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

Anais

VI SEMANÍSTICA

VI Semana Acadêmica do Departamento de Estatística

da UFRGS

http://www.ufrgs.br/semanistica

Porto Alegre - 19 a 22 de outubro de 2015

Organização:

Promoção:

Conteúdo

1 Cartaz da VI SEMANÍSTICA

2 Introdução

A VI Semana Acadêmica da Estatística (VI SEMANÍSTICA) será realizada de 19 a 22 de outubro de 2015, no Instituto de Matemática e Estatística - IME, Campus do Vale da UFRGS, Porto Alegre, RS. O evento engloba os mais variados temas dentro da área acadêmica e profissional.

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

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

Para maiores informações sobre a VI SEMANÍSTICA (Semana Acadêmica da Estatística 2015) ver www.ufrgs.br/semanistica.

3 Agradecimentos

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

- ABE - Associação Brasileira de Estatística
- DEST-UFRGS - Departamento de Estatística da UFRGS
- IME-UFRGS - Instituto de Matemática e Estatística da UFRGS
- StatSoft South America

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

Comissão Organizadora

4 Comissão Organizadora Docente

- Cleber Bisognin (Coordenador - Departamento de Estatística-UFRGS)
- Danilo Marcondes Filho (Departamento de Estatística-UFRGS)
- Guilherme Pumi (Departamento de Estatística-UFRGS)
- Márcio Valk (Departamento de Estatística-UFRGS)

5 Comissão Científica

- Cleber Bisognin (Coordenador - Departamento de Estatística-UFRGS)
- Danilo Marcondes Filho (Departamento de Estatística-UFRGS)
- Guilherme Pumi (Departamento de Estatística-UFRGS)
- Márcio Valk (Departamento de Estatística-UFRGS)

6 Comissão Organizadora Discente

- Bruna Martini Dalmoro (Curso de Estatística - UFRGS)
- Cinthia Becker (Curso de Estatística - UFRGS)
- Gabriel da Cunha (Curso de Estatística - UFRGS)
- Guilherme Machado Mansson (Curso de Estatística - UFRGS)
- Luana Giongo Pedrotti (Curso de Estatística - UFRGS)
- Matias Segelis Vieira (Curso de Estatística - UFRGS)

7 Apresentação

O programa da VI SEMANÍSTICA - Semana Acadêmica do Departamento de Estatística da Universidade Federal do Rio Grande do Sul engloba as seguintes atividades

- 6 Conferências envolvendo pesquisas realizadas em diversas áreas da Estatística proferidas por pesquisadores convidados de Universidades do Rio Grande do Sul e do Brasil e ainda profissionais em Estatística;
- 1 Minicurso relacionado ao tema Data Mining;
- Comunicações orais apresentadas pelos participantes do evento;
- Colóquio relacionado ao tema Estatística aplicada à Engenharia Industrial.

8 Programação

Horário	19/10/2015	20/10/2015	21/10/2015	22/10/2015
	Segunda-Feira	Terça-Feira	Quarta-Feira	Quinta-Feira
08:30-09:15	C1	COMGRAD/EST	M1	Apresentação Oral
09:15-10:00	C2			C6
10:00-10:30	Coffe Break	Coffe Break	Coffe Break	Coffe Break
10:30-11:15	C3	COMGRAD/EST	C5	M1
11:15-12:00	C4	Apresentação Oral	Colóquio	
12:00-12:15				

Minicurso: *Data Mining: Classificação vs Agrupamento (Clustering)*

Ministrante:

Dra. Taiane Schaedler Prass

Colóquio: *Estatística aplicada à Engenharia Industrial*

Ministrante:

Prof. Dr. Ângelo Márcio Oliveira Sant'Anna - UFBA

Conferências:

(C1) Conferência 1 - Dierê Fernandez e Yasmine Caxeiro
Título: Empoderamento da Estatística

(C2) Conferência 2 - Vinícius Cassol - Tino
Título: Use a matemática para ganhar na bolsa

(C3) Conferência 3 - Renan Xavier - FEE
Título: Sistema de Exportações FEE – Metodologia e Interface

(C4) Conferência 4 – Rodrigo Coster - CONRE 4
Título: Mercado de Trabalho para Estatísticos

(C5) Conferência 5 - Josias Oliveira - StatSoft South America
Título: O mundo dos negócios com Analytics!

(C6) Conferência 6 - Prof. Drª. Sídia M. Callegari Jacques - DEST - UFRGS
Título: Aplicação de Técnicas de Agrupamento em Variáveis Biológicas

9 Minicurso

Data Mining: Classificação vs Agrupamento (Clustering)

Drª. Taiane Schaedler Prass
StatSoft South America

Resumo

Você saberia explicar qual a diferença entre classificar e agrupar? Para alguém que é novo em mineração de dados, classificação e agrupamento podem parecer semelhantes pois, em ambos os casos, os algoritmos de mineração de dados essencialmente "dividem"os conjuntos de dados em sub-conjuntos de dados. Neste minicurso abordaremos os conceitos básicos envolvendo classificação e agrupamento. Em particular, discutiremos as diferenças e semelhanças entre essas técnicas e como cada uma delas se enquadra em termos de aprendizado supervisionado e não-supervisionado. Apresentaremos alguns exemplos utilizando o software Dell STATISTICA, bem como um resumo de todas as técnicas de classificação e agrupamento disponíveis no software indicando os prós e contras de cada uma delas.

10 Colóquio

Estatística aplicada à Engenharia Industrial

Prof. Dr. Ângelo Márcio Oliveira Sant'Anna
UFBA

Resumo

A Estatística aplicada à Engenharia Industrial consiste em realizar estudos para solucionar problemas em processos industriais. A estatística é vital para compreender a variabilidade do processo, estabelecer a melhoria contínua e aumentar a qualidade dos produtos. A medida que os processos se tornam mais complexos e automatizados cresce a demanda pelo desenvolvimento e aplicação de técnicas estatísticas mais adequadas para analisar tais processos. Este seminário ilustra a aplicação dessas técnicas na solução de problemas industriais envolvendo controle estatístico de processo, planejamento de experimentos, dentre outros.

11 Conferências

Conferência 1

Empoderamento da Estatística

Dierê Fernandez e Yasmine Caxeiro
Flowing Up

Resumo

Nesta palestra vamos falar sobre o poder da Estatística no cenário atual de geração exponencial de dados. Serão debatidos aspectos sobre o posicionamento do Estatístico no mercado e de como transformar este profissional em protagonista da era da informação.

Conferência 2

Use a Matemática para Ganhar na Bolsa

Vinícius Cassol
Tino

Resumo

Análise técnica para operações na Bolsa de Valores, utilizando de ferramentas matemáticas.

Conferência 3

Sistema de Exportações FEE - Metodologia e Interface

Renan Xavier
FEE - Fundação de Economia e Estatística

Resumo

O Sistema Exportações FEE (SisExp) representa uma inovadora ferramenta das estatísticas de exportações brasileiras que realiza cálculos de índices de valor, volume e preço das exportações de todas as Unidades da Federação do Brasil para qualquer país de destino, para diferentes classificações como a Classificação Nacional de Atividades Econômicas e Intensidade Tecnológica. Além disso, apresenta as informações de valores e de participações dos mesmos setores. A primeira metade do seminário contará com a concepção metodológica estatística que os indicadores foram construídos e a segunda metade será dedicada à apresentação da interface do sistema SisExp que foi desenvolvido.

Conferência 4

Mercado de Trabalho para Estatísticos

Rodrigo Coster
Presidente CONRE 4

Resumo

Apresentação sobre mercado de trabalho para Estatísticos, mostrando as legislações pertinentes à profissão, a quantidade de alunos no país e a situação do mercado de trabalho no Rio Grande do Sul e Santa Catarina, com relatos de estatísticos das regiões.

Conferência 5

Big Data & Analytics Concepts: Aprendendo e Aplicando Big Data Analytics

Josias Oliveira
StatSoft South America

Resumo

(i) Três Elementos que mudaram o poder de processamento

(ii) A Internet das Coisas: o que há aí?

(iii) O foco na Experiência do Cliente

(iv) Prática de Analytics

(v) Cases de Sucesso

Conferência 6

Aplicação de Técnicas de Agrupamento em Variáveis Biológicas

Profª. Drª. Sídia M. Callegari Jacques
Departamento de Estatística - UFRGS

Resumo

A tremenda diversidade da natureza que nos rodeia torna difícil o estudo dos fenômenos biológicos sem métodos que permitam a organização e classificação dos seres vivos, sejam eles animais, plantas, bactérias, vírus ou o próprio homem e seus comportamentos. A classificação dos entes vivos ajuda o raciocínio e permite a interpretação dos fenômenos biológicos, em especial os evolutivos. As técnicas de agrupamento (cluster analysis) contribuíram para o desenvolvimento da Taxonomia Numérica e hoje são usadas em larga escala em estudos de Sistemática, Ecologia e Genética de populações. Nesta palestra serão apresentados alguns exemplos de aplicação destas técnicas em diferentes áreas de estudos da Biologia.

12 Seções de Comunicações

Comunicações Orais

Credit Scoring: atribuição do limite através do lucro previsto

Andressa Bruna Costa [1]
Lisiane Priscila Roldão Selau [2]

Resumo: O objetivo do presente trabalho é propor um modelo de previsão de crédito, através de regressão linear múltipla, em que a decisão da concessão é baseada na medida monetária do lucro esperado por cada proponente. Neste modelo, são reprovados aqueles com os quais se espera uma medida de lucro menor que zero, ou seja, prejuízo e, por outro lado, por representarem ganho, sugerir a atribuição do limite de crédito somente aos aprovados pelo modelo, de forma condizente com a medida monetária de lucro esperado. O desenvolvimento do modelo consiste de três grandes etapas: 1) pré-processamento, 2) construção e avaliação do modelo e 3) sistemática para a atribuição do limite. O estudo envolveu dados reais de concessão de crédito de uma rede de farmácias. De forma a identificar o aumento potencial nos ganhos através da utilização do modelo de previsão, são avaliados os cenários anterior e posterior à implementação do modelo, que demonstra uma inversão de resultados, passando de prejuízo a lucro. Nesse sentido, o modelo de previsão com variável resposta contínua, que avalia o lucro esperado, mostra-se uma ferramenta efetiva na concessão de crédito e atribuição do limite.

Palavras-chave: *Regressão linear múltipla, Decisão monetária, Atribuição de limite.*

1 Introdução

O crescimento da oferta de crédito à pessoa física nos últimos tempos tem impulsionado o comércio de produtos e serviços. Para que as empresas possam expor esse crédito, é também necessário que se avalie todas as propostas igualmente e, nesse sentido, torna-se essencial a utilização de ferramentas rápidas e eficazes que auxiliem na tomada de decisão. Neste cenário, os modelos de *Credit Scoring*, que classificam os proponentes a crédito quanto ao seu risco como cliente, estão cada vez mais importantes e necessários.

Porém, os modelos de concessão de crédito podem ir muito além de encaixar, através de sua pontuação resultante, o proponente a crédito no grupo de bons pagadores, concedendo-lhe crédito, ou no de maus pagadores, recusando-o. Gonçalves (2005) afirma que, na prática, as instituições utilizam este conceito devido à maior facilidade de trabalhar com modelos de resposta binária. Segundo Selau (2012), a utilização de uma escala dicotômica na definição do desempenho dos clientes quanto à

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

inadimplência constitui perda de informação, e sugere que escalas contínuas para determinar o comportamento de pagamento possam ter melhor aproveitamento de informação.

Steiner et al. (1999) destacam que qualquer erro na decisão de conceder o crédito pode significar que, em uma única operação, haja a perda do ganho obtido em dezenas de outras transações bem-sucedidas, já que o não recebimento representa a perda total do montante emprestado. Thomas (2000) argumenta ainda que, ao invés de procurar minimizar o percentual de clientes que não pagará, as empresas estão esperando poder identificar os clientes que são mais lucrativos.

Um desafio das empresas é oferecer a quantidade certa de recursos ao proponente a crédito. O limite de crédito é a definição do valor máximo que o concessor admite entregar ao cliente, em forma de produtos, serviços ou do próprio valor em espécie, diante da avaliação das suas características e do seu potencial de devolução de todo valor tomado dentro do prazo estipulado (SILVA, 2002).

Neste sentido, o objetivo deste trabalho é propor um modelo de concessão de crédito cuja variável resposta é uma medida monetária contínua, indicando lucro ou prejuízo com os clientes após a concessão do crédito. Com isso, pretende-se determinar, adequadamente, aos clientes com lucro esperado maior que o zero, o seu limite de crédito.

2 Método Proposto

A metodologia proposta é uma adaptação da sistemática proposta por Selau (2012). É composta de três grandes etapas e suas subetapas, conforme a Figura 1.

