term Structure Of Government Bond Yields. Journal of Econometrics*". v130, 337-364, 2006.

[3] LITTERMAN, R. E SCHEINKMAN, J., *"Common Factors Affecting Bond Returns",* The Journal of Fixed Income, pp. 54-61, 1991.

[4] NELSON C. and A. SIEGEL; *"Parsimonious Modeling of Yield Curves".* Journal of Business, 60, 4, 473-489. 1987.

[5] SANTOS J.; "*GESTÃO DE RISCO CAMBIAL NO AMBIENTE CORPORATIVO: Aplicação da análise de componentes principais para a gestão do risco cambial em Trading Companies brasileiras."* Dissertação apresentada à Escola de Economia da Fundação Getulio Vargas (FGV/EESP), 2014.

[6] SANTOS, D. R.; RAMALHETE, R. M. ; SANFINS, M. A. "*Modelagem Da Estrutura a Termo da Taxa de Juros - ETTJ."* IV congresso UFSC de Controladoria e Finanças, 2011.

[7] VALLI, M. e VARGA, G. *"Movimentos da Estrutura a Termo da Taxa de Juros Brasileira e Imunização".* Revista de Estudos Avançados da USP, 2001.

[8] VARGA, G. *"Interpolação por Cubic Spline para a Estrutura a Termo Brasileira".* Resenha BM&F no 140, pp. 29-35, 2000.

[9] VIEIRA NETO, C.A., *"Modelagem da Estrutura a Termo da Taxa de Juros."* Tese de doutorado, Universidade de São Paulo- USP, 2001.

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

SHINY EM GRÁFICOS DE CONTROLE ESTATÍSTICO DE PROCESSOS

Andréa Cristina Konrath[56]
Rodrigo Gabriel de Miranda[57]
Elisa Henning[58]
Olga Maria Formigoni Carvalho Walter[59]

Resumo

Este trabalho tem como objetivo desenvolver um aplicativo no ambiente Shiny com foco no controle estatístico da qualidade. O Shiny é um pacote para usuários do software R, um ambiente para computação estatística. No aplicativo desenvolvido são apresentadas algumas medidas descritivas como média e desvio padrão, além do histograma e box plot do conjunto de dados. Também é apresentado o gráfico de controle para a média e amplitude, podendo construir tanto a fase I quanto a fase II do gráfico de controle. Conclui-se que o desenvolvimento deste aplicativo é uma alternativa para expandir as aplicações de gráficos de controle neste ambiente e que o Shiny é mais amigável de ser utilizado quando comparado com o RStudio. Para estudos futuros, pretende-se implementar os gráficos de controle para atributos e os gráficos de controle avançados como o Exponentially Weighted Moving Average (EWMA), o Cumulative Sum (CUSUM), e os índices de capacidade e ARL.
Palavras-Chave: Software R, Gráficos de Controle, Estatística, Shiny

Abstract

The purpose of this paper is to develop an application in the Shiny environment focused on the quality statistical control. Shiny is a package for software R users, an environment for statistical computing. A few descriptive measures are provided in the application, such as average and standard deviation, as well as the histogram and the box plot of the data set. The control chart is also presented for the average and the amplitude, allowing the user to build phase I as well as phase II of the control charts. The conclusion reached is that the development of this application is an alternative to expand the usage of control charts in this environment. Also, Shiny was considered more user friendly when compared to RStudio. In future research, is possible to implement the control charts for attributes, as well as the advanced control charts, such as Exponentially Weighted Moving Average (EWMA), Cumulative Sum (CUSUM), capability indices and ARL.
Keywords: Software R, Control Charts, Statistics, Shiny

Introdução

Gráficos de controle são amplamente utilizados para visualizar e compreender os diferentes tipos de variação de um processo. Um gráfico de controle, é constituído por uma linha central (LC) que representa o valor médio da característica da qualidade correspondente à situação do processo sob controle e dois limites de controle: um

[56] Universidade Federal de Santa Catarina, andrea.ck@ufsc.br
[57] Universidade do Estado de Santa Catarina, rgabrieldemiranda@yahoo.com.br
[58] Universidade do Estado de Santa Catarina, elisa.henning@gmail.com
[59] Universidade do Estado de Santa Catarina, olgaformigoni@gmail.com

deles situado abaixo da LC denominado limite inferior de controle (LIC) e, outro acima da LC, denominado limite superior de controle (LSC). Formalmente, ambos os limites ficam a uma distância de três erros-padrão da média ($\mu \pm 3\sigma$) (SAMOHYL, 2009).

Com o avanço tecnológico, a utilização de pacotes computacionais tem sido mais frequente tanto no meio acadêmico, como em ambientes empresariais (HENNING, ALVES; VIEIRA, 2007). Tratando de gráficos de controle, o pacote qcc (SCRUCCA, 2004), desenvolvido no *software* R (R DEVELOPMENT CORE TEAM, 2017) permite a construção os gráficos de controle. O R possui uma *interface* gráfica amigável, o RStudio. Com o RStudio é possível viabilizar o uso de R diretamente na internet por meio do pacote Shiny (CHANG et al., 2016), que ainda é pouco explorado na literatura nacional (KONRATH et al., 2013; HENNING et al., 2016; MULLER; ZABALA, 2016). Com o Shiny é possível substituir tarefas complexas como a programação de funções em R em aplicações interativas na web.

Objetivo

Este artigo tem como objetivo desenvolver um aplicativo no ambiente Shiny com foco no controle estatístico da qualidade.

Material e Métodos

Esta pesquisa quantitativa é considerada de natureza aplicada, uma vez que se caracteriza pelo interesse prático, ou seja, seus resultados são utilizados para solucionar problemas reais (GIL, 2010). A análise estatística foi realizada pelo *software* R (R CORE TEAM, 2017), utilizando os pacotes Shiny (CHANG et al., 2016) e qcc (SCRUCCA, 2004). O conjunto de dados está disponível em MONTGOMERY (2009) e também no R no pacote qcc. São 40 amostras de tamanho $n = 5$ do diâmetro interno (em mm) dos anéis de pistão do motor de automóvel.

Resultados e Discussão

Na Figura 1 é apresentada a tela principal da primeira versão do aplicativo desenvolvido, que pode ser acessada no seguinte link: https://r-nnq.shinyapps.io/index. O aplicativo é composto pelas seguintes abas: Gráfico de Controle; Dados; Resumo de Dados; Gráficos de Controle (fase1 e fase 2). São também apresentados, alguns *sidebar,* como por exemplo, o usuário poderá importar seu dados de um arquivo (xls, xlsx), selecionar o tipo de gráficos de controle que

podem ser por variáveis ou atributos e selecionar o Gráfico de Controle, conforme a seleção do item anterior, de acordo que é apresentado no *sidebar.*

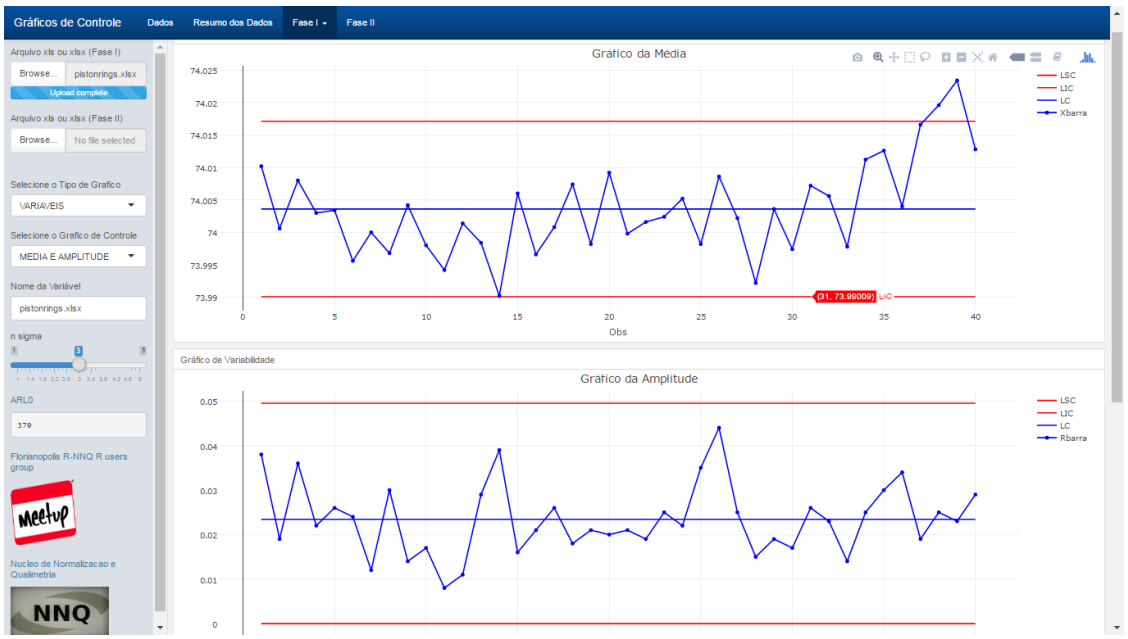


Figura 1: Gráfico de Controle para média e amplitude.

Conclusão

Este artigo teve o objetivo desenvolver um aplicativo no ambiente Shiny com foco no controle estatístico da qualidade, mais especificamente nos gráficos de controle. Foi apresentado o gráfico de controle para a média e amplitude. Constata-se que o Shiny se torna um meio mais amigável de ser utilizado quando comparado com o RStudio. Além de que a interação do usuário com o aplicativo, permite visualizar as informações instantaneamente a medida que são modificados os parâmetros das aplicações, o que aumenta o interesse na exploração dos gráficos, sem necessitar conhecimento de código de programação para realizar a alterações. Para continuidade deste trabalho, será realizada a implementação dos gráficos de controle avançados como o *Exponentially Weighted Moving Average* (EWMA), o *Cumulative Sum* (CUSUM), bem como índices de capacidade, e ARL.

Referências

CHANG, W.; CHENG, J.; ALLAIRE, J.J; XIE, Y.; MCPHERSON, J. 2016. **shiny: Web Application Framework for R**. R package version 0.14. Disponível em: https://CRAN.R-project.org/package=shiny. Acesso em: 13 mar. 2017.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4.ed. São Paulo: Atlas, 2010.

HENNING, E.; ALVES, C. C.; VIEIRA, V. O ambiente R como uma proposta de renovação para aprendizagem e monitoramento de processos em Controle Estatístico de Qualidade. **In:** Simpósio de Engenharia de Produção, 14. 2007, Bauru-SP. Anais... 2007.

HENNING, E.; RAMOS, M.S.; AGUIAR, R.; SANTOS, L.M.; SIPLE, I.Z. Para Além da Computação Estatística: o Uso do Ambiente R para o Ensino de Métodos Numéricos. **RENOTE**, v. 14, n. 1, p. 1 – 10, 2016.

KONRATH, A. C.; WALTER, O. M. F. C.; ALVES, C. C. ; HENNING, E.; SAMOHYL, R. W. Applications in teaching Statistical Quality Control with different R interfaces. **In**: IEEE Global Engineering Education Conference (EDUCON), 2013, Berlim. Anais…Berlim: 2013, p. 146-155.

MONTGOMERY, D. C**. Introdução ao Controle Estatístico de Qualidade**. 4ª ed. Rio de Janeiro: LTC, 2009.

MÜLLER, T. J.; ZABALA, F. J. Avaliação e Correção Automática no Software Livre RStudio. **RENOTE**, v. 14, n. 1., p. 1-10, 2016.

R CORE TEAM. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. Disponível em: <http://www.R-project.org/>. Acesso em: 13 Mar. 2017.

SAMOHYL, R. W. **Controle Estatístico de Qualidade**. Rio de Janeiro: Elsevier, 2009.

SCRUCCA, L. Qcc: An R Package for Quality Control Charting and Statistical Process Control. **R News**, v. 4, n. 1, p. 11-17, 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

DISTÂNCIA EUCLIDIANA COMO FERRAMENTA NA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA DE VARIEDADES DE GÉRBERAS

Tarcisio Rangel do Couto[60]
João Sebastião de Paula Araújo[61]
Leandro Miranda de Almeida[62]
Pedro Corrêa Damasceno Junior[63]

Resumo

A gérbera é uma das mais importantes flores ornamentais comercializadas em todo o mundo. Informações sobre a variabilidade genética desta espécie representa um importante recurso para programas de melhoramento que contam com a seleção de genótipos promissores. Objetivou-se estimar a divergência genética entre 12 genótipos de gérbera utilizando a distância Euclidiana para avaliar os descritores morfo-agronômicos quantitativos. Os genótipos foram caracterizados mediante as normas para a execução dos ensaios de distinguibilidade, homogeinidade e estabilidade (DHE) do Ministério da Agricultura (MAPA) para cultivares de Gérbera. Os agrupamentos dos genótipos foram obtidos pelos métodos de UPGMA, Ward, Vizinho Mais Próximo e Vizinho Mais Distante. A validação dos agrupamentos foi determinada pelo Coeficiente de Correlação Cofenético. Os dados foram analisados pelo programa R. Os genótipos mais distantes geneticamente, com uma distância de 209.58681 foram G2 e G4 e os genótipos mais próximos foram G2 e G7 (distância de 19.11917). A distância Euclidiana mostrou-se uma eficiente ferramenta na validação de divergência genética entre genótipos de gérbera, corroborando com descritores morfo-agronômicos quantitativos.
Palavras-Chave: Dissimilaridade, análise multivariada, recursos genéticos vegetais.