1 – AMOSTRA E PRÉ-PROCESSAMENTO

- Delimitar a população alvo para o modelo
- selecionar amostra e separar em análise e teste
- analisar preliminarmente do dados (limpeza e categorização)

2 – CONTRUÇÃO DO MODELO DE PREVISÃO DO LUCRO

- construir o modelo (Regressão Linear Múltipla)
- avaliar qualidade do modelo (Valor de KS e % de acerto)

3 – ATRIBUIÇÃO DE LIMITE DE CRÉDITO

- transformar o lucro em valor de venda

$$Lucro = Valor\ de\ Venda * Margem\ de\ Lucro$$

- sugerir o limite com base no lucro estimado
- fazer análise de sensibilidade: limite concedido X sugerido

Figura 1. Etapas do Método Proposto

3 Resultados

3.1 Amostra e pré-processamento

O banco de dados utilizado é composto por informações cadastrais (sexo, idade, escolaridade, profissão, etc.) de clientes de uma rede de farmácias, com unidades distribuídas em várias cidades do Rio Grande do Sul, que oferece, como uma das formas de pagamento e parcelamento, o cartão próprio (*private label*). A variável resposta é contínua e foi definida como o lucro observado após 12 meses de utilização do cartão, considerando que a empresa estima sua margem de lucro sobre o valor de venda dos produtos em 30%. Dessa forma, clientes com Lucro Observado maior que zero compõe a categoria "bom" e menor ou igual que zero, a categoria "mau".

Antes de iniciar a construção do modelo, as observações foram separadas aleatoriamente em duas amostras proporcionais, uma com 80% dos casos, que constitui a amostra de análise, utilizada para a criação do modelo, e os demais 20% utilizados para posteriormente testar o poder de predição do modelo. A amostra de análise ficou constituída de 9.981 observações, sendo 3.826 do grupo mau e 6.155 do bom e a de teste com 956 clientes do grupo mau e 1.539 do grupo bom, totalizando 2.495 observações.

Após o tratamento dos dados (verificação de preenchimento e consistência das informações), as variáveis foram transformadas de modo a estarem aptas a ingressarem no modelo, utilizando a técnica de criação de variáveis *dummy*, calculando o risco relativo (RR) para cada nível de cada atributo.

3.2 Construção do modelo de previsão do lucro

Na seleção de variáveis, foi utilizado o método automático *stepwise* para a Regressão Linear Múltipla, que traz, no modelo final, apenas as variáveis que mais são significativas e influenciam o resultado da variável dependente. Além disso, este método tende a funcionar como ação corretiva nos casos de multicolinearidade. Após a definição do modelo final, verificou-se o atendimento das suposições da regressão linear múltipla (homoscedasticidade, normalidade e independência dos erros).

Para mensurar o desempenho e a qualidade do modelo final, são utilizadas a taxa de acerto de classificações nos grupos bom e mau e o valor do teste de Kolmogorov-Smirnov (KS) para duas amostras. O percentual de classificação correta na amostra de análise foi de 60,59% e na amostra de teste foi 60,08%. Estes valores mostram que o modelo tem desempenho homogêneo para os clientes já existentes (utilizados na construção do modelo) e para os futuros proponentes a crédito deste negócio. Através do resultado do teste não paramétrico de Kolmogorov Smirnov (KS), pode-se determinar se duas amostras provêm de populações distintas, o que significaria que o modelo consegue separar os grupos bom e mau. O valor resultante do teste KS foi de 26,5, o que significa que o modelo consegue separar razoavelmente bem os dois grupos.

Com o objetivo de avaliar os resultados monetários obtidos pelo modelo, a Tabela 1 apresenta um resumo da comparação dos resultados da variável de lucro no cenário sem a utilização de nenhum modelo de risco de crédito e com a utilização do modelo proposto. Pode-se observar que a implementação do modelo de previsão do lucro inverte o quadro de prejuízo observado na empresa na ordem de R$ 83.448,08 para um lucro estimado de R$ 285.539,98 através dos valores e desempenhos praticados, em que cada cliente negado pelo modelo representaria um prejuízo de R$ 52,44 no resultado da empresa, se este fosse aprovado.

Tabela 1. Comparativo dos cenários sem e com modelo de previsão de crédito

Lucro previsto	**Sem modelo**	**Com modelo**	
		Aprovados	**Negados**
Clientes aprovados	12.476	5.440	7.036
Média (R$)	-6,69	52,49	-52,44
Total (R$)	-83.448,08	285.539,98	-368.988,06

3.3 Atribuição de limite de crédito

A atribuição do limite através do lucro previsto consiste do redimensionamento da variável resposta do modelo para seu valor de venda correspondente. Para isso, o valor de venda com cada cliente é obtido dividindo-se o lucro previsto pela margem de lucro, que é estimada pela empresa em 30%. Os valores obtidos representam, em valor monetário, quanto o cliente deve adquirir em produtos para que se obtenha o lucro esperado através da margem de lucro estimada e, portanto, correspondem ao limite recomendado ao cliente. A Tabela 2 mostra o limite médio e o total concedido pela empresa e os limites sugeridos pelo estudo em cada faixa de lucro previsto. Os resultados mostram que os limites médios concedidos pela empresa variam muito pouco entre cada classe de lucro esperado.

Os clientes com previsão de lucro de mais de R$ 160,00 (última classe) têm apenas R$ 40,00 a mais de limite que aqueles com previsão de lucro menor que R$ 20,00 (primeira classe). Através do método sugerido para a atribuição do limite, cada faixa de lucro esperado passa a ter um limite médio condizente com a classe.

Tabela 2. Limite concedido e limite sugerido pelo modelo

Lucro previsto (R$)	**Clientes**	**Limite concedido (R$)**		**Limite sugerido (R$)**	
		Médio	**Total**	**Médio**	**Total**
de 0 a 20	1006	166,98	167.980,00	20,81	20.932,94
de 20 a 40	1499	164,25	246.210,00	94,41	141.519,90
de 40 a 60	909	178,59	162.340,00	173,87	158.047,54
de 60 a 80	569	173,55	98.750,00	221,22	125.876,17
de 80 a 100	857	172,66	147.970,00	286,97	245.932,27
de 100 a 120	359	174,29	62.570,00	377,38	135.478,14
de 120 a 140	106	196,42	20.820,00	427,39	45.303,19
de 140 a 160	92	208,59	19.190,00	495,61	45.596,53
mais de 160	43	206,28	8.870,00	604,50	25.993,51
TOTAL	**5440**	**171,82**	**934.700,00**	**173,65**	**944.680,20**

4 Considerações Finais

Foi apresentada, neste estudo, uma ferramenta em que é possível prever, em medida monetária, o lucro que se espera com cada cliente, e através dessa resposta, atribuir um limite de crédito que corresponda em valor de venda de mercadorias, ao ganho previsto dada a margem de lucro. De posse do modelo construído, é realizada a análise dos resultados obtidos sem a utilização de nenhum modelo, e compara-se com os resultados obtidos com a utilização do modelo que prevê o lucro com cada cliente e atribui o limite adequado em relação ao seu lucro previsto. Com isso, tem-se um cenário em que, mesmo concedendo crédito para menos da metade dos clientes previamente aprovados, o resultado acumulado passa de prejuízo para lucro. Nesse sentido, o modelo de previsão com variável resposta contínua, que avalia o lucro esperado, mostra-se uma ferramenta efetiva na concessão de crédito.

Apesar do bom desempenho do modelo proposto, algumas limitações podem ser melhor estudadas futuramente, como i) verificação dos resultados através da cobrança de multas por atraso e possíveis alterações dos conceitos de atrasos aceitáveis e inaceitáveis, considerando, por exemplo, que atrasos de até 90 dias são aceitáveis pois, nesse contexto, há uma cobrança de juros que compensa o atraso; ii) utilização apenas de contratos encerrados, pois a maturação pode impactar no cálculo do lucro e iii) definição da margem de lucro diferente da estimada pela empresa.

Referências

GONÇALVES, E. B. *Análise de risco de crédito com o uso de modelos de regressão logística, redes neurais e algoritmos genéticos*. Dissertação de Mestrado, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, 2005.

SELAU, L. P. R. *Modelagem para Concessão de Crédito a pessoas físicas em empresas comerciais: da decisão binária para a decisão monetária*. Tese de Doutorado, Programa de Pós-Graduação em Administração, Universidade Federal do Rio Grande do Sul, 2012.

SILVA, J. A. *Análise do estabelecimento do limite de crédito - Um estudo de caso*. Dissertação de Mestrado, Departamento de Economia, Contabilidade, Administração e Secretariado, Universidade de Taubaté, 2002.

STEINER, M. T. A.; CARNIERI, C.; KOPITTKE, B. H.; STEINER NETO, P. J. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. *Revista de Administração*, São Paulo, v. 34, n. 3, p. 56-67, 1999.

THOMAS, L. C., A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, v. 16, n. 2, p. 149-172, 2000.

Identificação de *outliers* em modelos *k-Factor* Gegenbauer, resultados de simulação a partir do algoritmo SODA

Ian Danilevicz [1] [3]

Cleber Bisognin[2] [3]

Resumo: Neste trabalho seguimos aprimorando a identificação de *outliers* em processos com longa dependência. O modelo que estamos investigando, são os modelos da classe *k-Factor* Gegenbauer e o método de identificação é o SODA. Nesse algoritmo de identificação é atribuída uma estatística de teste para classificar cada uma das posições da série temporal como outlier aditivo, outliers inovador, ou não outlier. Para avaliar a utilidade dessa estatística Γ procedemos com estudos de simulação em que vários arranjos de modelos *k-Factor* Gegenbauer foram propostos e avaliamos a dispersão dessa estatística em cada um dos casos.

Palavras-chave: *processos estocásticos, longa dependência, Gegenbauer, Outliers*.

1 Introdução

No presente momento, a identificação de outliers ocupa uma posição de grande preocupação entre analistas e estatísticos, dada a dificuldade de encontrá-los e os grandes desequilíbrios que eles acarretam em modelos não robustos. No contexto de séries temporais essa questão se agrava, pois estamos lidando com dados correlacionados e as ferramentas para identificá-los deve levar em conta essa estrutura. O algoritmo SODA é desenhado de forma a levar em conta um modelo de série temporal e a partir desse filtro identificar os possíveis outliers por uma estatística Γ_i para as i observações em uma série. Dessa forma, valores altos da Γ_i indicam maior chance de estarmos em uma posição i que seja outlier.

A função SODA foi originalmente desenhada para processos da classe ARMA. No entanto, redesenhamos essa função para os modelos $k - Factor$ Gegenbauer, um tipo particular de processo estocástico sobre os quais estamos trabalhando a algum tempo. Neste resumo discutimos o comportamento dessa estatística Γ para diversos casos simulados, pois nos interessa saber o que é um valor alto ou baixo, para termos uma ferramenta acurada na identificação de outliers. Além disso, a presente estatística tem duas subdivisões a Γ_{AO} e a Γ_{IO}, valores adaptados para outliers aditivos e inovadores, respectivamente. Um

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

[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *cleberbisognin@google.com*

[3]Agradecimento ao CNPq pela bolsa de Iniciação Científica

Outlier Aditivo é uma observação anômala que aparece pontualmente na série, já um Outlier Inovador instaura um subperíodo de observações anômalas na série causando uma mudança estrutural na média ou mesmo na variância do processo estocástico.

2 Resultados de Simulação

Nesta seção apresentamos os resultados de simulação para a identificação de outliers pelo método SODA. Todas as simulações tomam como base modelos $k-Factor$-GARMA (0,**u**,λ,0) com **u**={-0.7,0.5,0.8} e λ={0.1,0.2,0.3}, no entanto as contaminações não são sempre as mesmas. Nossa primeira figura 1(a) é apenas um exemplo de uma série contaminada por sete outliers do tipo AO e ω, magnitude de contaminação, igual a 5. Ao lado, 1(b) temos a mesma estrutura de contaminação para um caso multivariado, ou seja, mil séries. Escolhemos ω=5 por entendermos que esse é um outlier de tamanho médio.

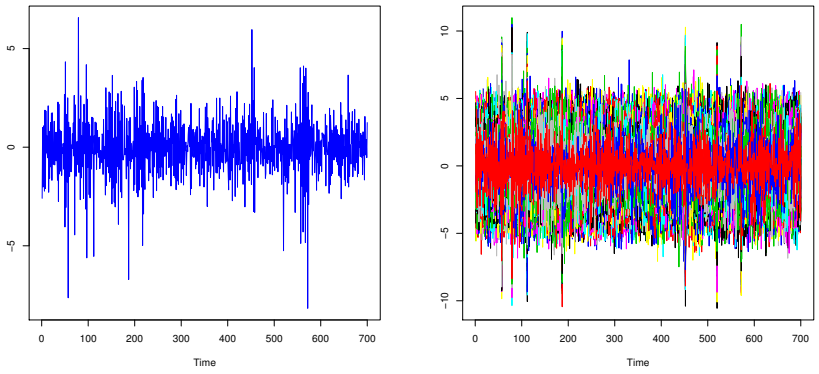


Figura 1: Séries Simuladas que seguem o modelo K-FACTOR-GARMA (0,**u**,λ,0) com **u**={-0.7,0.5,0.8} e λ={0.1,0.2,0.3}, com n=700 e 7 observações contaminadas por outliers AO de magnitude ω=5: (a) Modelo Univariado (b) Modelo Multivariado, ou seja, mil séries

Para testar a eficiência do algoritmo SODA geramos mil repetições de cada tipo de combinação de parâmetros, isto é, mil séries para cada ω = {0.5, 1, 3, 5}, para cada número de observações contaminadas, quais sejam 7, 14 ou 28, e, finalmente para cada tipo de contaminação, quais sejam aditivos e inovadores. No entanto, para poupar espaço apresentamos apenas os casos de ω extremos, ou seja, ω={0.5,5}. Como estamos procedendo com uma contaminação paramétrica, sabemos quais as posições contaminadas e por que tipo de outlier, em cada grupo de séries. Portanto, podemos separar as observações conforme a sua classe e avaliar os valores da estatística Γ.

Apresentamos uma tabela descritiva 2, média e desvio padrão, da estatística Γ do SODA que nos ajuda a classificar corretamente as observações de uma série como outliers ou não. Além disso po-

Tabela 1: Descritivas de Γ para 1000 séries de tamanho 700 com 7, 14 e 28 observações contaminadas por outliers tipo AO ou IO de magnitude ω=5

Série com Outliers Aditivos									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	3.369	2.873	0.465	3.369	2.866	0.477	3.360	2.847	0.495
Desv Pad	0.507	0.640	0.441	0.524	0.642	0.460	0.586	0.721	0.508
Série com Outliers Inovadores									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	1.622	1.448	0.249	1.636	1.445	0.261	1.635	1.441	0.283
Desv Pad	0.254	0.320	0.243	0.259	0.330	0.263	0.268	0.367	0.297

demos especificar esta estatística para outliers do tipo I e II. Para processos da classe ARMA, o valor de referência é 2, sendo valores superiores a dois fortes candidatos a serem corretamente classificados como outliers, e, se $\Gamma_{AO} > \Gamma_{IO}$ a observação deve ser do tipo I e, analogamente, se o contrário. Para os modelos K-FACTOR-GARMA ainda não temos esse valor de referência, nem temos certeza se o comportamento Γ_{AO} e Γ_{IO} é o mesmo.

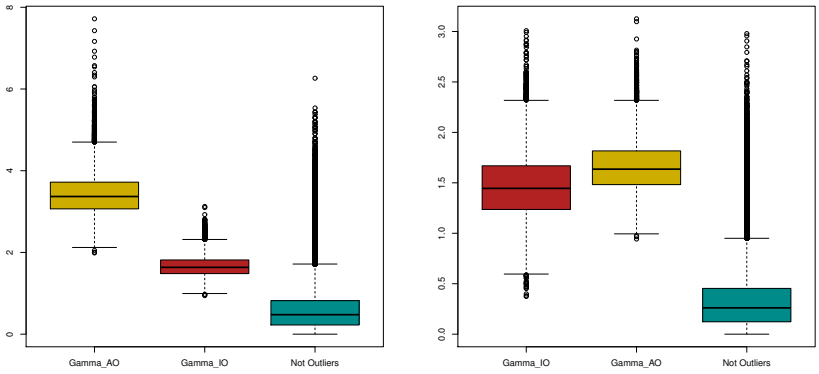


Figura 2: Box-Plot de Γ para 1000 séries de n=700 com 14 observações contaminadas por outliers de magnitude ω=5: (a) Séries contaminadas por AO, (b) Séries contaminadas por IO

Para ilustrar os resultados da tabela 1, apresentamos dois gráficos 2(a) e 2(b) para séries contaminadas por outliers do tipo I e II, respectivamente. Escolhemos apresentar somente o caso com um número intermediário de contaminações, ou seja, 14 observações contaminadas. Enfatizamos que o número de observações contaminadas pouco afeta a distribuição da estatística Γ, ou seja, os demais casos são semelhante. O mesmo não vale para quando mudamos os valores de magnitude, ω, dos outliers, como mostraremos mais adiante.

Pelo gráfico 2(a), fica claro que se a contaminação é do tipo I, podemos separar as observações

contaminadas das não contaminadas, além disso Γ_{AO} apresenta valores superiores a 3, Γ_{IO} valores intermediários entre 1.5 e 2, enquanto as observações não contaminadas valores abaixo de 1.5, com raras exceções. Já para o caso das séries contaminadas por outliers do tipo II 2(a), a separação entre observações contaminadas e não contaminadas funciona bem, no entanto a classificação quanto ao tipo de outlier não é tão simples, pois Γ_{AO} e Γ_{IO} ocupam regiões semelhantes. Sendo que a classe AO apresenta valores inclusive superiores em média, uma contradição, portanto, apenas ficamos seguros de estarmos diante de verdadeiros outliers do tipo I se Γ_{AO} for bem maior do que Γ_{IO}.

Tabela 2: Descritivas de Γ para 1000 séries de tamanho 700 com 7, 14 e 28 observações contaminadas por outliers tipo AO ou IO de magnitude ω=0.5

Série com Outliers Aditivos									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	0.624	0.590	0.378	0.615	0.592	0.380	0.613	0.593	0.379
Desv Pad	0.416	0.385	0.339	0.420	0.387	0.339	0.421	0.387	0.339
Série com Outliers Inovadores									
N outliers		7			14			28	
Tipo	AO	IO	NC	AO	IO	NC	AO	IO	NC
Mediana	0.302	0.279	0.173	0.296	0.282	0.173	0.297	0.282	0.173
Desv Pad	0.180	0.172	0.155	0.183	0.174	0.155	0.184	0.174	0.155

Os gráficos nos ajudam a perceber que as distribuições de todas as estatísticas Γ são assimétricas positivas, ou seja, com grande presença de valores extremos na cauda direita. Por esse motivo escolhemos trabalhar com a mediana ao invés da média em nossas tabelas descritivas, pois essa medida é mais robusta, ou seja, menos influenciável por valores extremos.

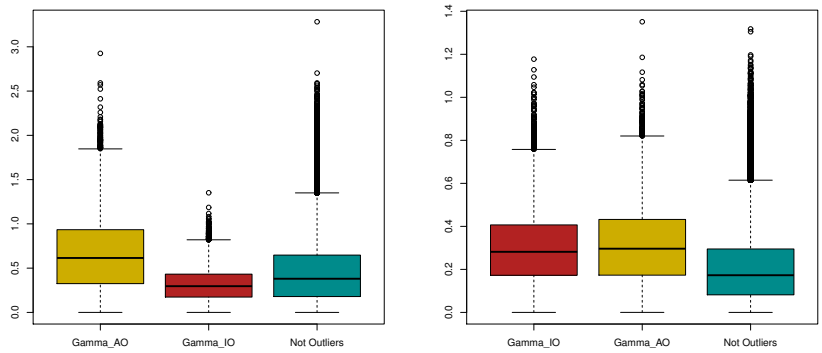


Figura 3: Box-Plot de Γ para 1000 séries de n=700 com 14 observações contaminadas por outliers de magnitude ω=0.5: (a) Séries contaminadas por AO, (b) Séries contaminadas por IO

Já a tabela e gráfico sobre os dados em que a contaminação é da ordem de ω=0.5 temos menos otimismo em identificar os outliers. Pois para o caso de contaminação do tipo AO, todos os tipos de Γ

flutuam na mesma região, ou seja, entre 0.2 e 1.0, sendo difícil a separação dos mesmos. A situação não melhora para o caso IO, em que Γ praticamente tem distribuições idênticas entre 0.15 e 0.4. Temos dessa forma, uma indicação dos limites do nosso trabalho, ou seja, outliers muito pequenos não são bem identificados pelo método proposto, qual seja o SODA especialmente desenhado para processos $k - Factor$ Gegenbauer.

3 Conclusões Parciais

Até este momento, sugerimos que outliers de magninitude media como ω=5 não são dificies de serem identificados pelo algoritmo SODA devidamente ajustado para o respectivo modelo $k - Factor$ Gegenbauer. No entanto, se a magnitude do outlier é muito pequena, como ω=0.5 temos maiores dificuldade de identificar esse outlier pois a estatística Γ das observações anômalas e não anômalas é muito semelhante. Porém, acreditamos que outliers assim pequenos não acarretam muitos problemas de alteração de parâmetros da série temporal, claro está que precisamos ainda investigar essa suposição.

Referências

[1] Fox, R.; Taqqu, M.S. "Large-Sample Properties of Estimates for Strongly Stationary Gaussian Time Series". *Annals of Statistics*, Vol. **14** (2), p.517-532, 1986.

[2] Woodward, W.A., Q.C. Cheng e H.L. Gray (1998). "A k-Factor GARMA Long-Memory Model". *Journal of Time Series Analysis*, Vol. **19**(4), pp. 485-504.

Admissões e demissões de trabalhadores de saúde do Rio Grande do Sul, por meio da metodologia Box e Jenkins

Afonso Valau de Lima Junior[1]
Jonatan da Rosa Pereira da Silva[2]
Adriano Mendonça Souza[3]

Resumo: O objetivo desta pesquisa é analisar as séries referente as admissões e demissões de enfermeiros e técnicos em enfermagem no estado do Rio Grande do Sul. Dentre os modelos selecionados para cada série, obteve-se o modelo ARIMA (0,1,1) para série referente a admissão de enfermeiros; ARIMA (0,1,1) para a demissão de enfermeiros; ARIMA (0,1,2) para a demissão de técnicos em enfermagem e um modelo SARIMA (1,1,1) $(1,0,1)_8$ para a admissão de técnicos em enfermagem.

Palavras-chave: *Enfermeiros, Técnicos em enfermagem, ARIMA, SARIMA.*

1 Introdução

A enfermagem tem suas práticas de trabalho centradas no cuidado ao paciente e realiza funções de cunho técnico a saúde. O mercado de trabalho desta profissão tem sofrido inúmeras mudanças, devido às novas tecnologias e o aperfeiçoamento dos profissionais, tornando a disputa por empregos mais acirrada. O objetivo deste estudo é prever o comportamento da dinâmica de admissão e demissão, além de descrever e verificar e o comportamento destas variáveis em relação a enfermeiros e técnicos em enfermagem no estado do Rio Grande do Sul, por meio da metodologia Box e Jenkins.

2 Metodologia

Os dados utilizados para a execução deste estudo foram obtidos na página do ministério do trabalho e Emprego (MTE) e referem-se as admissões e demissões de enfermeiros e técnicos em enfermagem no estado do Rio Grande do Sul, no período de janeiro de 2007 a junho de 2015, totalizando 102 observações mensais de cada variável. Utilizou-se o auxílio do software Eviews 8 SV para realização deste estudo.

Sabendo-se da condição de uma série temporal (MORETTIN, TOLOI, 1986), espera-se que a mesma seja estacionária. Neste estudo optou-se, para verificar a estacionariedade das séries, os testes:

[1] UFSM - Universidade Federal de Santa Maria. Email: avljunior*@yahoo.com.br*
[2] UFSM - Universidade Federal de Santa Maria. Email: jonatanprd*@gmail.com*
[3] UFSM - Universidade Federal de Santa Maria. Email: Segundo-autor*@ufrgs.br*

Augmented Dickey-Fuller (ADF) e *Kwiatkowski-Phillips-Schmidt-Shin* (KPSS). No ADF tem-se as seguintes hipóteses de teste: H_0: a série é não-estacionária, isto é, I(1); H_1: a série é estacionária, isto é, I(0). No teste KPSS as hipótese apresentadas são inversas ao teste ADF. Logo: H_0: a série é estacionária, isto é, I(0); H_1: a série é não-estacionária, isto é, I(1). Estes testes são amplamente discutidos na literatura por exemplo, Enders (1995).

Suplantada esta condição passa-se para a etapa da identificação da estrutura do modelo, pela análise das funções: autocorrelação (ACF) e autocorrelação parcial (PACF). Onde espera-se identificar um modelo da classe geral ARIMA (*p, d, q*), onde AR(*p*) corresponde a parte autorregressiva de ordem *p*, MA(*q*) corresponde ao processo de médias móveis de ordem *q*, *d* o número de diferenças necessárias para tornar a série estacionária. O modelo para se tornar, um modelo concorrente, deve apresentar resíduos com a característica de *ruído branco*, isto é, um conjunto de variáveis aleatórias com média igual a zero, distribuição normal, variância constantes e não autocorrelacionados.