Abstract

Gerbera is one of the most important ornamental flowers marketed worldwide. Information on the genetic variability this species represents important resource for breeding programs that rely selection of promising genotypes. The Objective was to estimate the genetic divergence among 12 genotypes of gerbera using the Euclidian distance to evaluate the quantitative morpho-agronomic descriptors. The genotypes were characterized by norms for performance of the distinguishing, homogeneity and stability of Ministry of Agriculture (MAPA) assays for gerbera cultivars. Genotype clusters were obtained by UPGMA, Ward, Single and Complete methods. The validation of the clusters was determined by Cofenetic Coefficient Correlation. Data were analyzed by program R. The most genetically distant genotypes with a Euclidean distance 209.58681 were G2 and G4 and the closest genotypes were G2 and G7 (Distance 19.11917). The Euclidian distance proved to be an efficient tool in the validation of genetic divergence among genotypes of gerbera, corroborating with quantitative morpho-agronomic descriptors.
Keywords: Dissimilarity, multivariate analysis, plant genetic resources.

[60] Universidade Federal Rural do Rio de Janeiro - UFRRJ e Universidade Federal Fluminense - UFF, E-mail: tarcisiorcouto@yahoo.com.br;
[61] Universidade Federal Rural do Rio de Janeiro - UFRRJ, E-mail: araujoft@ufrrj.br
[62] Universidade Federal Rural do Rio de Janeiro - UFRRJ, E-mail: leandromirandaufrrj@gmail.com
[63] Universidade Federal Rural do Rio de Janeiro - UFRRJ, E-mail: damascenojunior2009@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 gérbera é planta perene, herbácea, da Família Asteraceae. O gênero é composto por cerca de quarenta espécies, mas apenas uma espécie (*Gerbera jamesonii*) é atualmente cultivada comercialmente. As cultivares mais utilizadas são resultantes da hibridação entre a *Gerbera jamesonii* x *Gerbera viridifolia*. O híbrido resultante do cruzamento é conhecido por *Gerbera hybrida*, com grande variabilidade de flor, forma e cores (KLOSS et al., 2005).

A divergência genética tem sido avaliada com o objetivo de identificar combinações híbridas de maior efeito heterótico e maior heterozigose, de tal sorte que, em suas gerações segregantes, tenha-se maior possibilidade de recuperação de genótipos superiores (AMARAL JÚNIOR et al., 2010).

Por outro lado, a distância Euclidiana é, sem dúvida, uma das medidas mais utilizada para a análise de agrupamentos. Considerando o caso mais simples, no qual existem “n” indivíduos, onde cada um dos quais possuem valores para “p” variáveis, a distância Euclidiana entre eles é obtida mediante o Teorema de Pitágoras, para um espaço multidimensional, sendo equivalente ao comprimento da hipotenusa do triângulo retângulo projetado (CRUZ et al., 2012).

Objetivo

Objetivou-se estimar a divergência genética entre 12 genótipos de gérbera (*Gerbera* sp.), utilizando a distância Euclidiana para avaliar os descritores morfo-agronômicos quantitativos.

Material e Métodos

Foram utilizadas plantas matrizes de 12 genótipos de gérbera (*Gerbera* sp.) sendo cinco de origem conhecida: ‘Kozak’, ‘Orca’, ‘Pacific’, ‘Lionela’, ‘Igor’ e sete de origem não conhecida: G1, G2, G3, G4, G5, G6 e G7. Amostras iniciais foram obtidas através da parceria com produtores rurais da região serrana do Rio de Janeiro, trazidas e mantidas no Setor de Horticultura do Departamento de Fitotecnia da Universidade Federal Rural do Rio de Janeiro – UFRRJ, Seropédica/RJ. O delineamento experimental empregado foi inteiramente ao acaso, com 12 tratamentos (genótipos) e cinco repetições de cada genótipo. As avaliações foram feitas de outubro a dezembro de 2016.

Os genótipos foram caracterizados mediante as normas para a execução dos ensaios de distinguibilidade, homogeinidade e estabilidade (DHE) para cultivares de Gérbera, propostos pelo Serviço Nacional de Proteção de Cultivares (SNPC), vinculado ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA). Foram avaliados oito características quantitativas, mensuradas em milímetros: a) comprimento da folha, b) largura da folha, c) comprimento do pedúnculo, d) diâmetro do capítulo, e) largura da flor ligulada do raio interno, f) comprimento da flor ligulada do raio externo, g) largura da flor ligulada do raio externo e h) diâmetro do disco.

Para a obtenção do dendrograma de agrupamento, foi utilizada a distância Euclidiana como medida de dissimilaridade para análise do dados. Seja X_{ij} a observação referente a *j*-ésima característica (*j* = 1, 2,..., n) no *i*-ésimo genótipo (*i* = 1, 2,..., p), define-se, segundo Cruz et al. (2012), a distância Euclidiana entre dois genótipos *i* e *i*' por meio da expressão:

$$d_{ii'} = \sqrt{\sum_{j} (x_{ij} - x_{i'j})^2}$$

Os agrupamentos dos genótipos foram obtidos pelos métodos de UPGMA (*Unweighted Pair-Group Method Using an Arithmetic Average*), Ward, Vizinho Mais Próximo (VMP) e Vizinho Mais Distante (VMD). A validação dos agrupamentos foi determinada pelo Coeficiente de Correlação Cofenético (CCC) (CRUZ et al., 2014). Os dados foram analisados pelo programa R (http:// www.r-project.org/).

Resultados e Discussão

De acordo com a visualização gráfica, confirmado nas medidas de dissimilaridade e com base na distância Euclidiana, foram identificados os genótipos mais similares e os mais dissimilares. Os genótipos mais distantes geneticamente, com uma distância de 209.58681 foram G2 e G4 e os genótipos mais próximos foram G2 e G7 (distância de 19.11917).

A partir da matriz de dissimilaridade realizou-se o agrupamento dos genótipos através do método hierárquico UPGMA. A distorção produzida no processo de agrupamento é estimada por um CCC proposto por Sokal e Rohlf (1962). O CCC é um produto momento que quantifica a concordância entre os valores originais da matriz de dissimilaridade e os elementos da matriz cofenética (CRUZ et al., 2014). O agrupamento hierárquico UPGMA obteve maior valor para CCC (0,78) que aqueles

verificados para os métodos de agrupamento utilizando-se Ward (0,64), VMP (0,69) e VMD (0,67). Sendo que, Sokal e Rohlf (1962) consideram o ajuste do coeficiente de correlação cofenético bom, quando o mesmo apresenta valores maiores ou igual a 0,8.

Conclusão

A distância Euclidiana mostrou-se uma eficiente ferramenta na validação de divergência genética entre genótipos de gérbera, corroborando com descritores morfo-agronômicos quantitativos.

Referências

AMARAL JÚNIOR, A.T.; VIANA, A.P.; GONÇALVES, L.S.A.; BARBOSA, C.D. Procedimentos multivariados em recursos genéticos vegetais. In: PEREIRA, T.N.S. (Org.). **Germoplasma: conservação, manejo e uso no melhoramento de plantas**. 1ª Edição. Viçosa: Arka, 2010. p.205-254.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. v.1. 4 ed. Viçosa: UFV, 2012. 514p.

CRUZ, C.D.; CARNEIRO, P.C.S.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. v.2. 3 ed. Viçosa: UFV, 2014. 668p.

KLOSS, W.E.; GEORGE, C.G.; SORGE, L.K. Dark disk color in the flower of Gerbera hybrida is determined by a dominant gene, Dc. **HortScience**, Virginia, v.40, p.1992-1994, 2005.

SOKAL, R.R.; ROHLF, F.J. The comparison of dendograms by objetive methods. **Taxonomy**, Berlin, v.11, n.1, p.33-40, 1962.

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

CULTIVO DA AVEIA PRETA E AGREGAÇÃO DO SOLO EM ÁREAS DE AGRICULTURA DE MONTANHA EM NOVA FRIBURGO, RJ

Sandra Santana de Lima[64]
Eduardo de Carvalho Silva Neto [65]
Adriana Maria de Aquino[66]
Marcos Gervasio Pereira[67]

Resumo

O uso da aveia preta (*Avena strigosa* Schreber.) como planta de cobertura tem sido muito utilizado pelos diversos benefícios que esse manejo promove ao sistema. O objetivo deste estudo foi avaliar a influência do uso da aveia preta na estrutura do solo pela análise da formação de agregados biogênicos e fisiogênicos no programa R. O estudo foi realizado no Município de Nova Friburgo-RJ em áreas de encosta em recuperação, com aveia preta adubada e não adubada, e uma área com floresta nativa. Foram coletados torrões de solo e submetidos a peneiramento, por peneiras de 9,7 e 8,0 mm. Os agregados foram identificados a partir de 200 g de agregados contidos no intervalo de 9,7 a 8,0 mm de diâmetro, e assim quantificadas as frações de agregados fisiogênicos e biogênicos contidas na massa inicial. Na análise dos dados, foram testadas as premissas da análise de variância e as médias comparadas pelo teste T de Bonferroni ao nível de significância de 5%, utilizando o programa R. A porcentagem de massa dos agregados biogênicos e fisiogênicos indica que o cultivo de aveia preta está favorecendo a estrutura do solo. O programa R foi suficiente para realizar todas as análises necessárias.
Palavras-Chave: Plantas de cobertura, agregados biogênicos, conservação do solo.

Abstract

The use of black oats (*Avena strigosa* Schreber) as a cover crop has been widely used for the many benefits that this management promotes to the system. The objective of this study was to evaluate the influence of the use of black oats on soil structure by the analysis of the formation of biogenic and physiogenic aggregates in the R program. The study was carried out in Nova Friburgo-RJ, Brazil, in sloping areas with black oats fertilized and not fertilized and an area with native forest. Soils were collected and submitted to sieving, by sieves of 9.7 and 8.0 mm. The aggregates were identified from 200 g of aggregates contained in the range of 9.7 to 8.0 mm in diameter, and thus quantified the physiogenic and biogenic aggregate fractions contained in the initial mass. In the analysis of the data, the premises of the analysis of variance and the means were compared by the test of Bonferroni at the level of significance of 5%, using the program R. The percentage of mass of the biogenic and physiogenic aggregates indicates that the cultivation of black oats is Favoring the structure of the soil. The R program was sufficient to perform all the necessary analyzes.
Keywords: Cover plants, biogenic aggregates, soil conservation.

[64] Universidade Federal Rural do Rio de Janeiro; E-mail: sandra.biologa@hotmail.com
[65] Universidade Federal Rural do Rio de Janeiro; E-mail: netosceduardo@yahoo.com.br
[66] Embrapa Agrobiologia – NPTO; E-mail: adriana.aquino@embrapa.br
[67] Universidade Federal Rural do Rio de Janeiro; E-mail: gervasio@ufrrj.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 plantas de cobertura tem sido amplamente difundido como uma importante prática conservacionista para o manejo do solo. A aveia preta (*Avena strigosa* Schreber) é uma das principais plantas utilizadas pelos produtores com a finalidade de formação de cobertura do solo (Melo et al., 2011) com grande aporte de matéria orgânica. A influência da matéria orgânica na agregação do solo é um processo dinâmico, sendo necessário o acréscimo contínuo de material orgânico para a manutenção da estrutura adequada para o desenvolvimento das plantas (Cunha et al., 2011). A estrutura do solo é um fator chave no funcionamento do solo, sua capacidade de suportar a vida vegetal e animal, e qualidade ambiental (Tavares Filho et al., 2012). Além disso, tem estreita relação com os agregados, que são definidos por Bochner et al. (2008) como um conjunto de partículas do solo com forma e tamanho definidos, constituindo uma unidade estrutural. Os agregados, por sua vez, podem ser classificados de acordo com suas vias de formação de acordo com Velasquez et al. (2007), sendo biogênicos os agregados de origem biológica e fisiogênicos os formados a partir dos processos fisicos e químicos do solo.

Objetivo

Avaliar a influência do uso da aveia preta na estrutura do solo pela análise da formação de agregados biogênicos e fisiogênicos no programa R.