Caso a variável em análise não siga um modelo ARIMA, investiga-se então um modelo sazonal da forma SARIMA (p,d,q) $(P,D,Q)_S$, nos quais *p* e *q* refere-se às ordens auto-regressiva e de média móvel, respectivamente e as ordens auto-regressiva e de média móvel sazonais é representado por *P* e *Q*, respectivamente (VICINI e SOUZA, 2007). Os valores *d* e *D* representam respectivamente as diferenças de ordem simples e sazonais. A equação na íntegra do modelo SARIMA(p,q,d) $(P,D,Q)_S$ é dado por:

$$Y_t = \delta + \sum_{i=1}^{p} \phi_i \Delta^d Y_{i-1} + \sum_{i=1}^{P} \Phi_i^P \Delta^D Y_{i-1} + \mu_j + \sum_{j=1}^{q} \theta_j \mu_{j-1} + \sum_{j=1}^{q} \Theta_j^Q \mu_{j-1} + \varepsilon_t$$

, onde:

$\delta + \sum_{i=1}^{p} \phi_i \Delta^d Y_{i-1}$, é a parte auto-regressiva não-sazonal de ordem p;

$\sum_{i=1}^{P} \Phi_i^P \Delta^D Y_{i-1}$, é a parte auto-regressiva sazonal de ordem P e estação sazonal s;

$\mu_j + \sum_{j=1}^{q} \theta_j \mu_{j-1}$, é parte de integração não-sazonal de ordem d;

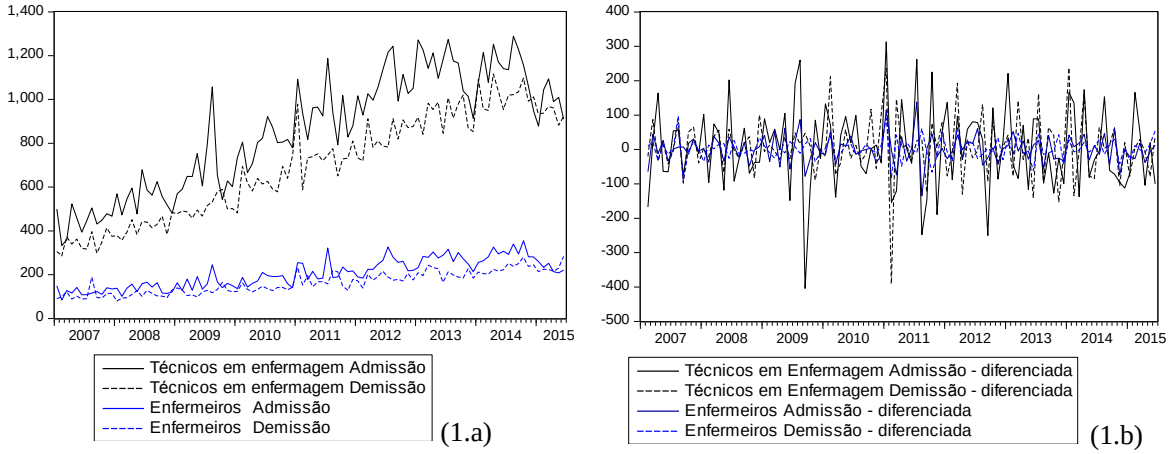
$\sum_{j=1}^{q} \Theta_j^Q \mu_{j-1}$, é a parte sazonal de médias móveis de ordem Q e estação sazonal s,

ε_t , ruído branco.

A seleção dos modelos será por meio dos critérios AIC (*Akaike Information Criteria*) e BIC (*Bayesian Information Criteria*) através das equações AIC= $\ln\sigma_e^2 + (2(p + q))/n$ e BIC = $\ln\sigma_e^2 + ((p + q) \ln n)/n$; *p* e *q* são os parâmetros conhecidos, n é o tamanho da amostra, ln é o logaritmo neperiano e σ_e^2 a variância estimada dos erros, levando em conta que quanto menor for o AIC e BIC mais adequado estar o modelo para a projeção dos valores futuros da série (MORETTIN, 2008).

3 Resultados e Discussões

As séries originais do estudo estão apresentadas na Figura 1 (a), observa-se que as séries referentes aos técnicos em enfermagem (demissão e admissão) e aos enfermeiros (admissão e demissão) possuem uma tendência ascendente.



(1.a) (1.b)

Figura 1 – (a) Séries originais, (b) Séries diferenciadas

Visualmente é possível observar que as séries não são estacionária em nível, na Figura 1 (b), observa-se as séries com 1 diferença possivelmente se tornam estacionárias. Para confirmar como analise visual, conforme foi mencionada na metodologia, utilizou-se dos testes ADF e KPSS, para verificar a estacionariedade das séries, os resultados dos testes encontra-se na Tabela 1.

TABELA 1 – Teste de raiz unitária por meio dos testes ADF e KPSS.

Variável	Série em nível		Série 1ª diferença	
	ADF[a]	KPSS[b]	ADF[a]	KPSS[b]
Técnico (Admissão)	0.4770 (I(1))	1.1273 (I(1))	**0.0000 (I(0))**	**0.1273 (I(0))**
Enfermeiro (Admissão)	0.3631 (I(1))	1.1227 (I(1))	**0.0000 (I(0))**	**0.3314 (I(0))**
Técnico (Demissão)	0.6409 (I(1))	1.2051 (I(1))	**0.0000 (I(0))**	**0.2747 (I(0))**
Enfermeiro (Demissão)	0.7431 (I(1))	1.2127 (I(1))	**0.0000 (I(0))**	**0.1441 (I(0))**

Notas: [a] Valores críticos de MacKinnon (1996): -3.493.129 (1%); -2.888.932 (5%) e -2.581.453 (10%).
[b] Valores críticos de Kwiatkowski-Phillips-Schmidt-Shin(1992, Table 1): 0.739 (1%); 0.463 (5%) e 0.347 (10%).

Na Tabela 1, observa-se que, as quatro variáveis utilizadas no estudo, analisadas por meio do teste ADF, onde a hipótese de nulidade é de que a variável possui uma raiz unitária I(1), isto é, ordem de integração igual a 1, não é rejeitada. Quando aplicado o teste de hipótese nas séries diferenciadas, há a evidência estatística de não aceitação da hipótese nula e, portanto as séries diferenciadas apresentam-se estacionárias, isto é, I(0). Esses resultados são apoiados pelos resultados do teste KPSS que possui como hipótese de nulidade de que a série é estacionária, isto é, I(0), portanto o inverso do que diz as hipóteses do teste anteriormente aplicado. Conclui-se assim, por meio dos testes, que as séries em nível são não estacionárias e após a primeira diferença as mesmas tornam-se estacionárias.

Ao efetuar o próximo passo metodológico, que consiste na identificação de modelos significativos que representem o comportamento das séries, conforme descrito na metodologia, foram selecionados dentre os modelos concorrentes, em cada série, os que apresentaram os menos valores de AIC e BIC. Os modelos selecionados para cada séries estão descritos na Tabela 2, juntamente com os coeficientes, *p-value*, AIC e BIC:

TABELA 2 – Modelos selecionados para as variáveis em estudo.

Série	Modelo	Coeficientes	*p –value*	AIC	BIC
Enfermeiros Admissão	**ARIMA (0,1,1)**	$\theta_1 =$ -0,65	*<0,05*	9,92	9,94
Enfermeiros Demissão	**ARIMA (0,1,1)**	$\theta_1 =$ -0,76	*<0,05*	9,31	9,34
Técnico em enfermagem Demissão	**ARIMA (0,1,2)**	$\theta_1 =$ -0,86 $\theta_2 =$ 0,21	*<0,05*	11,32	11,38
Técnico em enfermagem Admissão	**SARIMA (1,1,1) (1,0,1)$_8$**	$\varphi_1 =$ -0,38 $\theta_1^* =$ -0,44 $\Phi_1 =$ 0,54 $\Theta_1 =$ -0,90	*<0,05*	12,02	12,13

*Parâmetro significativo no lag =2.

Diante dos modelos selecionados, na Figura 2 são apresentadas as séries em estudos (diferenciadas), os resíduos produzido pelos modelos escolhidos e também série prevista com base no modelo proposto.

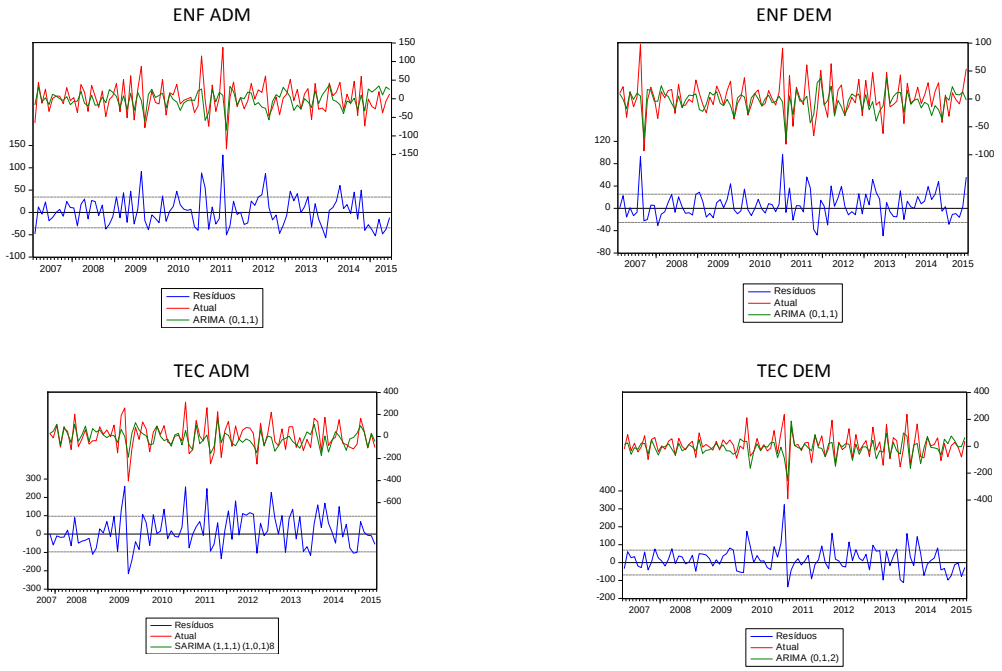


Figura 2 – Séries original em primeiras diferenças, a série ajustada e os resíduos dos modelos

Os valores previstos através dos modelos selecionados para cada variável estão expressos na Figura 3, onde também é apresentado os intervalos de confiança - de mais ou menos dois desvios padrão - é possível verificar visualmente que os modelos escolhidos se ajustam bem as séries em estudo.

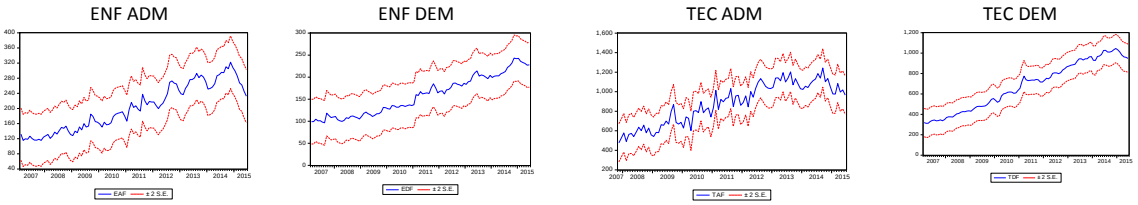


Figura 3 – Previsão das séries em estudo.

As previsões encontradas mostram que as séries analisadas mantém o seu comportamento ascendente, apesar de que nos últimos períodos são observados uma queda em todas as variáveis, sendo as séries referente as admissões (enfermeiros e técnicos) que apresentam uma queda mais acentuada.

4 Conclusão

Ao realizar o estudo referente as séries de admissão e demissão de enfermeiro e técnicos em enfermagem no Rio Grande do Sul entre janeiro de 2007 e junho de 2015 encontrou-se o modelo ARIMA (0,1,1) para série referente a admissão de enfermeiros; ARIMA (0,1,1) para a demissão de enfermeiros; ARIMA (0,1,2) para a demissão de técnicos em enfermagem e um modelo SARIMA $(1,1,1)\ (1,0,1)_8$. Nas quatro séries analisadas os modelos Box e Jenkins, conseguiram captar os movimentos das séries, inclusive a referente a demissão de técnicos em enfermagem que apresentou sazonalidade. Deixa-se como sugestão de estudos futuros a estimação conjunta destas variáveis como forma de verificar o seu interelacionamento por meio de um modelo de vetores autorregressivos.

Referências

[1] ENDERS, W. *Applied Econometric Time Series*. John Wiley & Sons, New York, 1995.

[2] MINISTÉRIO DO TRABALHO E EMPREGO – MET. Disponível em: < http://portal.mte.gov.br/ >. Acesso em: 01 out. 2015.

[3] MORETTIN, Pedro A.. *Enconometria financeira*: um curso de séries temporais financeiras. São Paulo: Blucher, 2008.

[4] MORETTIN, Pedro A.; TOLOI Clélia M.. *Métodos quantitativos*: séries temporais.São Paulo: Atual, 1986.

[5] VICINI, L. e SOUZA, A. M.. Geração de subsídios para a tomada de decisão na cadeia produtiva da bovinocultura do Brasil. *Gestão de Produção, Operações e Sistemas*, 4, 49-64. 2007.

Estudo de simulação para comparar métodos de estimação da distribuição do consumo alimentar usual: ISU, NCI, MSM e SPADE.

G H C Laureano[1]
V B L Torman[2]
S P Crispim[3]
A L M Dekkers[4]
S A Camey[5]

Resumo: Muitos métodos estão disponíveis para estimar as distribuições de consumo alimentar usual e, por isso, há uma necessidade de estudos de simulação para compará-los. Os métodos ISU (*Iowa State University*), NCI (*National Cancer Institute*), MSM (*Multiple Source Method*), e SPADE (*Statistical Program to Assess Dietary Exposure*) foram comparados anteriormente em um estudo realizado por SOUVEREIN et al., 2011, mas alguns resultados não foram conclusivos devido ao pequeno número de replicações utilizadas na simulação. Buscando superar esta limitação, o presente estudo utilizou 1.000 replicações na simulação para doze cenários diferentes, com foco em amostras pequenas e grandes razões de variâncias, para comparar a precisão e acurácia das estimativas geradas pelos métodos acima mencionados. O vício e o vício relativo foram utilizados como medidas de acurácia, enquanto que o erro quadrático médio foi utilizado como a medida da precisão. Para pequenas amostras, os métodos ISU, MSM e SPADE obtiveram estimativas mais acuradas e precisas do que o método NCI, particularmente para os percentis 5 e 95.

Palavras-chave: *Distribuição de consumo alimentar usual, ISU, NCI, MSM, SPADE.*

Introdução

A mensuração do consumo alimentar usual é assunto vigente na área da saúde, uma vez que várias doenças são influenciadas, ou até mesmo causadas, pelos hábitos alimentares do indivíduo (DONG et al., 2011). Uma das formas de estudar o consumo alimentar de uma população é por meio da estimação da distribuição do consumo alimentar usual. O estudo dessas distribuições traz um melhor entendimento a respeito do consumo alimentar em populações, como também, pode ajudar na identificação de possíveis grupos de risco para o desenvolvimento de doenças.

Foram propostas diversas metodologias de estimação dessa distribuição (GAY, 200AD; SLOB, 1993, 2006; WALLACE; DUAN; ZIEGENFUS, 1994; BUCK; HAMMERSTROM; RYAN, 1995; NUSSER et al., 1996; GUENTHER; KOTT; CARRIQUIRY, 1997; HOFFMANN et al., 2002; TOOZE; GRUNWALD; JONES, 2002; TOOZE et al., 2006, 2010; WAIJERS et al., 2006; KIPNIS et al., 2009; HAUBROCK et al., 2011; DEKKERS; SLOB, 2012; NUSSER; FULLER; GUENTHER, 2012), e que alguns autores realizaram estudos onde

[1]Programa de pós graduação em Epidemiologia, Universidade Federal do Rio Grande do Sul (UFRGS). Email: greice.laureano@ufrgs.br
[2]Programa de pós graduação em Epidemiologia e Departamento de Estatística, UFRGS. Email: vanessa.leotti@ufrgs.br
[3]Departamento de Nutrição, Universidade Federal do Paraná (UFPR). Email: crispim@ufpr.br
[4]Netherlands National Institute for Public Health and the Environment. Email: Arnold.Dekkers@rivm.nl
[5]Programa de pós graduação em Epidemiologia e Departamento de Estatística, UFRGS. Email: camey@mat.ufrgs.br

comparavam métodos de estimação da distribuição (GOEDHART et al., ; NUSSER et al., 1996; GUENTHER; KOTT; CARRIQUIRY, 1997; HOFFMANN et al., 2002; DODD et al., 2006; TOOZE et al., 2010; SOUVEREIN et al., 2011).

No trabalho de SOUVEREIN et al., 2011, os métodos ISU (*Iowa State University*), NCI (*National Cancer Institute*), MSM (*Multiple Source Method*) e SPADE (*Statistical Program to Assess Dietary Exposure*) foram comparados, mas alguns resultados foram inconclusivos devido ao pequeno número de replicações na simulação. Com o objetivo de superar essa limitação, este trabalho utilizou um número maior de replicações dando enfoque aos cenários que apresentaram maior incerteza nas conclusões que foram os com razões de variâncias iguais a 4 e 9, tamanhos amostrais iguais a 150, 500 e distribuições de consumo assimétricas. Para melhor entendimento de alguns resultados foi adicionado tamanho de amostra igual a 300.

O objetivo deste trabalho é comparar os métodos ISU, NCI, MSM e SPADE em um estudo de simulação e comparar suas acurácias e suas precisões na estimação da distribuição de consumo alimentar usual.

Métodos

O consumo alimentar usual, em geral, tem distribuição assimétrica (DODD et al., 2006), o que impossibilita a utilização de métodos estatísticos baseados na distribuição Normal. Nesses casos uma das estratégias utilizadas é a transformação dos dados.

Quase todos os métodos existentes (GAY, 200AD; SLOB, 1993, 2006; WALLACE; DUAN; ZIEGENFUS, 1994; BUCK; HAMMERSTROM; RYAN, 1995; NUSSER et al., 1996; GUENTHER; KOTT; CARRIQUIRY, 1997; HOFFMANN et al., 2002; TOOZE; GRUNWALD; JONES, 2002; TOOZE et al., 2006, 2010; WAIJERS et al., 2006; KIPNIS et al., 2009; HAUBROCK et al., 2011; DEKKERS; SLOB, 2012; NUSSER; FULLER; GUENTHER, 2012) estimam as variâncias (intraindividual e interindividual) e a média do consumo da população, na escala transformada (normal). Posteriormente faz-se a transformação inversa para a escala original e então se obtém a média e os percentis da distribuição do consumo alimentar usual.

Esses métodos diferem em relação às formas de transformação, aos modelos para estimar as variâncias e a média na escala transformada, às formas de se obter a transformação inversa e aos métodos de obtenção da média e dos percentis da distribuição.

O ISU proposto pela *Iowa State University* (NUSSER et al., 1996; GUENTHER; KOTT; CARRIQUIRY, 1997; NUSSER; FULLER; GUENTHER, 2012) tem duas implementações: uma que está em SAS (“SAS | Business Analytics and Business Intelligence Software”) e outra que pode ser obtida por meio de solicitação para os autores no site: http://www.side.stat.iastate.edu/pc-side.php/.

O NCI proposto pelo *National Cancer Institute* (TOOZE; GRUNWALD; JONES, 2002; TOOZE et al., 2006; KIPNIS et al., 2009; TOOZE et al., 2010), está implementado no SAS ("SAS | Business Analytics and Business Intelligence Software") e pode ser encontrado no site: http://riskfactor.cancer.gov/diet/usualintakes/.

O MSM foi proposto na Europa por uma equipe da Alemanha (HARTTIG et al., 2011; HAUBROCK et al., 2011) e está implementado na *web* no site: https://msm.dife.de/.

O SPADE (SOUVEREIN et al., 2011; DEKKERS; SLOB, 2012; DEKKERS et al., 2014) está implementado no *software* R (R CORE TEAM, 2015) e é baseado no programa AGEMODE (WAIJERS et al., 2006) onde a estimação dos consumos é modelada em função da covariável idade. A informação de idade pode ser omitida após ajustes no código do software. O SPADE é um pacote do R chamado SPADE.RIVM que pode ser obtido no endereço: www.spade.nl.

Neste estudo foram simulados 12 cenários. Na simulação foram utilizados os programas R (para a geração dos dados e para o SPADE), SAS (para o ISU e NCI) e o AutoHotkey ("AutoHotkey") (para o MSM).

Para comparação das estimativas se calculou para cada método o vício (V), o vício relativo em percentagem (VR) e o erro quadrático médio (EQM).

Para o cálculo dessas medidas foi necessário obter a média e os percentis verdadeiros e, para isso, utilizou-se a aproximação por quadratura Gaussiana (obtida pela função *f.gauss.quad* implementada no pacote SPADE.RIVM do *software* R).

Resultados

Em termos de acurácia e precisão os métodos ISU, MSM e SPADE obtiveram estimativas mais precisas e acuradas o método NCI em quase todos cenários, particularmente para os percentis 5 e 95 quando o tamanho da amostra é pequeno e a razão de variância é grande.

Discussão

Neste trabalho foi realizado um estudo de simulação que visava comparar quatro métodos utilizados para estimação da distribuição de consumo alimentar usual. Os resultados obtidos nos cenários simulados mostraram que, em termos de precisão e acurácia, o tamanho de amostra interferiu na qualidade das estimações de todos métodos.

Os resultados também mostraram que há uma estimação de pior qualidade em todos os métodos quando se trata da estimação dos percentis 5 e 95, sendo o percentil 95 o pior caso, pois obteve estimações com menor precisão.

Conclusões

Em conclusão, este estudo mostrou a importância do tamanho da amostra, que deve ser grande o suficiente para evitar problemas numéricos nos métodos utilizados.

Além disso, verificou-se também que os métodos tiveram uma performance semelhante entre si quando se utiliza tamanhos de amostra iguais a 150, 300 ou 500, apesar do método NCI ser menos preciso e acurado que os outros dois métodos.

Referências

AutoHotkey. Disponível em: <http://www.autohotkey.com/>.

BUCK, R. J.; HAMMERSTROM, K. A.; RYAN, P. B. Estimating Long-Term Exposures from Short-Term Measurements. **Journal of exposure analysis and environmental epidemiology**, v. 5, n. 3, p. 359–373, set. 1995.

DEKKERS, A. L. M.; SLOB, W. Gaussian Quadrature is an efficient method for the back-transformation in estimating the usual intake distribution when assessing dietary exposure. **Food and Chemical Toxicology**, v. 50, n. 10, p. 3853–3861, out. 2012. . Acesso em: 2 out. 2013.

DEKKERS, A. L.; VERKAIK-KLOOSTERMAN, J.; VAN ROSSUM, C. T.; OCKE, M. C. SPADE, a New Statistical Program to Estimate Habitual Dietary Intake from Multiple Food Sources and Dietary Supplements. **Journal of Nutrition**, v. 144, n. 12, p. 2083–2091, 1 dez. 2014.

DODD, K. W.; GUENTHER, P. M.; FREEDMAN, L. S.; SUBAR, A. F.; KIPNIS, V.; MIDTHUNE, D.; TOOZE, J. A.; KREBS-SMITH, S. M. Statistical Methods for Estimating Usual Intake of Nutrients and Foods: A Review of the Theory. **Journal of the American Dietetic Association**, v. 106, n. 10, p. 1640–1650, out. 2006.

DONG, J.-Y.; ZHANG, L.; HE, K.; QIN, L.-Q. Dairy Consumption and Risk of Breast Cancer: A Meta-Analysis of Prospective Cohort Studies. **Breast cancer research and treatment**, v. 127, n. 1, p. 23–31, maio 2011.

GAY, C. Estimation of population distributions of habitual nutrient intake based on a short-run weighed food diary. **British Journal of Nutrition**, p. 287–293, 200AD.

GOEDHART, P. W.; VOET, H.; SVEN, K.; DEKKERS, A. L. M.; DOOD, K. W.; BOEING, H.; KLAVEREN, J. A comparison by simulation of different methods to estimate the usual intake distribution for episodically consumed foods. **EFSA Journal**, [s.d.]

GUENTHER, P. M.; KOTT, P. S.; CARRIQUIRY, A. L. Development of an Approach for Estimating Usual Nutrient Intake Distributions at the Population Level. **The Journal of nutrition**, v. 127, n. 6, p. 1106–1112, jun. 1997.

HARTTIG, U.; HAUBROCK, J.; KNÜPPEL, S.; BOEING, H.; EFCOVAL CONSORTIUM. The MSM Program: Web-Based Statistics Package for Estimating Usual Dietary Intake Using the Multiple Source Method. **European journal of clinical nutrition**, v. 65 Suppl 1, p. S87–91, jul. 2011.

HAUBROCK, J.; NÖTHLINGS, U.; VOLATIER, J.-L.; DEKKERS, A.; OCKÉ, M.; HARTTIG, U.; ILLNER, A.-K.; KNÜPPEL, S.; ANDERSEN, L. F.; BOEING, H.; EUROPEAN FOOD CONSUMPTION VALIDATION CONSORTIUM. Estimating Usual Food Intake Distributions by Using the Multiple Source Method in the EPIC-Potsdam Calibration Study. **The Journal of nutrition**, v. 141, n. 5, p. 914–920, maio 2011.

HOFFMANN, K.; BOEING, H.; DUFOUR, A.; VOLATIER, J. L.; TELMAN, J.; VIRTANEN, M.; BECKER, W.; DE HENAUW, S.; EFCOSUM GROUP. Estimating the Distribution of Usual Dietary

Intake by Short-Term Measurements. **European journal of clinical nutrition**, v. 56 Suppl 2, p. S53–62, maio 2002.

KIPNIS, V.; MIDTHUNE, D.; BUCKMAN, D. W.; DODD, K. W.; GUENTHER, P. M.; KREBS-SMITH, S. M.; SUBAR, A. F.; TOOZE, J. A.; CARROLL, R. J.; FREEDMAN, L. S. Modeling Data with Excess Zeros and Measurement Error: Application to Evaluating Relationships between Episodically Consumed Foods and Health Outcomes. **Biometrics**, v. 65, n. 4, p. 1003–1010, dez. 2009.

NUSSER, S. M.; CARRIQUIRY, A. L.; DODD, K. W.; FULLER, W. A. A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions. **Journal of the American Statistical Association**, v. 91, n. 436, p. 1440–1449, dez. 1996. . Acesso em: 2 out. 2013.

NUSSER, S. M.; FULLER, W. A.; GUENTHER, P. M. Estimating Usual Dietary Intake Distributions: Adjusting for Measurement Error and Nonnormality in 24-Hour Food Intake Data. In: LYBERG, L.; BIEMER, P.; COLLINS, M.; DE LEEUW, E.; DIPPO, C.; SCHWARZ, N.; TREWIN, D. (Ed.). **Survey Measurement and Process Quality**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012. p. 689–709.

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

SAS | Business Analytics and Business Intelligence Software. Disponível em: <http://www.sas.com/>.

SLOB, W. Modeling Long-Term Exposure of the Whole Population to Chemicals in Food. **Risk analysis: an official publication of the Society for Risk Analysis**, v. 13, n. 5, p. 525–530, out. 1993.