Material e Métodos

O estudo foi realizado no Município de Nova Friburgo-RJ em áreas de encosta em recuperação, com dois sistemas de produção conservacionistas com aveia-preta (*Avena strigosa* Schreber.) adubada (AV-A) e não adubada (AV-N), além de uma área com floresta nativa. Foram coletados torrões de solo na profundidade de 0-10 cm e submetidos a peneiramento, por um conjunto de peneiras de 9,7 e 8,0 mm. A identificação dos agregados, segundo suas vias de formação, foi feita com auxílio de lupa, a partir de 200 g de agregados contidos no intervalo de 9,7 a 8,0 mm de diâmetro, e assim quantificadas as frações de agregados fisiogênicos e biogênicos contidas na massa inicial. Para a análise dos dados, considerou-se o modelo estatístico para o esquema de parcelas subdivididas, no delineamento inteiramente casualizado. Inicialmente, os dados foram submetidos à análise de variância. Foram investigadas as premissas fundamentais da análise de variância, a normalidade e

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

homocedasticidade da variância por meio dos testes de Shapiro-Wilk e Bartlett, respectivamente. As variáveis que não atenderam as premissas foram transformadas pelo método de Box & Cox (1964). As médias dos fatores foram comparadas pelo teste T de Bonferroni ao nível de significância de 5%, utilizando o programa estatístico R, versão 3.1.1 (R Development Core Team 2015) e pacote *ExpDes.pt*. (Ferreira et al., 2013).

Resultados e Discussão

A análise dos valores médios de porcentagem da massa de cada via de formação dos agregados (biogênico e fisiogênico) revela que houve desdobramento com interação a 5% de significância (Tabela 1). A comparação entre as áreas mostra o maior valor médio de agregados biogênicos na área com aveia adubada (AV-A). Por outro lado, o maior valor de agregados fisiogênicos foi observado na área com aveia não adubada (AV-N). Quanto à análise comparativa das vias de formação dentro de cada área, observou-se que na área de AV-N a massa de agregados fisiogênicos foi maior (P>0,05), contrariamente a área de AV-A, que apresentou maior valor de agregados biogênicos. Quando se considera a floresta como referência, verifica-se que os valores observados na área AV-A foram praticamente iguais, tanto para os agregados biogênicos, como para os fisiogênicos.

Tabela 1. Porcentagem da massa (200g) de agregados em áreas de cultivo e floreta, considerando as vias de formação dos agregados.

ÁREAS	**Massa** ---------- **%** ----------	
	Biogênico	Fisiogênico
AV-N	47,34 bB	52,66 aA
AV-A	51,70 aA	48,30 bB
FL	51,67	48,34
CV% Áreas	2,72	
CV% Classes	0,02	

Médias seguidas de mesma letra maiúscula, na mesma linha, não diferem entre si; médias seguidas de mesma letra minúscula, na mesma coluna, não diferem entre si pelo teste de Bonferroni (P<0,05). AV-N: aveia preta não adubada. AV-A: aveia preta adubada. FL: Floresta. CV:coeficiente de variação.

Esses resultados podem estar relacionados à maior densidade e melhor distribuição do sistema radicular da aveia preta, favorecendo a aproximação das

partículas minerais, contribuindo para a formação e a estabilização dos agregados (Silva & Mielniczuk, 1997).

Conclusão

A porcentagem de massa dos agregados biogênicos e fisiogênicos indica que o cultivo de aveia preta está favorecendo a estrutura do solo. O programa R favoreceu a análise dos dados por apresentar várias ferramentas imprescindíveis para análise dos dados.

Referências

BOCHNER, J. K.; FERNANDES, M. F.; PEREIRA, M. G.; BALIEIRO, F. C.; SANTANA, I. K. S. Matéria orgânica e agregação de um Planossolo sob diferentes, coberturas florestais Cerne, Lavras, v. 14, n. 1, p. 46-53, 2008.

BOX, G. E. P.; COX, D. R. Na analyis of transformations. Journal of the Rayal Statistical society, London, 26(2): 42-56. 1964.

CUNHA, E. Q.; STONE, L. F.; MOREIRA, J. A. A.; FERREIRA, E. P. B.; AGOSTINHO DIRCEU DIDONET, A. D.; LEANDRO, W. M. Sistemas de preparo do solo e culturas de cobertura na produção orgânica de feijão e milho. R. Bras. Ci. Solo, 35:589-602, 2011.

FERREIRA, E.B.: CAVALCANTI, P.P.: NOGUEIRA, D.A. ExpDes.pt: Experimental Designs pacakge (Portuguese). R package version 1.1.2. 2013.

MELO, A. V.; GALVÃO, J. C. C.; BRAUN, H.; SANTOS, M. M.; COIMBRA, R. R.; SILVA, R. R.; REIS, W. F. Extração de nutrientes e produção de biomassa de aveia-preta cultivada em solo submetido a dezoito anos de adubação orgânica e mineral. Semina: Ciências Agrárias, Londrina, 32(2):411-420, 2011.

SILVA, I.F.& MIELNICZUK, J. Ação do sistema radicular de plantas na formação e estabilização de agregados do solo. R. Bras. Ci. Solo, 21:113-117, 1997.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015.

TAVARES FILHO, J.; FELTRAN, C. T. M.; OLIVEIRA, J. F.; ALMEIDA, E.; GUIMARÃES, M. F. Atributos de solo determinantes para a estimativa do índice de estabilidade de agregados. Pesq. agropec. bras., 47(3):436-441, 2012.

VELASQUEZ, E.; PELOSI, C.; BRUNET, D.; GRIMALDI, M.; MARTINS, M.; RENDEIRO, A. C.; BARRIOS, E.; LAVELLE, P. This ped is my ped: Visual separation and near infrared spectra allow determination of the origins of soil macroaggregates. Pedobiologia. 51:75-87, 2007.

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

MATÉRIA ORGÂNICA EM NINHOS DE TÉRMITAS SOB AMBIENTES DE MAR DE MORROS NA REGIÃO SUDESTE DO RIO DE JANEIRO

Sandra Santana de Lima[68]
Renato Nunes Pereira [69]
Marcos Gervasio Pereira[70]

Resumo

Áreas densamente ocupadas por ninhos de térmitas são comumente obsevadas em diferentes biomas. A espécie construtora influencia na composição orgânica dos termiteiros. Nesse sentido, objetivou-se avaliar os teores de matéria orgânica nas diferentes partes do termiteiro e solo do entorno, utilizando o programa R. Foram selecionadas duas topossequências com termiteiros, sendo estas divididas em terço superior, médio e inferior, em cada uma das secções amostrados 4 termiteiros (subdividido em topo, centro e base) e na área do entorno coletadas amostras de terra nas profundidades de 0-5 cm, nas distâncias de 50 e 150 cm da base dos termiteiros. A caracterização da matéria orgânica através do fracionamento granulométrico. Na análise dos dados, foram testadas as premissas da análise de variância. As variáveis que não atenderam aos pressupostos foram transformadas. Posteriormente as médias foram comparadas pelo teste Scott-Knott ao nível de significância de 5%. Os termiteiros construídos por *Cornitermes c.* apresentam em suas bases maiores teores de carbono orgânico total, assim como o carbono particulado e o carbono associado as frações minerais. O uso do programa R, bem como o pacote utilizado, foi de grande relevância por possibilitar todas as análises necessárias e favorecer melhor interpretação dos resultados.
Palavras-Chave: Cupinzeiros, fracionamento granulométrico, conservação do solo.

Abstract

Areas that are densely occupied by termites mounds are commonly observed in different biomes. The constructor species influences the organic composition of the mounds. In this sense, the objective was to evaluate the organic matter contents in the different parts of the mounds and surrounding soil, using the R program. Two toposequences were selected with mounds, which were divided into upper, middle and lower third, in each of the four sampled sections (subdivided into top, center and base) and in the surrounding area. Depths of 0-5 cm in the distances of 50 and 150 cm of the base of the mounds. A characterization of the organic matter was accomplished by the granulometric fractionation. In the analysis of the data, the assumptions of the analysis of variance were tested. The variables that did not meet the assumptions were transformed. Subsequently the averages were compared by the Scott-Knott test at a significance level of 5%. The termites constructed by *Cornitermes c.* Present in their bases higher levels of total organic carbon, as well as the particulate carbon and the carbon associated with the mineral fractions. The use of the R program, as well as the package used, was of great relevance because it allows all the necessary analyzes and favors a better interpretation results.
Keywords: Mounds, granulometric fractionation, soil conservation.

[68] Universidade Federal Rural do Rio de Janeiro; E-mail: sandra.biologa@hotmail.com
[69] Universidade Federal Rural do Rio de Janeiro; E-mail:rnpmoc@gmail.com
[70] Universidade Federal Rural do Rio de Janeiro; E-mail:gervasio@ufrrj.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 ocorrência de térmitas ou cupins em uma determinada área ou paisagem pode ser observada através da presença de suas construções de ninhos epígeos conhecidos como termiteiros, cupinzeiros ou montículos. No Brasil é comum observar grandes áreas de pastagens densamente ocupadas com ninhos de térmitas. Essa ocupação resulta na depreciação da propriedade pelos agricultores, pelo fato de ser atribuída a quantidade de termiteiros com um indicativo da degradação do solo, ou mesmo a processos como diminuição da fertilidade, em especial ao aumento da acidez do solo, padrões que não foram observados por Lima et al. (2011). Apesar da presença dos térmitas não seja bem vista nos ecossistemas, esse grupo tem importante função no solo, ao realizar suas atividades na construção dos ninhos, o que lhes confere a denominação de "engenheiros do solo". Esses promovem modificações na estrutura do solo com a criação de estruturas biogênicas que podem alterar os atributos físicos dos solos. Também podem atuar na decomposição da matéria orgânica e disponibilizar, pela fragmentação da serapilheira, nutrientes aos demais organismos (Lavelle et al., 1997; Jouquet et al., 2006). Adicionalmente promovem impactos sobre os atributos químicos solo. A espécie construtora do termiteiro influencia no material orgânico presente no solo do termiteiro (Rückamp et al., 2009). A espécie *Cornitermes cumulans* é comumente observada em áreas de pastagens e apresenta em seu interior uma parte central com maior concentração de material de origem vegetal (Sanchez et al., 1989).

Objetivo

Avaliar os teores de matéria orgânica nas diferentes partes do termiteiro e solo do entorno, utilizando o programa R.

Material e Métodos

As áreas de estudo estão localizadas no município de Pinheiral, na sub-bacia do ribeirão Cachimbal, na região do Médio Paraíba Fluminense sob o relevo conhecido como "Mar de Morros". As áreas apresentam sinais de degradação, e ocorrência de termiteiros. Foram selecionadas duas topossequências caracterizadas pela ocorrência de termiteiros, sendo estas divididas em terço superior, médio e inferior, em cada uma das secções pré-estabelecidas, foram amostrados 4 termiteiros (subdividido em topo, centro e base) e na área do entorno coletadas amostras de terra

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

nas profundidades de 0-5 cm, nas distâncias de 50 e 150 cm da base dos termiteiros. As amostras de terra, bem como as provenientes dos ninhos, foram preparadas, sendo secas ao ar, para a caracterização da matéria orgânica através do fracionamento granulométrico. Na análise dos dados considerou-se o modelo estatístico para o esquema de parcelas subdivididas, no delineamento blocos casualizados, em que, as topossequências foram consideradas como blocos. Os pontos de amostragens foram denominados solo, o que corresponde às partes dos ninhos epígeos (topo, centro e base) e o solo adjacente, nas distâncias de 50 e 150 cm da base. Na análise dos dados, inicialmente, foi realizado a análise de variância. Foram investigadas, a normalidade e homocedasticidade da variância por meio dos testes de Shapiro-Wilk & Bartlett, respectivamente. As variáveis que não atenderam aos pressupostos foram transformados pelo método de Box & Cox (1964). As médias dos fatores foram comparadas pelo teste Scott-Knott ao nível de significância de 5%, por ser robusto e eficiente por separar as média em grupos distintos, utilizando o programa estatístico R, versão 3.1.1 (R Development Core Team 2015) e pacote *ExpDes.pt*. (Ferreira et al., 2013).

Resultados e Discussão

Todos os termiteiros das duas topossequências foram construídos pela espécie *Cornitermes c.*, fato que favorece a comparação e acurácia aos resultados. Os teores de carbono orgânico do solo (COT) na profundidade de 0-5 cm (Tabela 1) diferiram ($P<0,05$) entre as partes do termiteiros, sendo nessas observados os teores mais elevados. Os teores quantificados na base diferiram por apresentarem-se mais elevados (38,46 g kg^{-1}) quando comparado ao topo (20,37 g kg^{-1}) e centro (21,99 g kg^{-1}), e especialmente em comparação aos valores observados no solo, nas duas distâncias 50 cm (13,51 g kg^{-1}) 150 cm (12,88 g kg^{-1}). Quando se considerou a distribuição de COT na paisagem, não foram observadas diferença entre os terços. Quanto ao carbono orgânico particulado (COp), que está associado a fração areia, verifica-se o mesmo padrão observado para o COT, com os maiores ($P<0,05$) valores no termiteiro. Contudo para esse atributo, os teores referentes a base do termiteiro são superiores aos demais (49,32 g kg^{-1}), sendo aproximadamente seis vezes maior que o observado no topo (8,26 g kg^{-1}). Os teores no solo foram menores nas profundidades de 50 (2,54 g kg^{-1}) e 150 cm (2,24 g kg^{-1}). Os maiores valores de COp na base do termiteiro também estão relacionados a grande concentração de material

vegetal na parte interior de sua base. Quando se observa o carbono associados aos minerais (Coam) não foram verificadas diferenças entre as partes do termiteiro, embora os teores correspondentes a essas partes tenham seguido o padrão de aumento do topo (13,14 g kg^{-1}) para a base (21,97 g kg^{-1}). Contudo, os teores no ninho foram superiores ($P<0,05$) quando comparados ao solo em ambas as distâncias da base, 50 cm (1,50 g kg^{-1}) e150 cm (1,23 g kg^{-1}). A maior concentração de carbono na base do termiteiro está relacionada com a espécie construtora.