SLOB, W. Probabilistic Dietary Exposure Assessment Taking into Account Variability in Both Amount and Frequency of Consumption. **Food and chemical toxicology: an international journal published for the British Industrial Biological Research Association**, v. 44, n. 7, p. 933–951, jul. 2006.

SOUVEREIN, O. W.; DEKKERS, A. L.; GEELEN, A.; HAUBROCK, J.; DE VRIES, J. H.; OCKÉ, M. C.; HARTTIG, U.; BOEING, H.; VAN 'T VEER, P.; EFCOVAL CONSORTIUM. Comparing Four Methods to Estimate Usual Intake Distributions. **European journal of clinical nutrition**, v. 65 Suppl 1, p. S92–101, jul. 2011.

TOOZE, J. A.; GRUNWALD, G. K.; JONES, R. H. Analysis of Repeated Measures Data with Clumping at Zero. **Statistical methods in medical research**, v. 11, n. 4, p. 341–355, ago. 2002.

TOOZE, J. A.; KIPNIS, V.; BUCKMAN, D. W.; CARROLL, R. J.; FREEDMAN, L. S.; GUENTHER, P. M.; KREBS-SMITH, S. M.; SUBAR, A. F.; DODD, K. W. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: The NCI method. **Statistics in Medicine**, v. 29, n. 27, p. 2857–2868, 30 nov. 2010. . Acesso em: 16 jan. 2013.

TOOZE, J. A.; MIDTHUNE, D.; DODD, K. W.; FREEDMAN, L. S.; KREBS-SMITH, S. M.; SUBAR, A. F.; GUENTHER, P. M.; CARROLL, R. J.; KIPNIS, V. A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application to Their Distribution. **Journal of the American Dietetic Association**, v. 106, n. 10, p. 1575–1587, out. 2006.

WAIJERS, P. M. C. M.; DEKKERS, A. L. M.; BOER, J. M. A.; BOSHUIZEN, H. C.; VAN ROSSUM, C. T. M. The Potential of AGE MODE, an Age-Dependent Model, to Estimate Usual Intakes and Prevalences of Inadequate Intakes in a Population. **The Journal of nutrition**, v. 136, n. 11, p. 2916–2920, nov. 2006.

WALLACE, L. A.; DUAN, N.; ZIEGENFUS, R. Can Long-Term Exposure Distributions Be Predicted from Short-Term Measurements? **Risk Analysis**, v. 14, n. 1, p. 75–85, fev. 1994. . Acesso em: 2 out. 2013.

Uma experiência de pesquisa de campo com alunos do ensino médio noturno

Daniel Ânderson Müller[1]

Luciana Neves Nunes[2]

Resumo: Este trabalho consiste no desenvolvimento e execução de uma pesquisa de campo realizada pelos alunos de uma turma de ensino médio noturno de uma escola pública estadual situada em uma pequena cidade do interior do Rio Grande do Sul. A pesquisa foi concebida a partir de questões de interesse dos alunos e teve como principal objetivo promover o aprendizado de conceitos de Estatística através da obtenção de dados relevantes a respeito da comunidade. Ao final da atividade, foi possível perceber que estimular que os alunos vivenciem todo o processo de pesquisa de campo qualifica o currículo e o aprendizado da teoria estatística.

Palavras-chave: *Pesquisa de campo, ensino médio noturno, educação estatística.*

1 Introdução

Os Parâmetros Curriculares Nacionais (PCNs) recomendam que a coleta e análise de dados deve ser ensinado desde as séries iniciais. Deve-se ver a Estatística como "um conjunto de ideias e procedimentos que permitem aplicar a Matemática em questões do mundo real" (BRASIL, 2002, p. 123). A Estatística é entendida, portanto, como um ramo da Matemática direcionado à solução de problemas do cotidiano. Mesmo que tópicos de estatística permeiem os estudos em praticamente todas as áreas do conhecimento, o ensino da teoria estatística acaba ficando sob a responsabilidade dos professores de Matemática.

Atualmente, há uma grande quantidade de informações disponíveis em diversos meios. Isso é campo fértil para as mais variadas atividades relacionadas à Estatística na escola. Porém, "diante desse ambiente saturado de informações, poucas pessoas questionam a forma como esses dados foram coletados, tratados e trabalhados até chegarem no formato 'acabado' em que são apresentados". (ROSETTI, 2007, p. 37). Essa é uma das razões para se incluir o ensino de Estatística no currículo da educação básica, pois propicia o desenvolvimento de uma leitura crítica dos dados.

Cazorla e Santana (2006) nos dizem que os alunos gostam muito de coletar dados, construir gráficos e interpretar fenômenos, o que torna a Estatística a parte mais divertida e cidadã da Matemática. Uma possibilidade para que se faça da aprendizagem estatística algo que seja de fato significativo é a coleta, resumo e análise de dados da realidade dos alunos. Isso pode ser feito através de projetos de

[1] Mestrando do PPGEMAT – UFRGS – Universidade Federal do Rio Grande do Sul. E-mail: danielmuller@live.com

[2] Departamento de Estatística – UFRGS – Universidade Federal do Rio Grande do Sul. E-mail: lununes@mat.ufrgs.br

pesquisa interdisciplinares. Com os procedimentos de coleta e sistematização de informações fornecidos pela Estatística, surgem oportunidades de realizar atividades interdisciplinares da Matemática com outras áreas do conhecimento. As diversas áreas do conhecimento contribuiriam com as questões de pesquisa. A área da Matemática, por sua vez, ficaria responsável pela normatização do processo, através da teoria estatística. A partir da análise dos dados obtidos pelos alunos, pode-se elaborar propostas de intervenção na realidade, algo tão almejado e valorizado na concepção de educação atual, presente inclusive em avaliações de larga escala, como o Enem.

A atividade aqui descrita consiste na realização de uma pesquisa de campo fundamentada num trabalho colaborativo de alunos do ensino médio noturno, com vistas a fazer diagnósticos da realidade da comunidade onde a escola está inserida. Como objetivo principal, pretende-se estimular o aprendizado de conteúdos de Estatística através da realização de uma atividade prática e significativa.

2 Metodologia e desenvolvimento da proposta

2.1 Concepção da proposta

Em 2011, motivada pelos elevados índices de reprovação e evasão escolar que eram verificados nas escolas da Rede Estadual de Ensino, a Secretaria de Estado da Educação do Rio Grande do Sul (SEDUC) divulgou uma proposta que instituiria o Ensino Médio Politécnico como forma de reestruturação do Ensino Médio da Rede Estadual de Ensino do Rio Grande do Sul a partir do ano seguinte. Essa nova concepção de ensino médio impôs mudanças profundas, desde a estrutura curricular até, principalmente, o processo de avaliação da aprendizagem.

Mesmo com uma forte resistência por parte dos professores, estava claro que o processo seria irreversível. Mas houve quem, como eu, enxergasse ali uma bela oportunidade de desenvolver estratégias inovadoras de ensino. A disciplina de Seminário Integrado, integrada ao currículo pela proposta da SEDUC, é destinada ao desenvolvimento de "projetos vivenciais" dos alunos. Poderíamos direcionar o trabalho na disciplina para o diagnóstico de problemas, partindo do interesse dos alunos (dando sentido e significado ao trabalho escolar), para pensar ações de intervenção na realidade (esfera crítica da educação).

2.2 Preparação e realização da pesquisa

A atividade foi realizada no segundo semestre de 2014. A turma, do 2º ano do ensino médio noturno, era composta por 17 alunos com idades entre 16 e 18 anos. Desde meados do primeiro trimestre letivo, foram trabalhados os conceitos relacionados à Estatística Descritiva. A partir de setembro, as aulas de Seminário Integrado desta turma foram integralmente destinadas ao desenvolvimento e realização da mencionada pesquisa junto aos moradores da cidade onde está localizada a escola. As

atividades foram organizadas em encontros de três períodos de 40 minutos às terças feiras. Excepcionalmente, alguns encontros aconteceram em outros dias da semana.

Todo o projeto foi dividido em três etapas, a saber:

I. Elaboração do instrumento de pesquisa: Durante o mês de setembro, foram destinadas 15 aulas de 40 minutos de duração para as atividades de:

a) Levantamento das temáticas a serem abordadas na pesquisa: Numa roda de conversa, motivados pela minha preocupação com a evasão escolar e a aparente pequena importância que é dada à educação naquela cidade, os alunos também expressaram suas preocupações em relação à cidade onde moram. Assuntos como educação, emprego, consumo de drogas pelos jovens, espaços para lazer e atividades físicas, saneamento básico do município, entre outros, foram recorrentes.

b) Elaboração e testagem do instrumento de pesquisa: Um instrumento com 31 questões foi elaborado. Na parte inicial, questões para traçar o perfil do entrevistado (idade, sexo, escolaridade, trabalho, etc.), na parte intermediária, questões sobre o estilo de vida (prática de atividade física, visitas ao médico, crença religiosa, time de futebol, etc.) e, na parte final, a avaliação do entrevistado quanto a diversos aspectos do município (saúde, educação, saneamento, etc.). Foram entrevistados inicialmente 49 pessoas para testar o instrumento de pesquisa. Posteriormente, algumas adaptações foram feitas nas questões para que fosse dado início ao trabalho de campo.

c) Cálculo do tamanho da amostra: Calculamos o tamanho da amostra necessária pela fórmula $n = p \cdot q \cdot z^2/d^2$, onde: n é o tamanho da amostra; p é a proporção assumida para a variável pesquisada; $q = 1 - p$; z é um valor oriundo de uma distribuição normal e relacionado ao nível de confiança da pesquisa; d é a diferença aceitável ou margem de erro da pesquisa. Definindo $p = 0{,}5$, $z = 1{,}96$ (relativo a um nível de confiança de 95%), e $d = 0{,}05$, chegamos a um tamanho mínimo de 385 pessoas para a amostra. Foram feitas simulações com outros valores, mas os alunos consideraram viável trabalhar o tamanho de amostra inicialmente calculado.

d) Treinamento dos alunos pesquisadores e organização da pesquisa de campo: Antes de ir a campo, os alunos simularam entrevistas, incluindo algumas situações que poderiam acontecer (recusa em responder, ninguém na residência, endereço incorreto, etc.). Com as simulações, os alunos procuraram aperfeiçoar o processo, reduzindo o tempo de cada entrevista e definindo alguns procedimentos para contornar eventuais dificuldades. Por exemplo, em caso de impossibilidade de realizar a entrevista no endereço pré-determinado, os alunos decidiram que iriam para a residência imediatamente à direita da sorteada. Ficou definido, ainda, que as visitas seriam feitas em duplas, com os alunos identificados por crachá.

II. Pesquisa de campo: Os meses de outubro e novembro foram destinados à pesquisa propriamente dita. Atividades desenvolvidas:

a) Sorteio da amostra aleatória simples: Foi obtida junto à Prefeitura Municipal a lista completa e um mapa de endereços cadastrados na zona urbana referentes ao IPTU. Quanto à zona rural, nos foi fornecida a relação de prontuários da Secretaria da Saúde (cada um deles fazia referência a uma família

cadastrada). No total, foram obtidos 1936 endereços, que foram todos listados e numerados numa planilha do Microsoft Excel. Os alunos sortearam as 385 residências que seriam visitadas usando o algoritmo de geração de números aleatórios do site Random.org. Para facilitar o trabalho, as duplas de pesquisadores visitariam os endereços mais próximos de suas residências.

b) Pesquisa de campo: Os alunos decidiram esperar o término das eleições para iniciar as entrevistas. Como trabalhavam durante o dia e estudavam à noite, utilizaram principalmente os fins de semana para dar andamento à pesquisa. Com isso, a maior parte dos endereços sorteados puderam ser visitados, uma vez que havia alguém na residência que aceitara responder às questões dos alunos.

c) Compilação dos dados coletados: Assim que os instrumentos eram preenchidos, os alunos iam preenchendo uma planilha do Microsoft Excel com as respostas dos entrevistados.

d) Relatos das experiências dos alunos durante a pesquisa de campo: Nesta fase, realizamos debates e discussões para contornar dificuldades e imprevistos, reorganizando o trabalho. Tivemos, inclusive, que sortear alguns endereços adicionais para dar conta da amostra mínima necessária.

III. Análise dos dados e elaboração de relatórios: Esta etapa foi realizada entre o final do mês de novembro e o início de dezembro, culminando na apresentação dos resultados da pesquisa no dia 9 de dezembro de 2014. Atividades desenvolvidas:

a) Resumo dos resultados da pesquisa: Com os dados pesquisados por todas as duplas unidos em uma única planilha, usamos o Microsoft Excel para representar os resultados obtidos. Cada dupla de estudantes definiu duas ou três variáveis que analisariam e representariam em gráficos. A turma escolheu gráficos de setores (chamados de pizza no software utilizado) para representar os resultados.

b) Elaboração de relatórios da pesquisa: foi montada uma apresentação de slides com o software Microsoft PowerPoint e um documento de texto do Microsoft Word para relatar os resultados. Cada dupla fez o trabalho sobre as variáveis previamente escolhidas e tudo foi reunido em arquivos únicos ao final.

c) Apresentação dos resultados à comunidade escolar: Reunidos todos os alunos do turno da noite e os professores presentes na escola, os alunos apresentaram o processo de pesquisa, dificuldades, aprendizagens e os resultados propriamente ditos.