Conclusão

Os termiteiros construídos por *Cornitermes c.* apresentam em suas bases maiores teores de COT, assim como o COP e Coam. O uso do programa R, bem como o pacote utilizado, foram de grande relevância por possibilitar todas as análises necessárias e favorecer melhor interpretação dos resultados.

Referências

BOX, G.E.P.; COX, D.R. Na analyis of transformations. Journal of the Rayal Statistical society, London, 26(2): 42-56. 1964.

FERREIRA, E.B.: CAVALCANTI, P.P.: NOGUEIRA, D.A. ExpDes.pt: Experimental Designs pacakge (Portuguese). R package version 1.1.2. 2013.

JOUQUET, P.; DAUBER, J.; LAGERLÖF, J. Soil invertebrates as ecosystemengineers: intended and accidental effects on soil and feedback loops. Appl Soil Ecol., 32: 153-164, 2006.

LAVELLE, P.; BIGNELL, D.; LAPAGE, M. Soil function in changing world: the role of invertebrate ecosystems engineers. Eur. J. Soil Biol., 33(4):159-193, 1997.

LIMA, S.S.; ALVES, B.J.R.; AQUINO, A.M.; MERCANTE, F.M.; PINHEIRO, É.F.M.; SANT'ANNA, S.A.C.; URQUIAGA, S.; BODDEY, R.M. Relação entre a presença de cupinzeiros e a degradação de pastagens. Pesq agropec. bras. 46(12): 1699-1706, 2011.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2015.

RÜCKAMP, D.; AMELUNG, W.; BORMA, L.S.; NAVAL, L.P.; MARTIUS, C. Carbon and nutrient leaching from termite mounds inhabited by primary and secondary termites. Applied Soil Ecology, 43: 159-162, 2009.

SANCHEZ, G.; PERES FILHO, O.; SALVADORI, J.R.; NAKANO, O. Estrutura e sistema de aeração do cupinzeiro de Cornitermes cumulans (Kollar, 1832) (Isoptera: Termitidae). Pesq. agropec. bras., 24(8): 941-943, 1989.

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

DESENVOLVIMENTO DE UM APLICATIVO EM SHINY PARA PREVISÃO DE SÉRIES TEMPORAIS

Rodrigo Gabriel de Miranda[71]
Robert Wayne Samohyl[72]
Andréa Cristina Konrath[73]
Gueibi Peres Souza[74]

Resumo

O objetivo principal deste trabalho foi apresentar a estrutura de um aplicativo para web, feito na linguagem R, para a previsão de séries temporais. Para isto foi utilizado o framework do pacote "shiny" em conjunto com outros cinco pacotes. Os pacotes utilizados foram: "flexdashboard" (construção da interface gráfica), "readxl" (leitura dos dados no formato .xls ou .xlsx), "plotly" (gráficos interativos), "DT" (tabelas) e "forecast" (previsões de séries temporais). A programação e compilação do aplicativo foi realizada utilizando a IDE (ambiente de desenvolvimento integrado) desenvolvida pela empresa rstudio. Os modelos utilizados no aplicativo foram: modelo ingênuo e ingênuo sazonal, suavização exponencial (considerando 5 tipos de tendência e 3 tipos de sazonalidade) e modelos ARIMA (não sazonal e sazonal). O resultado foi o desenvolvimento de um aplicativo que permite ao usuário de uma forma simples e intuitiva, sem nenhum conhecimento na linguagem de programação R, utilizar os modelos disponíveis no R.
Palavras-Chave: shiny, previsão, séries temporais

Abstract

The main objective of this work was to present the structure of a web application, made in the R language, for time series forecast. For this, the framework of the "shiny" package was used in conjunction with five other packages. The packages used were: "flexdashboard" (construction of the graphical interface), "readxl" (reading data in .xls or .xlsx format), plotly (interactive graphics), "DT" (tables) and "forecast" (time series forecasts). The programming and compilation of the application was performed using the IDE (integrated development environment) developed by the company rstudio. The models used in the application were: naive and naive seasonal model, exponential smoothing (considering 5 types of trend and 3 types of seasonality) and ARIMA models (non-seasonal and seasonal). The result was the development of an application that allows the user in a simple and intuitive way, without any knowledge in the programming language R, to use the models available in R.
Keywords: shiny, forecast, time series

Introdução

Atualmente no software R (R Core Team, 2016) existem muitos pacotes e modelos para a previsão de séries temporais (https://cran.r-project.org/web/views/TimeSeries.html), porém todo este conteúdo está restrito a

[71] UDESC – Universidade Estadual de Santa Catarina, rgabrieldemiranda@gmail.com
[72] UFSC – Universidade Federal de Santa Catarina, dinhopixupreju@gmail.com
[73] UFSC – Universidade Federal de Santa Catarina, andreack@gmail.com
[74] UFSC – Universidade Federal de Santa Catarina, gpssouza@yahoo.com.br

pessoas com alguma experiência na linguagem de programação R, tornando difícil a aplicação destes modelos tanto no meio acadêmico quanto no meio empresarial. Pensando nesta situação, foi desenvolvido um aplicativo para web utilizando o pacote "shiny", que permite ao usuário de uma forma simples e intuitiva, sem nenhum conhecimento na linguagem de programação R, utilizar os modelos disponíveis no R. Neste aplicativo foram considerados a implementação de três modelos, decomposição clássica, suavização exponencial (Hyndman et al., 2002) e ARIMA (Hyndman e Khandakar, 2008).

Objetivo

O objetivo deste trabalho foi apresentar a estrutura de um aplicativo para web feito na linguagem R para a previsão de séries temporais utilizando o framework do pacote "shiny".

Material e Métodos

Para construção do aplicativo web de previsão de séries temporais foram utilizados seis pacotes do R, cada um com uma função especifica listada a seguir. A base para o desenvolvimento deste aplicativo para web foi o pacote "shiny" (Chang et al., 2017), que é um framework para o desenvolvimento de aplicações em web. A principal função deste pacote é permitir a comunicação entre um servidor (que executa as funções do R) e uma interface gráfica em HTML. A interface em HTML do aplicativo foi construída utilizando o pacote "flexdashboard" (Allaire, 2016) que de maneira simples (só é necessário conhecimento de códigos de programação em R) e rápida permite elaborar vários layouts de páginas para visualização dos dados e gráficos. Para a entrada dos dados no aplicativo utilizou-se o pacote "readxl" (Wickham, 2016) que permite a leitura de arquivos no formato .xls e .xlsx (padrões do Excel). Esta escolha ocorreu devido a dificuldade que muitos usuários tem em trabalhar com arquivos .csv e estão mais habituados a utilizar Excel. Os gráficos do aplicativo foram gerados utilizando o pacote "plotly" (Sievert et al., 2016) que possuí a facilidades para gerar vários tipos de gráficos, incluindo o de séries temporais. Para apresentar os dados de forma tabular foi utilizado o pacote "DT" (Xie, 2016), que é uma interface do R para biblioteca DataTables (https://datatables.net) onde é possível construir tabelas interativas (filtros, paginação, ordenação e outras funções). Por fim, o pacote que é responsável por processar os cálculos dos modelos de séries temporais é o "forecast"

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

(Hyndman, 2016). Toda a programação foi realizada utilizando o ambiente de programação Rstudio (https://www.rstudio.com).

Resultados e Discussão

A estrutura básica de um aplicativo em shiny compreende duas partes, uma interface gráfica para o usuário (UI) que possuí os Inputs (parâmetros inseridos pelo usuário) e Outputs (resultados em gráficos e tabelas) e um servidor (server), que processa os Inputs da UI no R e devolve os resultados para mesma UI para serem exibidos nos Outputs. Esta estrutura pode ser vista na figura 1, que é uma das telas do aplicativo.

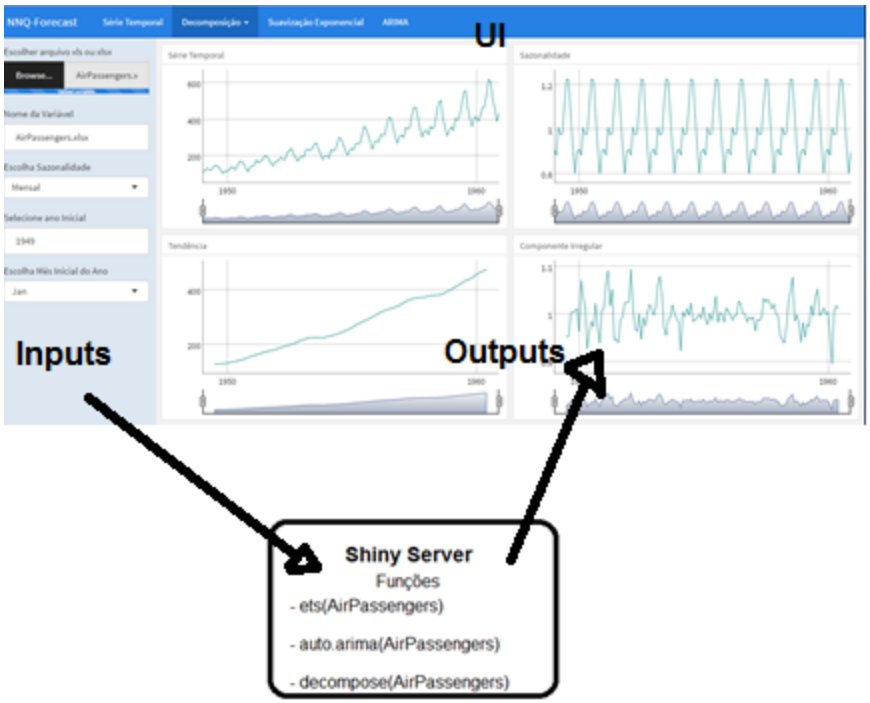


Figura 1 – Aplicativo em shiny para séries temporais

Nesta tela, por exemplo, os Inputs do usuário são: selecionar o arquivo com os dados, selecionar a periodicidade da série e a data inicial dos dados. Os Outputs são quatro gráficos, dados originais, tendência, sazonalidade e componente irregular.

Conclusão

A utilização do pacote “shiny” para o desenvolvimento de aplicativos interativos para a web pode ser uma forma de ampliar o uso do R tanto no meio acadêmico quanto no empresarial, pois possibilita a utilização de modelos estatísticos (ex. séries temporais) por usuários com pouca ou nenhuma experiência de programação na linguagem R. O pacote permite que toda a programação do aplicativo seja feita apenas

com comandos e funções do R, não requerendo conhecimento em HTML, CSS ou JavaScript.

Referências

ALLAIRE, JJ. flexdashboard: R Markdown Format for Flexible Dashboards. Rpackag e version 0.3. disponível em: https://CRAN.R-project.org/package=flexdashboard, 2016.

CHANG, Joe Cheng Winston; ALLAIRE, Yihui Xie JJ e MCPHERSON, Jonathan. shiny: Web Application Framework for R. R package version 1.0.0. disponível em: https://CRAN.R-project.org/package=shiny, 2017.

HYNDMAN, R.J. forecast: Forecasting functions for time series and linear models. R package version 7.3, disponível em: http://github.com/robjhyndman/forecast, 2016.
HYNDMAN, R.J. e KHANDAKAR, Y. Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3), 2008.

HYNDMAN, R.J.; KOEHLER, A.B.; SNYDER, R.D. e GROSE, S. A state space framework for automatic forecasting using exponential smoothing methods, *International J. Forecasting*, **18**(3), 439–454, 2002.

R Core Team. R: A language and environment for statistical computin R Foundation for Statistical Computing, Vienna, Austria. disponível em: https://www.R-project.org/, 2016.

SIEVERT, Carson; PARMER, Chris; HOCKING, Toby, CHAMBERLAIN, Scott; RAM, Karthik; CORVELLEC, Marianne e DESPOUY, Pedro (2016). plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 4.5.6. disponível em https://CRAN.R-project.org/package=plotly, 2016.

WICKHAM, Hadley. readxl: Read Excel Files. R package version 0.1.1. disponível em: https://CRAN.R-project.org/package=readxl, 2016.