3 Resultados e Considerações Finais

Como resultados relevantes da pesquisa, a partir da análise e opinião dos alunos, destacam-se a baixa escolaridade da população do município (55% dos entrevistados têm no máximo o ensino fundamental completo) e a satisfação da população no tocante à maioria dos aspectos analisados no município (com exceção ao serviço de saneamento básico, alvo de críticas por parte dos entrevistados). Os alunos também perceberam uma relação entre o grau de instrução dos entrevistados e sua renda mensal: maior nível de escolaridade parece resultar em salários melhores. Os professores presentes na

apresentação dos resultados consideram que os dados obtidos podem servir como subsídios para projetos de intervenção na realidade escolar, por exemplo, para reduzir a evasão escolar no ensino médio noturno.

Em se tratando de resultados da atividade, destaca-se o envolvimento dos alunos. A adesão da turma foi completa e a maioria dos alunos permaneceu focada e entusiasmada do início ao fim. Concluíram o ano letivo orgulhosos pela empreitada que realizaram. Alguns deles reportaram que não faziam ideia da complexidade que existe num processo de pesquisa, sobretudo quanto aos cuidados metodológicos na definição das questões de pesquisa e na seleção da amostra a ser pesquisada.

Para mim, esta atividade foi especialmente interessante por ter colocado os alunos em contato com todo o processo de pesquisa, desde a elaboração de questões de pesquisa, seleção da amostra a ser pesquisada, realização das entrevistas e, finalmente, a compilação, análise e apresentação dos dados coletados. Os alunos tiveram total autonomia para, inclusive, amenizar contratempos e encontrar soluções para todos os problemas que surgiram.

Destaco que esta atividade foi realizada com alunos de ensino médio noturno de uma escola pública. Estes alunos trabalhavam durante o dia e estudavam à noite, tendo dedicado boa parte de seu tempo livre para a realização da pesquisa. Mesmo assim, a motivação dos alunos ao participar de uma atividade significativa para a comunidade onde moram foi algo de inestimável valor.

Fica a sugestão para a realização de atividades semelhantes, de modo a qualificar o estudo de Estatística no ensino médio e, com isso, envolver os alunos em suas comunidades. Especificamente em se tratando do Ensino Médio Politécnico, a inclusão de tópicos da Teoria Estatística pode ser feita à disciplina de Seminário Integrado, de modo a qualificar também os projetos de pesquisa dos alunos.

Referências

[1] BRASIL. Ministério da Educação. Secretaria da Educação Média e Tecnológica. *Parâmetros Curriculares Nacionais + (PCN+) - Ciências da Natureza e suas Tecnologias*. Brasília: MEC, 2002. Disponível em: <http://portal.mec.gov.br/seb/arquivos/pdf/CienciasNatureza.pdf>. Acesso em 28 set. 2015.

[2] ROSETTI, Hélio Jr. Educação Estatística no ensino básico: uma exigência do mundo do trabalho. *Revista Capixaba de Ciência e Tecnologia*, Vitória, n. 2, p. 35-37, 1. sem. 2007. Disponível em: <http://recitec.cefetes.br/artigo/documentos/Artigo%205.pdf>. Acesso em: 01 out. 2015.

[3] CAZORLA, Irene Mauricio; SANTANA, Eurivalda Ribeiro dos Santos. *Tratamento da informação para o ensino fundamental e médio*. Série Alfabetização Matemática, Estatística e Científica. Itabuna, Via Litterarum, 2006.

[4] SECRETARIA DE ESTADO DA EDUCAÇÃO (RIO GRANDE DO SUL) (SEDUC). *Proposta Pedagógica para o Ensino Médio Politécnico e Educação Profissional Integrada ao Ensino Médio*. Porto Alegre, nov. 2011. Disponível em: <http://www.educacao.rs.gov.br/dados/ens_med_proposta.pdf>. Acesso em: 20 maio 2014.

Text Mining: Descrição da utilização do pacote Rfacebook

Carolina Peçaibes de Oliveira[1]

Guilherme Pumi[2]

Introdução

Com o advento da internet, tornou-se disponível uma grande quantidade de informações relevantes em forma de texto, e surgiu a demanda por processos e algoritmos capazes de obter, organizar, classificar, depurar e analisar esses dados não estruturados (sendo essas etapas que compõe o processo de *text mining*). Há interesse especial em entender os dados pertinentes ao comportamento do consumidor, sendo estes frequentemente expressos através das redes sociais. Com isso em mente, trazemos aqui uma introdução ao processo de obtenção de dados para aplicação posterior de *text mining*, utilizando o pacote *RFacebook* do software R.

Metodologia

Para a realização desse trabalho, executamos as etapas necessárias de autenticação para extrair dados utilizando o RFacebook. Posteriormente, utilizamos algumas funções do RFacebook aplicáveis à páginas públicas na página oficial do jornal Zero Hora (getPage e getPost) e outras aplicáveis à perfis pessoais na página pessoal da autora (getFriends, getLikes) no dia 05/10/2015, obtendo quatro base de dados com informações em forma de texto.

Desenvolvimento

Realizamos a instalação do pacote no software R usando os comandos apropriados. Em seguida, a documentação do pacote RFacebook apresenta como primeira função disponível o comando *fbOAuth* que exige as informações de *App ID* e *App Secret* para autenticação do acesso ao facebook através do R, informando que esses dados estão disponíveis no endereço www.developers.facebook.com/apps.Porém a partir dessa página não há instruções de como

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

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

obter esses dados e o que eles são.

Verificamos através de pesquisa no FAQ do Facebook que esses dados provém da criação de um aplicativo, e que o usuário precisa registrar-se como desenvolvedor de apps utilizando sua conta pessoal do Facebook, registrar a criação de aplicativo. Isto feito, as informações de ID (*App ID*) e Senha de Acesso (*App Secret*) aprecem disponíveis, e podem ser inseridas no comando do R que deve ser rodado nesse momento.

Esse comando nos retorna uma URL que deve ser inserida nas informações de registro do aplicativo no Facebook. Com isso está completa a autenticação de acesso.

Para executar os demais comandos do pacote, é exigida a informação do *token*, que está disponível no aplicativo criado e autenticado.

Realizamos a extração de informações do feed de notícias a página oficial da *Zero Hora* usando a função **getPage**. Por se tratar de uma página pública, temos acesso aos seus dados completos. Informamos o número de postagens que queremos extrair e obtemos uma base de dados completa em forma de lista, incluindo o texto do post, o tipo de post, o link compartilhado (se houver), a data e hora da postagem, número de curtidas e número de comentários. Com a função **getPost** extraímos informações detalhadas de cada post, como quantidade de comentários e curtidas, nome do perfil dos usuários que comentaram ou curtiram o post, quantidade de curtidas de cada comentário, e data e hora dos mesmos.

Executamos alguns testes de extração e verificamos que não há um limite de quantidade de postagens e comentários que podemos obter. Além disso, por tratar-se de uma página pública, independentemente do tipo de configuração de privacidade do usuário do facebook, seus comentários e curtidas ficam disponíveis para extração e, posteriormente, análise.

A partir da página pessoal da autora, executamos a função **getFriends** para obter uma listagem do nome dos seus amigos na rede social, além das outras informações que o comando fornece como data de nascimento, gênero, profissão informada e escolaridade. Com a função **getLikes** obtemos a relação de páginas curtidas por aquele perfil. Verificamos que a primeira função retornou o nome apenas de dois amigos do perfil, o que é incorreto segundo a conferência da página original. De acordo com as configurações do facebook, apenas usuários que autorizam a visualização irrestrita de suas informações tem seus dados disponíveis por esse procedimento. Também não é possível extrair a lista de amigos de um perfil com visualização restrita, tornando as informações escassas para uma análise de mercado, por exemplo.

O output de cada um desses quatr comandos nos traz uma lista do R, que convertemos para o formato de matriz e exportamos para o excel, para fins de visualização. Cada tipo de variável de texto fica dividida por colunas, facilitando a análise.

Partindo para a etapa de conferência e limpeza da base de dados, focando na análise de *text mining* que queremos aplicar, verificamos na literatura que a exploração de dados desse assunto comumente parte da identificação de palavras-chave e listagem de termos mais frequentes. Dependendo do foco do trabalho, podemos querer que termos com grafias incorretas sejam computados na mesma contagem dos termos com grafia correta, ou podemos querer ignorar palavras de pouco interesse. Para esses casos, os dados não estruturados obtidos da rede social apresentam dificuldades, por ser um espaço em que a expressão de acordo com a língua culta não é mandatória.

Concluindo-se que queremos encontrar a frequência de palavras agrupando as diferentes grafias disponíveis, temos que propor um método para execução desse procedimento. Esse processo acaba dividido em duas abordagens aplicadas juntamente: a listagem de termos equivalentes, de acordo com o conhecimento da norma culta, da forma de escrita na rede social, e do conhecimento específico do assunto e do público de interesse; e a listagem de termos equivalentes de acordo com a análise exploratória de dados. Ambos exigem a conferência manual do texto, uma vez que um algoritmo padrão não é capaz de captar todas as nuances de variação dos dados possível, e apesar de dispendioso é necessário para termos uma base de dados de boa qualidadade.

Conclusões

O pacote*RFacebook*apresenta um conjunto de funções úteis para a extração de dados, embora todas as etapas necessárias para o funcionamento dos comandos do pacote não estejam descritas no documento que descreve sua utilização. Ele também se limita a extração de dados, não apresentando alternativas para análise dos mesmos, e sem corrigir qualquer dado necessário para correta análise posterior. Também as funções não podem contornar as configurações de confidencialidade da rede social, que não permite o acesso às informações de seus usuários sem autorização dos mesmos, o que limita a utilização das informações.

Referências

[1] BARBERA, P. *Documentação do pacote 'RFacebook'*. Disponível em: <cran.r-project.org/web/packages/Rfacebook> Acesso em: 10 de outubro de 2015.

[2] Francis, L; Flynn, M. Text Mining Handbook. *Casualty Actuarial Society E-Forum*, Spring 2010 61

[3] Vários autores. Seção de dúvidas frequentes para desenvolvedores de apps no Facebook.

Avaliação do Risco de Crédito: Modelos de Regressão Logística com amostras de diferentes proporções

Mariana Nolde Pacheco [1]

Lisiane Priscila Roldão Selau [2]

Resumo: O objetivo do estudo é propor um modelo de risco de crédito, utilizando regressão logística binária, em que foram propostas duas amostragens de clientes: com proporção igual e desigual de clientes bons e maus, bem como diferentes pontos de corte para classificação desses clientes. O estudo foi realizado através da avaliação de dados reais de concessão de crédito. Nos modelos propostos pelo estudo, os valores do teste KS e da área abaixo da curva ROC foram muito semelhantes em ambas as amostras (igual e proporcional), bem como o percentual de acerto geral, com valores em torno de 60% de acerto para todos os pontos de corte. Já a avaliação da classificação das categorias (bons e maus), evidenciou resultados diferentes nas categorias. O estudo mostrou que ao aumentar o ponto de corte, melhoramos a classificação dos maus clientes, mas pioramos o percentual de acerto dos bons clientes. Dessa forma, os resultados do estudo sugerem que a proporção das categorias de clientes da amostra, bem como os pontos de corte da classificação do modelo devem ser considerados de acordo com o objetivo de classificação da empresa.

Palavras-chave: *Credit Scoring, Balanceamento Amostral, Regressão Logística.*

1 Introdução

No Brasil, com o crescente desenvolvimento econômico dos últimos anos, houve um aumento da demanda e da concessão de crédito à população. Segundo Brito (2008), o conceito de crédito consiste na atividade de colocar um valor à disposição de um tomador de recursos, com o compromisso do pagamento do valor emprestado em um determinado período de tempo previamente estabelecido. Dessa forma, a concessão de crédito envolve diversos riscos, uma vez que há a possibilidade de não cumprimento das obrigações financeiras estabelecidas.

De acordo com Steiner et al. (1999), a análise correta da concessão de crédito é essencial para a sobrevivência das instituições financeiras, pois um erro na decisão de conceder o crédito pode significar um grande prejuízo financeiro dentro de uma única operação, gerando a perda do ganho obtido em outras diversas operações bem-sucedidas. Dessa forma, tem aumentado a necessidade de as empresas buscarem diferentes formas de identificar e diferenciar o bom e o mau pagador gerando,

[1] UFRGS - Universidade Federal do Rio Grande do Sul. Email: marinolde*@yahoo.com.br*
[2] UFRGS - Universidade Federal do Rio Grande do Sul. Email: lisianeselau*@gmail.com*

consequentemente, uma minimização do prejuízo obtido com transações malsucedidas bem como um acréscimo na rentabilidade da instituição.

As empresas utilizam diferentes técnicas para concessão de crédito aos seus clientes. Geralmente a avaliação é realizada através do uso de uma variedade de informações vindas de diferentes fontes, como os dados cadastrais do cliente na instituição. Embora os gestores consigam muitas vezes identificar fatores que diferenciam o bom e o mau pagador, esses critérios geralmente são subjetivos e errôneos, gerando prejuízos financeiros e morais não somente para as instituições como também aos clientes. Sendo assim, surge a necessidade da substituição desses critérios de avaliação subjetivos pelo uso de técnicas quantitativas que melhorem a tomada de decisão das empresas, não apenas diminuindo a concessão de crédito aos maus pagadores como também aumentando o crédito aos potenciais bons pagadores (SELAU, 2011).