XIE, Yihui. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.2. disponível em: https://CRAN.R-project.org/package=DT, 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

DOENÇAS INFECTO-PARASITÁRIAS E SUAS INTER-RELAÇÕES COM VARIÁVEIS CLIMÁTICAS, VIA ANÁLISE DE COMPONENTES PRINCIPAIS, EM NATAL-RN

Julio Cesar Barreto da Silva[75]
Carlos José Saldanha Machado[76]

Resumo

Mudanças climáticas podem produzir impactos sobre a saúde humana por diferentes vias e contribuir indiretamente com o aumento da incidência de doenças infecciosas mediado por alterações no ambiente. A Análise de Componentes Principais (ACP) tem por finalidade básica a redução, eliminação de sobreposições e escolha das formas mais representativas de dados, a partir de combinações lineares das variáveis originais pela transformação de variáveis discretas em coeficientes correlacionados. Neste estudo, usou-se a ACP especificamente para reduzir a dimensionalidade do conjunto de dez variáveis climáticas e escolher as mais representativas junto à ocorrência de sete doenças infectoparasitárias, de modo a verificar a interdependência entre tais variáveis em Natal-RN, no período de janeiro de 2001 a dezembro de 2012. Foi observada inter-relação entre a hepatite, malária (altamente correlacionadas) e sífilis; porém, tais doenças não evidenciaram inter-relação com as variáveis climatológicas estudadas. As variáveis de temperatura, bem como as calculadas a partir desta (déficit de pressão de vapor saturado, evapotranspiração potencial, índice de calor e índice de conforto humano) não se mostraram representativas para nenhuma das doenças estudadas. Embora as doenças se apresentem bastante dispersas em relação umas às outras, bem como em relação às variáveis climáticas, sugere-se o uso das seguintes variáveis em estudos futuros: sífilis, umidade relativa, índice de conforto humano e dengue.
Palavras-Chave: análise multivariada, autovetores, clima, covariância

Abstract

Climate change can have an impact on human health by different pathways, and may contribute indirectly to an increase in the incidence of infectious diseases mediated by changes in the environment. Principal Component Analysis (PCA) has the basic purpose of reducing, eliminating overlaps and choosing the most representative forms of data, from linear combinations of the original variables by transforming discrete variables into correlated coefficients. In this study, PCA was used specifically to reduce the dimensionality of the set of ten climatic variables and to choose these more representative beside seven infectious and parasitic diseases, in order, to verify the interdependence between these variables in Natal-RN on period from January 2001 to December 2012. Intercorrelations between hepatitis, malaria (these highly correlated) and syphilis were observed; however, these diseases did not show an interrelation with the studied climatological variables. The temperature variables, as well as those calculated from this one (saturated vapor pressure deficit, potential evapotranspiration, heat index and human comfort index) were not representative for any of the diseases studied. Although the diseases are quite dispersed in relation to each other, as well as in relation to climatic variables, we suggest the use of the following variables in future studies: syphilis, relative humidity, human comfort index and dengue.
Keywords: multivariate analysis, eigenvectors, climate, covariance

[75] Programa de Pós-Graduação em Meio Ambiente. Universidade do Estado do Rio de Janeiro, barretojcs@gmail.com
[76] Fundação Oswaldo Cruz; saldanhamachado@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

Mudanças climáticas podem produzir impactos sobre a saúde humana por diferentes vias. Por um lado, impacta de forma direta, como no caso das ondas de calor, ou mortes causadas por outros eventos extremos como furacões e inundações. Mas muitas vezes, esse impacto é indireto, sendo mediado por alterações no ambiente como a alteração de ecossistemas e de ciclos biogeoquímicos, que podem aumentar a incidência de doenças infecciosas (BARCELLOS et al., 2009).

Proposto por Pearson (1901), a análise de componentes principais - ACP ou PCA, do inglês *Principal Component Analysis* - tem por finalidade básica a redução, eliminação de sobreposições e a escolha das formas mais representativas de dados, a partir de combinações lineares das variáveis originais, a partir da transformação de variáveis discretas em coeficientes correlacionados.

Objetivo

Neste estudo usou-se a ACP para reduzir a dimensionalidade do conjunto de variáveis climáticas e escolher as mais representativas junto à ocorrência de doenças infecto-parasitárias, de modo a verificar a interdependência entre tais variáveis junto à cidade de Natal-RN.

Material e Métodos

A ACP foi aplicada com a utilização do programa gratuito R-*Project* para tal estudo, no período de janeiro de 2001 a dezembro de 2012, sobre dados médios mensais de 17 variáveis, sendo: 1) casos notificados de sete doenças infecto-parasitárias: dengue (Den), esquistossomose (Esq), hepatite (Hep), leishmaniose visceral (Lei), malária (Mal), meningite (Men) e sífilis congênita (Sif) - coletados junto ao Departamento de Informática do Sistema Único de Saúde, via SINAN-net (BRASIL, 2016); e 2) dados climatológicos: precipitação (PRP); temperaturas média (T-med), máxima (T_max) e mínima (T_min); e umidade relativa (UR)– coletados do Instituto Nacional de Meteorologia; bem como, outros calculados a partir destes coletados: evapotranspiração potencial (ETP), segundo Allen et al. (1998); evapotranspiração de referência (ETO), propostos por Thornthwaite (1948); déficit de pressão de vapor saturado (SVPD) e índice de calor (IC), propostos por Steadman (1979); e índice de conforto humano (ICH), segundo Rosenberg (1983).

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 Alencar (2009), a ACP é um dos métodos estatísticos usados para analisar as inter-relações entre múltiplas variáveis de forma a condensar a informação contida nelas em um conjunto menor de variáveis estatísticas, observando uma perda mínima de informação. Segundo Mingoti (2005), os componentes principais são descritos de uma forma geral, por um conjunto de p variáveis $X_1, X_2, \ldots, X_P$ com médias $\mu_1, \mu_2, \ldots, \mu_P$ e variâncias $\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_P$, respectivamente. Tais variáveis não são independentes e, portanto, possuem covariância entre a *i-ésima* e *k-ésima* variável definida por σ_{ik}, para $i \neq k, k = 1, 2, \ldots, p$. Logo, as p variáveis podem ser expressas na forma vetorial por: $X = [X_1, X_2, \ldots, X_P]'$, com vetor de médias $\mu = [\mu_1, \mu_2, \ldots, \mu_P]'$, e matriz de covariância Σ. Encontram-se os pares de autovalores e autovetores $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_P, e_P)$ em que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P$, associados à Σ. *Logo,* o *i-ésimo* componente principal é definido por $Z_i = e_i'X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p$, sendo $i = 1, 2, \ldots, p$. A variável que apresentar o maior coeficiente (valor absoluto) no componente principal (CP) de menor autovalor (menor variância) deve ser menos importante para explicar a variância total; logo, deve ser descartada (Regazzi, 2002). Neste processo de descarte, considera-se o autovetor (coeficientes do CP) correspondente ao menor autovalor, e rejeita-se a variável associada ao maior coeficiente. Desta forma, o próximo menor autovetor é avaliado, continuamente, até que o autovetor associado ao autovalor inferior a 0,7 seja considerado (Jolliffe, 1973).

Resultados e Discussão

A Tabela 1 apresenta a variância explicada e cumulativa distribuída ao longo das CP. As componentes 1 e 2 dominam juntas 56,79% da variância total dos dados padronizados. A Figura 1 mostra a dispersão de tratamentos, originadas a partir da ACP entre as doenças infecto-parasitárias e as variáveis climáticas. Algumas doenças ficaram posicionadas bem próximas umas das outras formando um pequeno ângulo entre as setas representativas dos atributos, demonstrando correlação positiva entre tais, o que foi observado entre a hepatite, malária (altamente correlacionadas, vide setas sobrepostas) e sífilis; no entanto, tais doenças não evidenciaram correlação junto às demais variáveis. Quanto às variáveis climáticas, a precipitação e umidade relativa se mostraram altamente correlacionadas à meningite, bem como, determinada inter-relação à dengue, leishmaniose e esquistossomose, nesta ordem decrescente de correlação. Já a ETP, o IC, o ICH e a SVPD, calculadas em função da temperatura média, apresentaram-se altamente correlacionadas a esta variável, o que foi também

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

conjuntamente observado em relação à temperatura máxima. Tais variáveis apresentaram certa inter-relação junto à temperatura mínima.

Tabela 1. Variância explicada e cumulativa das componentes principais (CP)

CP	Variância explicada (%)	Variância cumulativa (%)
1	46,570	46,570
2	10,220	56,790
3	8,393	65,186
4	7,030	72,220
5	6,647	78,863
6	5,774	84,637
7	4,334	88,971
8	3,624	92,595
9	3,122	95,717
10	1,972	97,689
11	1,238	98,927
12	0,921	99,848
13	0,135	99,983
14	0,007	99,990
15	0,005	99,995
16	0,003	99,999
17	0,001	100,000

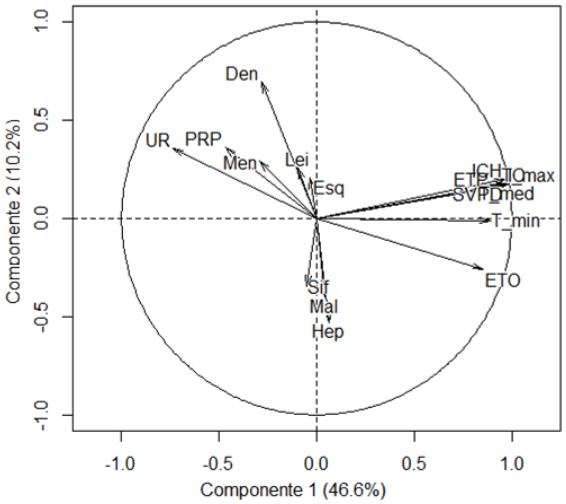


Figura 1. Pesos de PC1 e PC2 para as 17 variáveis estudadas.

As variáveis sugeridas como passíveis de descarte são apresentadas a seguir, a partir do último componente principal (CP-17) até a CP-05, inclusive, pela ordem de menor importância para explicar a variação total: déficit de pressão de vapor saturado, temperatura média, evapotranspiração de referência, índice de calor, temperatura máxima, evapotranspiração potencial, temperatura mínima, precipitação, meningite, leishmaniose visceral, malária, hepatite e esquistossomose.

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

Este estudo serviu para evidenciar que, embora as doenças se apresentem bastante dispersas umas em relação às outras, bem como em relação às variáveis climáticas; recomendam-se as seguintes variáveis para serem mantidas em estudos futuros: sífilis, umidade relativa, índice de conforto humano e dengue, uma vez que explicam a variância total, nesta ordem de importância.

Referências

Alencar, B. J. (2009). *A análise multivariada aplicada ao tratamento da informação espacial: uma abordagem matemático-computacional em análise de agrupamentos e análise de componentes principais.* Tese de Doutorado. Pontifícia Universidade Católica. Belo Horizonte – MG, Brasil.

Allen, R. G.; Pereira, L. S.; Raes, D.; Smith, M. (1998). *Crop Evapotranspiration (guidelines for computing crop water requirements).* Fao Irrigation and Drainege, 56, 297p.

Barcellos, C.; Monteiro, A. M. V.; Corvalán, C.; Gurgel, H. C.; Carvalho, M. S.; Artaxo, P.; et al. (2009). Febre amarela: reflexões sobre a doença, as perspectivas para o século XXI e o risco da reurbanização. *Epidemiol. Serv. Saúde*, 18 (3): 285-304.

Brasil (2017). Ministério da Saúde. Departamento de Informática do Sistema Único de Saúde – DATASUS. Disponível em: <*http://www2.datasus.gov.br/DATASUS/*>. Acesso em 06 março 2017.

Brasil (2017). Ministério da Agricultura, Pecuária e Desenvolvimento. Instituto Nacional de Meteorologia – INMET. Disponível em: <*http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep*>. Acesso em 06 março 2017.

Jolliffe, I.T. (1973). Discarding variables in a principal componente analysis II. Real data. *Aplied Statistics*, 22: 21-31.

Mingoti, S. A. (2005). *Análise de dados através de Métodos de Estatística Multivariada:* Belo Horizonte, MG. Editora UFMG.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 2 (6): 559–572.

Regazzi, A.J. (2002). *Análise multivariada*. Viçosa, MG: Universidade Federal de Viçosa. (INF-766 - notas de aula).

Rosenberg, N. J.; Bland, B. L.; Verma, S. B. (1983). *Microclimate: The Biological Environment.* New York: Jonh Wiley & Sons, 467p.

Steadman, R. G. (1979). The assessment of sultriness: part I: A temperature-humidity index based on human physiology and clothing science. *J. Appl. Meteor*., 18: 861-884.

Thornthwaite, W. C. (1948). An approach toward a rational classification of climate. *Geographycal Review*, 38 (1): 55-94.

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

OCORRÊNCIA DE CASOS DE FEBRE TIFOIDE NO ESTADO DO PARÁ E SUA ASSOCIAÇÃO COM ALGUMAS CARACTERÍSTICAS DA DOENÇA: UMA ANÁLISE USANDO MODELO DE REGRESSÃO LOGÍSTICA BINÁRIA MÚLTIPLA

José Ailton Nunes de Lima[77]
Sibelle Cristine Nascimento Vilhena[78]
Adrilayne dos Reis Araújo[79]
José Gracildo de Carvalho Júnior[80]

Resumo

A Febre Tifoide é uma doença infecciosa causada pela bactéria *Salmonella Typhi* que é disseminada pela água e alimentos mal tratados. Ela é uma doença que ainda afeta diversos países em condições deficientes de tratamento sanitário e atendimento à epidemia. Neste contexto, o objetivo deste trabalho é analisar a chance de ocorrência da Febre Tifoide no Estado do Pará, no período de 2007 a 2017, utilizando-se do modelo de Regressão Logística Binária Múltipla no *software* Estatístico R, para encontrar associação entre a confirmação da doença e características dos casos. Verificou-se no trabalho, que o modelo exposto tem significância relevante e que as características dos casos têm influência na ocorrência de Febre Tifoide, pois seus valores *p* do teste de Wald ficaram abaixo de 5%. A partir do *Odds Ratio* (Razão de Chances) dentro do intervalo de confiança de 95%, pôde-se perceber que o caso que teve *Contato* apresentou três vezes mais chance de contrair a doença do que aquele que não teve Contato, sendo que as demais variáveis permaneceram constantes.
Palavras-Chave: Contato, Intervalo de Confiança, Razão de Chances, Software Estatístico R.

Abstract

Typhoid Fever is an infectious disease caused by the bacterium Salmonella Typhi that is spread by water and poorly treated. It is a disease that still affects several countries in precarious conditions of sanitary treatment and attendance to the epidemic. In this context, the objective of this study is to analyze the odds of occurrence of Typhoid Fever in the State of Pará from 2007 to 2017, using the Multiple Binary Logistic Regression model in Statistical software R, for find an association between the confirmation of the disease and the characteristics of the cases. It was verified in the work that the exposed model has relevance and that the characteristics of the cases have influence in the occurrence of Typhoid Fever, because their p-values from the Wald test were below 5%. From the Odds Ratio within the 95% confidence interval, it was possible to notice that the case that had contact was three times more odds to contract the disease than the one that had no contact, and the other variable cases remained constant.
Keywords: Contact, Confidence Interval, Odds Ratio, R Statistical Software

[77] Universidade Federal do Pará, ailton.g1721@gmail.com
[78] Universidade Federal do Pará, sibellevilhena@gmail.com
[79] Universidade Federal do Pará, adrilayne@ufpa.br
[80] Universidade Federal do Pará, gracildo@ufpa.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 necessidade de avaliar de forma mais ponderada as informações em saúde torna necessária a tomada de estudos e medidas, principalmente em Epidemiologia, que estabeleçam novas formas de tratamento para as doenças.

As doenças epidemiológicas no Brasil é um assunto atual que preocupa a população em decorrência das necessidades e realidades observadas na efetivação da saúde. Dessa maneira, elas se tornam constantemente objetos de investigação por diversos profissionais da área e outros que queiram aplica-las aos seus estudos.

Segundo Bonita, Beaglehole e Kjellstrom (2010), a epidemiologia tem como objetivo principal favorecer os métodos de aplicação de saúde nos indivíduos.

A Febre Tifoide é uma doença infecciosa causada pela bactéria *Salmonella Typhi* que é disseminada pela água e alimentos eivados. Essa doença se caracteriza por apresentar sintomas como astenia, cefaleia, constipação, diarreia, decomposição pulso-temperatura, dores abdominais, esplenomegalia (aumento anormal do volume do baço) e febre resistente, podendo elevar-se a sintomas mais graves como a enterorragia (hemorragia nos intestinos).

Conforme Alecrim *et al*. (2002), a Febre Tifoide é um dos principais dilemas de saúde nos diversos países que apresentam condições deficientes de tratamento sanitário e atendimento à epidemia. Já no Brasil, Arruda e Araújo (1997) declaram que o Norte e o Nordeste são as regiões com maiores índices de ocorrência de Febre Tifoide, devido também à precariedade no saneamento.

Objetivo

O objetivo deste trabalho consiste em estudar a chance de ocorrência da Febre Tifoide no Estado do Pará, no período de 2007 a 2017, em função do *Contato* com a doença e dos sintomas *Diarreia* e *Esplenomegalia*, a fim de entender a proporção que essas características influenciam na ocorrência da doença.

Material e Métodos

Para este estudo, foram utilizados os Dados cedidos pela Secretaria de Estado de Saúde Pública do Estado do Pará (SESPA), em 31 de Janeiro de 2017, ao Grupo de Estudos e Pesquisas Estatísticas e Computacionais (GEPEC) e ao Laboratório de Sistema de Informação e Georreferenciamento (LASIG), ambos da Universidade Federal do Pará, os dados são referentes a 2.433 (dois mil quatrocentos e trinta e três) casos notificados de Febre Tifoide no Estado do Pará, no período de 2007 a 2017. As

variáveis em estudo são *Classificação Final, Contato, Diarreia* e *Esplenomegalia*, no qual as duas últimas são referentes a sintomas do caso.

Aplicou-se como técnica a Regressão Logística no *software* Estatístico R, pois segundo Fávero *et al*. (2009) a Regressão Logística é um método estatístico utilizado para encontrar associação entre uma variável dependente (Y) e uma ou mais variáveis explicativas (X). Neste caso, utilizou-se o modelo de Regressão Logística Binária Múltipla para retornar $P(Y_i \mid X_1, X_2, \dots, X_n)$, que é a probabilidade de ocorrer *Y* dado que já ocorreu *X*, sendo Y uma variável aleatória assumindo probabilidade de distribuição Bernoulli (valores 1 e 0) dada por

$$Y_i = \begin{cases} 1, & P(Y_i = 1) = \pi_i \\ 0, & P(Y_i = 0) = 1 - \pi_i \end{cases} \quad (1)$$

O ajuste do modelo foi feito pela função *glm*, na qual foi indicada a variável resposta e as explicativas e usada função de ligação logística e distribuição binomial, foi empregado o adquirido ajuste do modelo no comando *summary* para obter, além dos parâmetros, outros resultados importantes para análise. Desse modo, a função logística do modelo é demonstrada por

$$\pi(x_i) = \frac{e^{\left(\alpha + \sum_{i=1}^{n} \beta_i X_i\right)}}{1 + e^{\left(\alpha + \sum_{i=1}^{n} \beta_i X_i\right)}} \quad (2)$$

Sendo $\pi(x_i)$ a constante que determina a probabilidade de ocorrer determinado evento associado à variável Y; X_i é o vetor das variáveis independentes com $i = 1, 2, \dots, n$; α e β_i são os parâmetros.

Quando se pretende encontrar a chance de ocorrer determinado evento, a probabilidade do modelo observado em (2) não é o ideal. Já que a chance (*odds*) não é uma probabilidade, mas sim um valor que indica a proporção da ocorrência de interesse para cada não ocorrência, cujo valor é calculado pela razão entre a probabilidade (2) e seu complemento, como mostrado em (3)

$$chance_i = \frac{\pi(X_i)}{1 - \pi(X_i)}. \quad (3)$$

De acordo com Francisco *et al*. (2008), o critério que significa o termo Razão de Chances é o *Odds Ratio*, o qual expressa a chance de um caso evidenciado possuir a condição de interesse, comparado à chance do caso não evidenciado. Em (4) é expressa essa razão

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

$$\text{OR} = \frac{\frac{\pi(x_i=1)}{1-\pi(x_i=1)}}{\frac{\pi(x_i=0)}{1-\pi(x_i=0)}} = \frac{chance_1}{chance_0} \quad (4)$$

onde *OR* é o *odds ratio*.

Resultados e Discussão

A variável dependente binária *Classificação Final* foi codificada como *Presença* = 1 *Ausência* = 0. As variáveis explicativas *Contato*, *Diarreia* e *Esplenomegalia* também foram codificadas em presença e ausência (1;0).

Os p-valores do teste de Wald (*p*), ORs e intervalos de confiança de 95% para os ORs (I.C.) são observados na Tabela 01, onde se verifica que os valores *p* levam em consideração o máximo adotado de 5%, ou seja, como eles são menores que 0,05, apresentam evidências de que as variáveis influenciam na ocorrência de Febre Tifoide. Para a variável *Contato*, o *Odds Ratio* de 3,40 mostra que o indivíduo que teve *Contato* com caso suspeito ou confirmado de Febre Tifoide tem aproximadamente 3 (três) vezes mais chance de contrair a doença que o indivíduo que não teve Contato, desde que os sintomas *Diarreia* e *Esplenomegalia* permaneçam constante. O *Odds Ratio* de 1,84 para a variável *Diarreia*, indica que o indivíduo que apresentou este sintoma tem 84% a mais de possibilidade de contrair Febre Tifoide do que aquele que não apresentou o sintoma, desde que *Contato* e o sintoma *Esplenomegalia* permaneçam constante. No *Odds Ratio* de 0,53 é sugerido que apresentar o sintoma *Esplenomegalia* reduz em cerca de 47% a chance de ter Febre Tifoide, mantendo Contato e Diarreia constantes.

Tabela 01. *Resultados do p-valor*, *Odds Ratio e Intervalos de Confiança no software Estatístico R*, *para o Estudo dos Casos de Febre Tifoide no Estado do Pará*, *no Período de* 2007 *a* 2017.

Preditores	*p*	*OR*	I.C. LI	I.C. LS
Constante	0,000	-	-	-
Contato	0,000	3,404	2,323	4,989
Diarreia	0,001	1,839	1,295	2,610
Esplenomegalia	0,010	0,530	0,327	0,858

Nota: OR: *Odds Ratio*; I.C.: Intervalo de Confiança; LI: Limite Inferior; LS: Limite Superior.

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 *script* do *software* R estão dispostos abaixo.

```
> ajuste<-glm(Classificacao~Contato+Diarreia+Espleno,family = binomial(link
= "logit"))
> summary(ajuste)
> (Intervalo_Confianca.Parametros<-confint.default(ajuste,level = 0.95))
> (OR<-exp(ajuste$coefficients))
> (Intervalo_Confianca.OR<-exp(Intervalo_Confianca.Parametros))
```

Conclusão

Na análise da Razão de Chances, observou-se que as características dos casos tem influência na ocorrência de Febre Tifoide, sendo que os casos que teve contato com pessoas suspeitas ou confirmadas com a doença tem uma chance maior de adquiri-la. Desta forma, evidencia-se a importância do estudo como indicativo para a tomada de decisões e medidas que possibilitem um diagnóstico instantâneo de FT e seu combate.

Referências

ALECRIM, W. D.; LOUREIRO, A. C. S. P.; MORAES, R. S.; MONTE, R. L.; LACERDA, M. V. G. Febre Tifoide: recaída por resistência antimicrobiana. Relato de caso. **Revista da Sociedade Brasileira de Medicina Tropical**. Uberaba, v. 35, n. 6, p. 661-663, dez. 2002.

ARRUDA, A. H. S; ARAUJO, T. M. Epidemia de Febre Tifoide em Laranja da Terra/Espírito Santo: relato preliminar. **Informe Epidemiológico do SUS**, Brasília, v. 6, n. 2, p. 21-31, jun. 1997.

BONITA, R.; BEAGLEHOLE, R.; KJELLSTROM, T. **Epidemiologia básica**. São Paulo: Santos. 2. ed. 2010, 213 p.

FAVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. L. **Análise de Dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FRANCISCO, P. M. S. B.; DONALISIO, M. R.; BARROS, M. B. A.; CESAR, C. L. G.; CARANDINA, L; GOLDBAUM, M. Medidas de associação em estudo transversal com delineamento complexo: razão de chances e razão de prevalência. **Revista Brasileira de Epidemiologia.** 2008; 11(3):347-55.

R CORE TEAM. **R**: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 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

R AS A TOOL FOR PROMOTING UNDERGRADUATE STUDENTS ENGAGEMENT IN STATISTICS COURSES AND RESEARCH

Karla Patrícia Santos Oliveira Rodrigues Esquerre[81]
Adelmo Menezes de Aguiar Filho[82]
Robson Wilson Silva Pessoa[83]
Pedro Henrique Neri de Menezes[84]

Resumo

O foco desse estudo foi promover o ensino da estatística aos alunos de graduação utilizando o software R. Para alcançar esse objetivo, foram implementadas aulas em Markdown e Shiny, minicursos e seminários para incentivá-los a aprender estatística. Os minicursos foram organizados pelo grupo de pesquisa *Growing with Applied Modeling and Multivariate Analysis* (GAMMA) promoveu os minicursos, utilizando técnicas simples e de fácil entendimento.Portanto, observou-se um aumento da quantidade de estudantes em continuar a aprofundar-se no estudo da estatística através da realização de cursos intermediários em estatística e pesquisas de graduação e pós-graduação. Um outro objetivo foi facilitar a manipulação de dados através de estudos de simulações de problemas reais, estimulando assim, os alunos a desenvolverem o pensamento estatístico e técnicas transversais. De fato, o software R é uma ferramenta útil aos alunos de graduação para simular e visualizar a variabilidade estocástica e inferência, além de uma melhor compreensão dos métodos estatísticos individualmente ou combinados, evitando assim, cálculos cansativos, promovendo uma economia de tempo considerável.
Palavras-Chave: Statistics education, R programming,undergraduate students.

Abstract

The aim of this study was to fomenting undergraduate students by using statistics methods tightly coupled with computing in R. To achieve it, many practical lecture in Markdown and shiny,short-courses and seminars have been promoted to inspire students learn statistics. The Growing with Applied Modeling and Multivariate Analysis (GAMMA) group promoted those short-courses, focus on the basic skills of critical thinking, clear writing, and coherent speaking. Therefore, an increasing of the fraction of students interested in continuing studding statistics through taking intermediate statistics courses and on undergraduate and graduate research has been observed. Other goals were to facilitate computing with data through use of small simulation studies based on real problems and to foster students to develop a statistical thinking and transversal competences.Indeed, the open software environment R provide to undergraduate students an useful tool to simulate and visualize the stochastic variability and inference, for example, and a better understanding of statistical methods and its individual or combined assumption and applications avoiding tedious computations with a considerable saving of time.
Keywords: Statistics education, R programming, undergraduate students.

Industrial Engineering Program - PEI, Growing with Applied Modeling and Mutivariate Analysis - GAMMA, Universidade Federal da Bahia - UFBA, [1] karla.esquerre@gmail.com
[82] [2] adelmo.aguiar.filho@gmail.com
[83] [3] robsonpessoa2007@gmail.com
[84] [4] pedro_henrique_neri@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

Introduction

Practical lectures by using R programming have been recently introduced in the traditional presentation of the theory in Statistics undergraduate courses worldwide (Columbi et al, 2010; Gonzalez-Arteaga, 2011). When working individually or in group on exercises, projects or self-made simulations and programming, the students are required to actively collaborate focusing on the characteristics that are suggested by the professor. In this way, besides learning theory and programming, transversal competences as proactive problem solving and communication are addressed. The open software environment R provide to undergraduate students an useful tool to simulate and visualize the stochastic variability and inference, for example, and a better understanding of statistical methods and its individual or combined assumption and applications avoiding tedious computations with a considerable save of time. Here we briefly summarize and discuss the results of using R on education of statistics in Chemical, Production and Control and Automation undergraduate courses at Polytechnic School of Federal University of Bahia, Bahia State, Brazil. The results also embraces the research developed by some of these students who got interested in continuing learning statistics or R programming even after classes closing. Growing with Applied Modeling and Multivariate Analysis (GAMMA) research group has been the head of this teaching and researching activities.

Objective

Promoting undergraduate students engagement in statistics courses and research by using statistics methods tightly coupled with computing in R. Other goals were to facilitate computing with data through use of small simulation studies based on real problems and to foster students to develop a statistical thinking and transversal competences.

Material and Methods

A more attractive method in learning and applying statistics during undergraduate courses and research was introduced through the development of the following activities:

- R Studio was installed in about 35 computers located on 3 research and teaching computational laboratories at Polytechnic School of Engineering;

The classes were organized in such a way that the undergraduate students could reproduce the codes and the exercises applied in classroom;

- Lectures were organized using R Markdown;
- The proposed activities and exercises were built based on industry-university cooperation research and were intended to highlight a project based learning approach to statistical education focusing on data analysis, inference and modeling;
- Online open courses were presented to the students and they were encouraged to attend some of them;
- Research topics were disseminated to the students;
- Extra short-courses and seminars on R programming, R Markdown, Shiny and Multivariate Statistics were promoted to the students;
- Research projects focusing research grants to undergraduate students were written and submitted to the university (and later were approved).

Results and Discussion

GAMMA research group has been the main responsible for the practical lectures on R and for disseminating R programming in seminars and short courses for undergraduate and graduate students and researchers. This experience instilled the basic skills of critical thinking, clear writing, and coherent speaking, and challenged the research activity of the undergraduate students in utilizing more advanced programming on this research. As an example, R programming was use in the following research projects: Simulation of the process of hemocomponent production by using Monte Carlo method; Demand forecasting algorithms for evaluation of platelet stock management; Using numerical integration algorithm for second-order derivatives (Debye-Huckel equation); Evaluation of public safety data by using principal component analysis; Reliability analysis of electric transformers (Aguiar et al., 2013) and effluent overflow from accumulation basins and Statistical descriptive analysis and inference were used to evaluate the data structure of water consumption reduction (Oliveira et al, 2014) and of toxicity (Medeiros et al, 2017). The main results of using R on undergraduate statistics courses and research are summarized as follow:

- The students could solve data-oriented problems on R, regardless of their programming aptitude;

- The students got aware of the benefits of online education after their experience on taking online courses on R Programming such as the Coursera one from Roger Peng.
- The R programming has become one of the programs most used by undergraduate students when dealing with statistics due to its practicality of handling and well documented functions;
- The produced material by the undergraduate students, teaching assistants and professor were documented and shared by using Rpubs;
- Several shiny apps were elaborated in order to demonstrate statistical concepts interactively and to encourage self-learning;
- Graduate students have began to work with R as first tool or associated with other softwares.

The evaluation of advantages and disadvantages of using R in undergraduate statistics courses and research is summarized in a SWOT (Strengths, Weakness, Opportunities and Treats) Matrix shown in Figure 1.

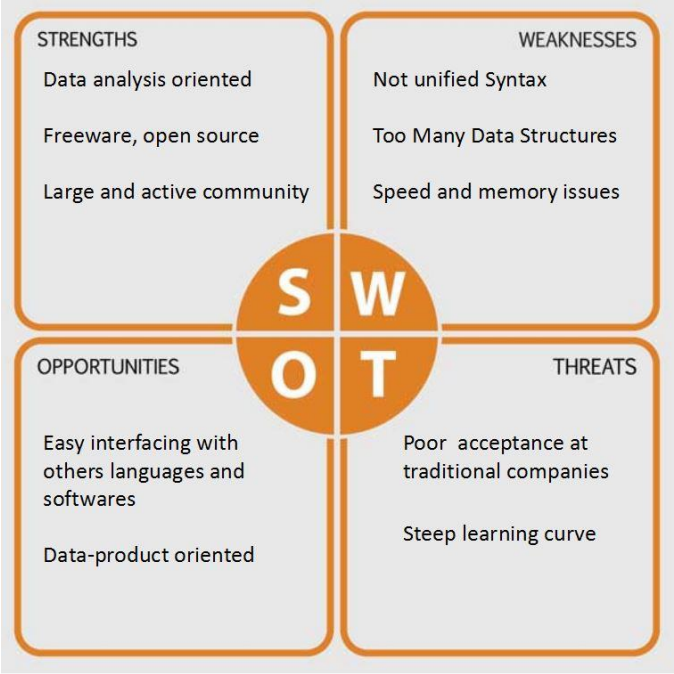


Figure1. SWOT Matrix of R application in undergraduate courses and research.

Conclusion

As a result of combining lectures of statistics theory with applied a real-life problems and R programming, an increasing of the fraction of students interested in continuing studding statistics through taking intermediate statistics courses and on undergraduate and graduate research has been observed. The experience on taking online R courses succeeded also in eliminating the students' barriers and borders, since they got aware they can learn anything they can imagine specially when guided through structured and flexible procedures. The students are required to be in contact with a new programming language and to start to learn how to implement their own statistical algorithms showing a valuable transference of competences in Computer Sciences and Engineering.

References

Aguiar Filho, A.M. de, Oliveira-Esquerre, K.P.S.R. Modelagem probabilística do tempo de vida dos transformadores de distribuição elétrica do estado da Bahia. Iniciação Científica Cesumar 15, 1–14, 2013.

Columbi, A.; Lubiano, M.A.; González-Rodriguez, G. Teaching Statistics for Computer Scienceswith R: The importance of visualizing the variability. 2nd International Conference on Education and New Learning Technologies, Barcelona, 2010.

Gonzalez-Arteaga, T. (2011). R for undergraduate statisticians. In ISI 58th World Statistics Congress of the International Statistical Institute, 2011 (pp. 6322–6325).

Medeiros Melo, T.; Bottlinger, M.; Schulz, E.; Leandro, W. .M; AguiarFilho, A. M.; Ok, Y. S.; Rinklebe, J. Effect of biosolid hydrochar on toxicity to earthworms and shrimps. Environmental Geochemistryand Health, 2017.

Oliveira, P. S.; Aguiar Filho, A.M.; Oliveira-Esquerre, K.P.S.R.; Kalid, R. A.; Contreras, C. H.; Braga, E. Aplicação de técnicas estatísticas multivariadas para compreensão do consumo de água de uma indústria de fertilizantes brasileira. 12º Congresso da Água / 16.ºENASB / XVI SILUBESA. Lisboa, Portugal, 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

CONSIDERAÇÕES SOBRE OFERTA CULTURAL NO RIO DE JANEIRO POR REGIÃO ADMINISTRATIVA, EM 2013, UTILIZANDO ANÁLISE DE AGRUPAMENTO

Daniele Cristina Dantas[85]

Resumo

Parte importante da infraestrutura para a oferta de serviços culturais existentes na cidade do Rio de Janeiro hoje foi instalada no período em que era capital do país. Este contexto favorece debates sobre o desequilíbrio na distribuição da oferta de infraestrutura de serviços culturais. Foi desenvolvido o Indicador de Oferta Cultural por Regiões Administrativas (IOC-RA), possibilitando um tipo de avaliação da oferta cultural na cidade que considera dados sobre capacidade e número de atividades realizadas (DANTAS, 2015) em equipamentos culturais sob a gestão da Secretaria Municipal de Cultura do Rio de Janeiro (SMC-RJ). O presente trabalho tem como objetivo principal processar os resultados do Indicador de Oferta Cultural por Regiões Administrativas (IOC-RA), com o uso do software estatístico R, favorecendo a análise do comportamento da oferta nas diferentes unidades de análise. São objetivos específicos fazer a análise de agrupamento, ou de *cluster*, através dos métodos de Ligação Simples e Ward e identificar se algum deles apresenta resultados que melhor auxiliem na apreciação do contexto estudado. Conclui-se que os dois métodos auxiliam na representação do agrupamento do perfil da oferta cultural a partir dos resultados do IOC-RA, sendo o método de Ward mais apropriado em função do perfil dos dados com variância elevada.
Palavras-Chave: Oferta cultural, Rio de Janeiro, análise de agrupamento

Abstract

An important part of the infrastructure for the provision of cultural services in the city of Rio de Janeiro was installed during the period when it was the capital of the country. This context favors debates about the imbalance in the distribution of cultural services infrastructure supply. The Cultural Offering Indicator by Administrative Regions (IOC-RA) was developed, allowing a type of evaluation of the cultural offer in the city that considers data on capacity and number of activities performed (DANTAS, 2015) in cultural equipment under the management of the Municipal Secretariat Of Culture of Rio de Janeiro (SMC-RJ). The present work has as main objective to process the results of the Cultural Offering Indicator by Administrative Regions (IOC-RA), using the statistical software R, favoring the analysis of the behavior of the offer in the different units of analysis. It is the specific objectives of the cluster analysis, through the methods of Single Linkage and Ward, and to identify if any of them presents results that best help in the appreciation of the studied context. It is concluded that the two methods help to represent the grouping of the cultural offer profile from the results of the IOC-RA, and Ward's method is more appropriate as a function of the profile of the data with high variance.
Keywords: Cultural offering, Rio de Janeiro, cluster analysis

[85] Doutoranda em Ciência da Informação (IBICT/UFRJ), cursando especialização em Estatística Aplicada (DEMAT/UFRRJ), danidantas1@yahoo.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

Dado o histórico da construção da cidade do Rio de Janeiro, grande parte da infraestrutura para a oferta de serviços culturais hoje existentes na cidade foi instalada em um contexto no qual a cidade foi capital do país. Município da região Sudeste, a cidade é capital do estado de mesmo nome e tem uma população de, aproximadamente, 6.320.446 habitantes, composta por 53% de mulheres e 47% de homens[0], vivendo em uma área de 1.224,56 km² [1] (DANTAS, 2015).

Dados do Instituto Municipal de Urbanismo Pereira Passos (IPP-RJ), indicam que, no ano de 2008, existiam 661 equipamentos culturais na cidade sob a gestão municipal (por fundações públicas e outras secretarias), estadual e federal, além de equipamentos privados e do terceiro setor, entre museus, bibliotecas, teatros, salas de cinema, galerias, espaços e centros culturais, escolas e sociedades musicais. Este ambiente favorece debates constantes sobre o desequilíbrio na distribuição da oferta de infraestrutura de serviços culturais e demandas por ações através das quais se busque favorecer o equilíbrio da oferta cultural nos bairros e regiões (Idem).

Neste contexto, foi desenvolvida uma proposta do Indicador de Oferta Cultural por Regiões Administrativas (IOC-RA) para avaliação da oferta cultural na cidade, considerando dados sobre capacidade e número de atividades culturais realizadas (DANTAS, 2015) em equipamentos culturais sob a gestão da Secretaria Municipal de Cultura do Rio de Janeiro (SMC-RJ).

Objetivo

O objetivo principal do trabalho é, com o uso do software estatístico R, realizar processamento dos resultados do Indicador de Oferta Cultural por Regiões Administrativas (IOC-RA) da cidade para analisar o comportamento da oferta cultural nas diferentes unidades de análise. São objetivos específicos fazer a análise de agrupamento, ou análise de *cluster*, através dos métodos hierárquicos de Ligação Simples e de Ward e identificar qual deles apresenta resultados que melhor auxiliem na análise do contexto estudado.

[0] Fonte Instituto Brasileiro de Geografia e Estatística (IBGE), disponível em <www.ibge.gov.br>.
[1] Fonte Instituto Municipal de Urbanismo Pereira Passos (IPP-RJ), disponível em <http://www.armazemdedados.rio.rj.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

Material e Métodos

São utilizados resultados do IOC-RA do ano de 2013 para arenas, lonas e centros culturais, bibliotecas, museus e teatros, além do resultado geral para cada uma das 33 (trinta e três) Regiões Administrativas da cidade. O IOC-RA foi gerado a partir do processamento de dados de registro administrativo sobre a capacidade dos 52 equipamentos culturais sob a gestão da SMC-RJ e as atividades neles realizadas entre os meses de janeiro e dezembro de 2013.

Utilizou-se análise de agrupamento, ou análise de *cluster*, através dos métodos de Ligação Simples e de Ward, utilizando o *software* estatístico R, para analisar a similaridade das regiões administrativas a partir do resultado do IOC-RA no período observado, conforme é apresentado na tabela a seguir:

Tabela 1 - Indicador de Oferta Cultural por tipo de equipamento e Região Administrativa, 2013

Região Administrativa	Arena	Biblioteca	Centro Cultural	Lona Cultural	Museu	Teatro	ICO-RA
I Portuária	0	0,59	0,14	0	0	0	0,73
II Centro	0	0	1,06	0	12,34	3,47	16,88
III Rio Comprido	0	0,72	2,42	0	0,37	0	3,51
IV Botafogo	0	2,7	1,16	0	3,24	2,55	9,65
V Copacabana	0	0	0	0	0	1,98	1,98
VI Lagoa	0	0	0	0	0,71	5,54	6,25
VII São Cristóvão	0	0	0	0	0	0	0
VIII Tijuca	0	1,98	3	0	0	2,29	7,28
IX Vila Isabel	0	0	0	0	0	0	0
X Ramos	0	0	0	0	0	0	0
XI Penha	4,95	0	0	0	0	0	4,95
XII Inhaúma	0	0	0	0	0	0	0
XIII Méier	0	0	4,41	0	0	0,83	5,24
XIV Irajá	0	2,6	0	1,17	0	0	3,77
XV Madureira	5,09	0	0	0	0	0	5,09
XVI Jacarepaguá	0	1,78	1,06	1,07	0	0	3,91
XVII Bangu	0	0	0	0,85	0	0	0,85
XVIII Campo Grande	0	1,22	0	1,32	0	0	2,54
XIX Santa Cruz	0	0,25	0	1,62	0	0	1,87
XX Ilha do Governador	0	3,26	0	1,65	0	0	4,91
XXI Paquetá	0	0	0	0	0	0	0
XXII Anchieta	0	0	0	5,81	0	0	5,81
XXIII Santa Teresa	0	0,91	3,4	0	0	0	4,31
XXIV Barra da Tijuca	0	0	0	0	0	0	0
XXV Pavuna	2,85	0	0	0	0	0	2,85
XXVI Guaratiba	3,77	0	0	0	0	0	3,77
XXVII Rocinha	0	0	0	0	0	0	0
XXVIII Jacarezinho	0	0	0	0	0	0	0
XXIX Complexo do Alemão	0	0	0	0	0	0	0
XXX Maré	0	0,66	0	1,01	0	0	1,67

(continua)

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

(continuação)

Região Administrativa	Arena	Biblioteca	Centro Cultural	Lona Cultural	Museu	Teatro	ICO-RA
XXXI Vigário Geral	0	0	0	0	0	0	0
XXXIII Realengo	0	0	0	2,16	0	0	2,16
XXXIV Cidade de Deus	0	0	0	0	0	0	0

As unidades de análise com resultado igual a zero representam as ausências de equipamentos culturais com atividade no ano de 2013 e quanto maiores os valores, maior é a representação da oferta cultural daquela unidade espacial de informação (região administrativa) para o tipo de equipamento cultural na cidade.

O método de Ligação Simples, ou de distância entre vizinhos, é um método aglomerativo hierárquico que utiliza a distância do valor mínimo considerando as iterações repetidas como apresentado no algoritmo padrão, sempre se fazendo o cálculo das distâncias mínimas entre os elementos ou grupos (DONI, 2004; MINGOTI, 2005). O método de Ward também é um método de agrupamento hierárquico que utiliza medidas de similaridade para fazer os agrupamentos calculando-os através da "soma de quadrados entre os dois agrupamentos feita sobre todas as variáveis" resultando em "agrupamentos de tamanhos aproximadamente iguais devido a sua minimização de variação interna. Em cada estágio, combinam-se os dois agrupamentos que apresentarem menor aumento na soma global de quadrados dentro dos agrupamentos" (SEIDEL et al, 2008, p. 10). O procedimento adotado também é chamado de "Mínima Variância" (MINGOTI, 2005, p.176).

Resultados e Discussão

A análise descritiva dos dados destaca o perfil dos dados utilizados apresenta as características básicas para cada tipo de equipamento cultural. Verifica-se que, em relação aos tipos de equipamento cultural, os museus têm os resultados mais expressivos enquanto arenas e lonas culturais têm valor máximo e média similares.

```
     Arena          Biblioteca     Centro Cultural   Lona Cultural
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.5048   Mean   :0.5052   Mean   :0.5045   Mean   :0.5048
 3rd Qu.:0.0000   3rd Qu.:0.6600   3rd Qu.:0.0000   3rd Qu.:0.8500
 Max.   :5.0900   Max.   :3.2600   Max.   :4.4100   Max.   :5.8100

      Museu            Teatro        ICO.RA
 Min.   : 0.0000   Min.   :0.0000   Min.   : 0.00
 1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.00
 Median : 0.0000   Median :0.0000   Median : 2.16
 Mean   : 0.5048   Mean   :0.5048   Mean   : 3.03
 3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 4.91
 Max.   :12.3400   Max.   :5.5400   Max.   :16.88
```

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 processamento dos dados no R ofereceu como resultado os agrupamentos apresentados abaixo para os métodos Ligação Simples e Ward. Pelo primeiro método criou quatro *clusters* diferentes com apenas uma Região Administrativa em cada. Pelo segundo, que considera a minimização das variâncias, os agrupamentos têm uma distribuição mais diversificada, com mais regiões em cada *cluster* criado.

Figura 1 - Dendograma | Método de Ligação Simples

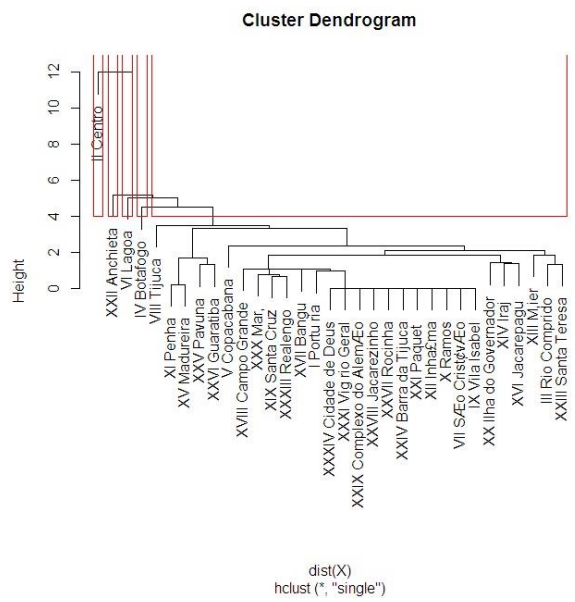
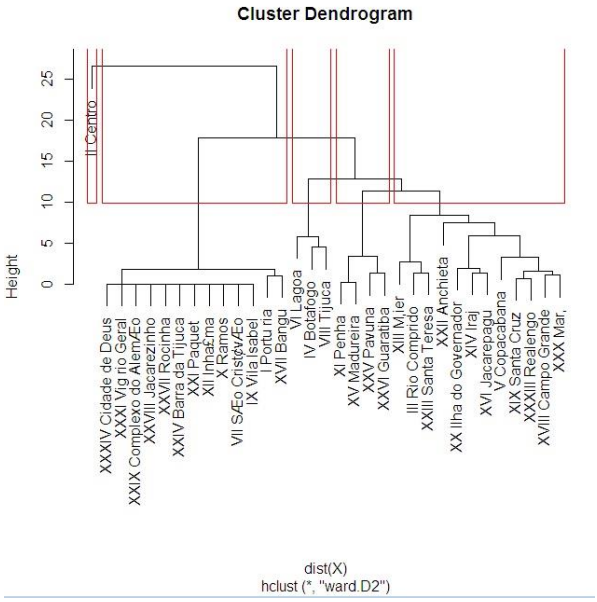


Figura 2 – Dendograma | Método de Ward



Verifica-se que cada método destaca um perfil de representação do agrupamento do perfil da oferta cultural a partir dos resultados do IOC-RA, considera-se o método de Ward o mais apropriado, visto que considera a minimização da variação interna e o conjunto de dados utilizado é significativamente influenciado por valores divergentes.

Conclusão

Os dois métodos auxiliam na representação do agrupamento do perfil da oferta cultural a partir dos resultados do IOC-RA. Pelo método de Ligação Simples a contribuição não se mostrou tão relevante, visto que, pelo perfil do conjunto de dados, poucas unidades de análise (Regiões Administrativas) figuraram como *clusters* isolados e um conjunto grande de unidades foi agrupada em um só *cluster*. O resultado serviu para reiterar que algumas regiões têm padrão de oferta cultural destacadamente diferentes de outros, acentuando que quatro regiões da cidade têm

uma dinâmica de oferta cultural destacada das outras vinte e nove; percepção que merece análises mais aprofundadas posteriormente.

Considerando que o método de Ward minimiza a variação interna no agrupamento das unidades, os clusters gerados permitem a melhor compreensão da dinâmica de distribuição dos cinco agrupamentos, a saber: Centro com valor mais extremo em relação aos demais; Regiões com oferta cultural significativamente boa (Botafogo, Lagoa e Tijuca); Regiões com oferta cultural boa (Guaratiba, Pavuna, Penha e Madureira); Regiões com oferta cultural de boa para baixa (Meier, Rio Comprido, Santa Teresa, Anchieta, Ilha do Governador, Irajá, Jacarepaguá, Copacabana, Santa Cruz, Realengo, Campo Grande e Maré); e Regiões que apresentam oferta cultural muito baixa (Bangu e Portuária) ou não apresentam oferta cultural (Cidade de Deus, Vigário Geral, Barra de Tijuca, Complexo do Alemão, Jacarezinho, Rocinha, Paquetá, Inhaúma, Ramos, São Cristóvão e Vila Isabel). O método de Ward mostrou-se mais apropriado, por agrupar as unidades de análise a partir de um perfil de similaridade um pouco mais equilibrado, através do qual é possível identificar semelhanças de resultados de Regiões distantes geograficamente, mas com perfis aproximados, evidenciadas com a minimização das discrepâncias dos resultados.

Observa-se que a análise de agrupamento é um recurso importante para a crítica dos resultados do IOC-RA e que o *software* estatístico R é uma ferramenta acessível, através da qual se alcança resultados objetivos no processamento de dados. Recurso aplicável a dados culturais com possibilidades de incremento das análises estatísticas.

Referências

DANTAS, D. C. **Indicadores para análise da oferta cultural na cidade do Rio de Janeiro: um estudo a partir de dados de registros administrativos da Secretaria Municipal de Cultura no ano de 2013**. Rio de Janeiro: ENCE/IBGE, 2015. 211p. Dissertação (Mestrado) - Programa de Pós-Graduação em Estudos Populacionais e Pesquisas Sociais, Escola Nacional de Ciências Estatísticas, Rio de Janeiro, 2015.

DONI, M. V. **Análise de Cluster: métodos hierárquicos e de particionamento**. São Paulo. Mackenzie, 2004, 93p. Monografia (Graduação), Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie, São Paulo, 2004.

MINGOTI, S. A. **Análise de Dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte, Editora UFMG, 2005. 297p.

SEIDEL, J. E., MOREIRA JUNIOR, F. de J., ANSUJ, A. P., NOAL, M. R. C. **Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite**. Ciência e Natureza. UFSM, v. 30, 2008, p. 7-15.

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.

Penúltima revisão em 09/06/2017 por Luciane às 17:49

Última revisão em 17/06/2017 por Luciane às 23:30