Os modelos de risco de crédito, também conhecidos como *Credit Scoring*, são ferramentas de avaliação para classificação dos clientes, e apresentam como principal objetivo identificar previamente o bom e o mau pagador evitando transações financeiras equivocadas. Uma das técnicas utilizadas para avaliação do risco de crédito é a regressão logística, que analisa o efeito de uma ou mais variáveis explicativas (categóricas ou métricas) sobre uma variável resposta binária, que nesse estudo é o tipo de cliente (bom e mau). A regressão logística atribui diferentes pesos para cada uma das variáveis explicativas do modelo que juntas fornecem a probabilidade de o cliente pertencer ao grupo de interesse (HOSMER; LEMESHOW, 2013).

Nas instituições financeiras é comum observar um desbalanceamento na proporção de bons e maus pagadores. Como consequência, há um prejuízo no desenvolvimento de modelos estatísticos, visto que as proporções diferentes de bons e maus pagadores podem influenciar nas variáveis preditoras do modelo, ocasionando erros de classificação. Muitos estudos já utilizaram a comparação de diferentes técnicas estatísticas para análise dos dados de crédito, porém poucos estudos avaliam o impacto do desequilíbrio amostral na correta classificação e predição do modelo (BROWN; MUES, 2012). Dessa forma, o objetivo desse estudo é identificar o impacto do desbalanceamento amostral na predição do risco de crédito por meio da utilização de um modelo de regressão logística.

2 Método

O estudo utilizou uma base de dados reais de uma instituição financeira, com informações do tipo: sexo, idade, escolaridade, profissão, etc. Tendo em vista a comparação dos modelos propostos, foram criadas duas composições amostrais com diferentes percentuais de bons e maus clientes. Primeiramente, foram tomadas amostras de análise e validação na proporção real de bons e maus pagadores da instituição. No segundo momento, foram criadas amostras de análise e validação com a

mesma proporção de clientes nas amostras (50% bons e 50% maus) e uma amostra de teste com as proporções reais do negócio para comparação com o primeiro modelo.

A construção dos modelos seguiu uma adaptação da sistemática proposta por Selau (2011), cujas etapas para desenvolvimento são: delimitação da população, seleção da amostra, análise preliminar, construção dos modelos, avaliação dos modelos. Para avaliação da qualidade dos modelos, foi observado o percentual de acerto de classificação, o valor de teste de Kolmogorov-Smirnov (KS), que revela a correta separação entre os grupos de bons e maus pagadores, e a área abaixo da curva ROC, que se baseia na sensibilidade do modelo em identificar os maus pagadores e na especificidade, que é a correta identificação de bons pagadores.

3 Resultados

Os modelos construídos foram avaliados utilizando a área abaixo da curva ROC e o valor do teste KS. As medidas foram realizadas tanto para as amostras de análise e validação do modelo de grupos proporcionais, quanto para as amostras de análise, validação e teste do modelo de grupos iguais, com a mesma proporção de bons e maus clientes. Os resultados obtidos foram bastante semelhantes nos dois modelos construídos e em todas as amostras, como pode ser observado na Tabela 1.

Tabela1. Avaliação dos modelos – Curva ROC e Teste KS.

Avaliação dos modelos – Teste KS e Curva ROC					
	Grupos Iguais			Grupos Proporcionais	
Avaliação	Análise	Validação	Teste	Análise	Validação
ROC	0,709	0,699	0,689	0,706	0,705
KS	30,26	30,11	27,18	30,70	30,04

O ponto de corte inicial para separação dos clientes bons e maus foi de 0,5. Dessa forma, clientes com probabilidade acima de 0,5 eram classificados como bons e abaixo de 0,5 eram classificados como maus. Para avaliação dos modelos propostos, foram fixados diferentes pontos de corte para classificação dos clientes (variando de 0,5 até 0,7). A Tabela 2 apresenta a proporção de acerto nos dois modelos, tomando como base os diferentes pontos de corte para classificação dos clientes.

Tabela 2. Percentual de acerto nos modelos de grupos iguais e proporcionais de clientes

Percentual de acerto geral (%)					
	Grupos Proporcionais		Grupos Iguais		
Ponto de Corte	Análise	Validação	Análise	Validação	Teste
0,5	65,4	66,1	65,3	65,6	64,7
0,55	65,3	65,5	64,7	64,7	63,6
0,6	63,9	63,7	64,2	63,6	60,4
0,65	62,0	61,5	62,6	62,7	59,2
0,7	59,3	59,0	59,9	59,5	55,9

Após a primeira avaliação geral do percentual de acerto dos modelos construídos, realizou-se a avaliação do percentual de acerto de cada categoria de clientes (bons e maus) nos modelos de grupos proporcionais e iguais de clientes. A Tabela 3 apresenta os resultados obtidos de percentual de acerto por categoria de cliente no modelo de grupos proporcionais de clientes, por meio do uso de diferentes pontos de corte para classificação dos bons e maus.

Tabela 3. Percentual de acerto das categorias no modelo de grupos proporcionais de clientes.

Percentual de acerto por categoria (%)				
Modelo de Grupos Proporcionais				
	Análise		Validação	
Ponto de Corte	Bons	Maus	Bons	Maus
0,5	75,7	50,9	80,2	46,7
0,55	66,7	63,2	66,0	63,2
0,6	57,1	73,5	59,4	70,0
0,65	49,2	80,3	47,2	79,4
0,7	40,2	86,4	38,7	85,9

Da mesma forma, como realizou-se a avaliação das categorias de clientes bons e maus no modelo de grupos proporcionais, avaliou-se também os percentuais de classificação correta das categorias no modelo de grupos iguais de clientes, cujo resultado pode ser visto na Tabela 4.

Tabela 4. Percentual de acerto das categorias no modelo de grupos iguais de clientes

Percentual de acerto por categoria (%)						
Modelo de Grupos Iguais						
	Análise		Validação		Teste	
Ponto de Corte	Bons	Maus	Bons	Maus	Bons	Maus
0,5	60,9	69,6	61,5	69,0	58,7	67,1
0,55	53,1	76,6	53,1	77,0	50,8	76,6
0,6	44,1	83,2	44,0	82,3	42,9	83,6
0,65	35,7	88,5	36,2	86,9	34,3	88,5
0,7	27,6	92,5	28,6	90,7	25,6	92,4

4 Discussão e Conclusões

Com da análise dos resultados obtidos com o modelo de regressão logística, pode-se observar que os valores do teste KS e da área abaixo da curva ROC foram muito semelhantes em ambos os modelos de grupos proporcionais ou iguais, o que sugere que os modelos conseguiram separar adequadamente os grupos de bons e maus clientes da empresa.

Os percentuais de acerto geral nos dois modelos construídos foram bastante semelhantes, com valores em torno de um acerto de 60%. Por meio da avaliação do percentual geral nos diferentes

pontos de corte observou-se que à medida que o ponto de corte aumenta, ocorre um percentual de acerto geral diminuído, o que gera uma piora na previsão do modelo proposto.

Já a avaliação do percentual de acerto das categorias de clientes (bons e maus) apresentam valores bem diferenciados de acordo com o ponto de corte. Embora de forma geral os modelos apresentem um percentual de acerto por volta dos 60% para ambos os pontos de corte, a avaliação do percentual de clientes por categoria sugere resultados diferentes. Observa-se que à medida que se aumenta o ponto de corte, há uma melhora na classificação dos maus clientes, mas há uma piora no percentual de acerto dos bons clientes. Da mesma forma, observa-se que no modelo de grupos iguais, a classificação inicial dos clientes foi feita de maneira equilibrada e assim que se aumenta o ponto de corte, o modelo foi melhorando a classificação dos maus clientes e piorando a classificação dos bons. Já o modelo de grupos proporcionais iniciou classificando melhor os clientes que estavam em maior proporção na amostra, os bons clientes, e à medida que o ponto de corte aumentou, passou a classificar melhor os maus clientes.

Os resultados sugerem que a proporção das categorias de clientes da amostra para a construções do modelo, bem como os pontos de corte da classificação devem ser considerados de acordo com o objetivo da empresa. Caso haja uma necessidade de classificação melhor dos maus clientes, tendo em vista que os mesmos causam um prejuízo maior para a empresa do que o lucro obtido com bons clientes, então o ponto de corte do modelo de regressão logística deve ser maior. Nesse sentido, uma sugestão de trabalho futuro é avaliar a construção de um modelo em que se tenham grupos com proporcionalidade maior de maus clientes.

Referências

BRITO, G. A. S.; NETO, A. A. Modelo de Classificação de Risco de Crédito de Empresas. *Revista Contabilidade e Finanças*, São Paulo, v. 19, n. 46, p. 18-29, 2008.

BROWN, I., MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with Applications*, v. 39, n. 3, p. 3446-3453, 2012.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. New York: Wiley, 3.ed., 2013.

SELAU, L. P. R.; RIBEIRO, J. L. D. Systematic Approach to Construct Credit Risk Forecast Models. *Pesquisa Operacional*, v.31, n.1, 2011.

STEINER, M. T. A.; CARNIERI, C.; KOPITTKE, B. H.; STEINER NETO, P. J. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. *Revista de Administração*, São Paulo, v. 34, n. 3, p. 56-67, 1999.

Análise crítica dos resultados de uma análise especial de varredura – Aplicação no estudo do herpesvírus bovino tipo 1 (BoHV-1) no Rio Grande do Sul, Brasil

G.S. Silva[1]
A.M. Perez[2]
L.G. Corbellini[1]

Resumo: O Rio Grande do Sul (RS) é responsável por 12% do total de leite produzido no Brasil, sendo o segundo maior produtor com um rebanho de aproximadamente 5,8 milhões de bovinos (6,5% do rebanho nacional de bovinos leiteiros). Esta perspectiva demanda pesquisas com objetivo de promover o progresso sanitário e à formação de recursos humanos por meio de estudos epidemiológicos que suportem programas de controle das enfermidades. O herpesvirus bovino tipo 1 (BoHV-1) é um importante patógeno na bovinocultura mundial. Sua infecção leva a perdas econômicas significativas principalmente devido a falhas reprodutivas e abortamentos. A identificação de rebanhos infectados pode ser baseada na detecção de anticorpos específicos anti-BoHV em amostras de soro ou leite, em adição a isso, a análise espacial pode ser usada como ferramenta complementar para a detecção de agregados ou áreas de maior ocorrência de doenças. O objetivo deste estudo foi avaliar a validade dos resultados gerados por uma análise do teste de varredura espacial para compreender melhor a epidemiologia do agente de transmissão entre rebanhos e estimar a prevalência do agente no Estado do Rio Grande do Sul através de uma amostragem planejada.
Palavras-chave: SaTScan, análise estatística de varredura, hespesvírus bovino tipo I.

Introdução: O Rio Grande do Sul é o segundo maior produtor de lácteos do Brasil, sendo o maior produtor da região sul, com registros de aumentos na produção ao longo do tempo. Dados recentes publicados no Relatório Socioeconômico da Cadeia Produtiva do Leite no Rio Grande do Sul contabilizam que a produção leiteira no estado, cujo rebanho é de 1.427.730 vacas, gera 9,13% do PIB gaúcho em quase 199.000 propriedades. No total, o Rio Grande do Sul produz mais de 11,5 milhões de litros de leite por dia, mas a capacidade industrial instalada é de 18,5 milhões de litros de leite/dia, o que demonstra potencial para ampliação de produtos lácteos processados. O herpesvírus bovino tipo 1 (BoHV-1) é um alfa herpesvírus que acomete os bovinos, principalmente os animais de produção, e provoca diversas síndromes incluindo doenças respiratórias, abortos e distúrbios genitais. O vírus tem um caráter de transmissão definido como portador assintomático, que passa despercebido aos olhos dos produtores pela apresentação subclínica da doença, o que ajuda na disseminação dentro e entre rebanhos. Como esse agente pode causar grande impacto comercial para as propriedades leiteiras, a identificação dos rebanhos positivos é de suma importância para elaboração de estratégias de controle e prevenção de futuras perdas. Uma das técnicas rápidas e eficazes para identificação de rebanhos positivos é a

1 EPILAB, Dep. Med. Vet. Preventiva, UFRGS, Brasil. Primeiro autor: gustavossvet@hotmail.com
2 Dep. de Med. Vet. Populacional, Universidade de Minnesota, EUA.

identificação de anticorpos específicos anti-BoHV via amostras de *pool* de leite, retiradas do tanque da propriedade, através da técnica de ELISA indireto. Além disso, a análise espacial de varredura pode ser usada para a detecção de agregados de rebanhos positivos ocorridos em determinados locais devido a transmissão do agente nas vizinhanças. Esta técnica vem sendo amplamente utilizada para avaliação de áreas onde o número da ocorrência (casos) das doenças está acima dos valores esperados, além de servir como ferramenta utilizada na investigação epidemiológica humana e animal, pois esses modelos possibilitam ajustes não só para as varreduras espaciais, mas também utilizar as informações espaço-temporais dos casos.

Materiais e métodos: Os dados utilizados nas simulações foram de um estudo transversal em rebanhos leiteiros no estado do Rio Grande do Sul (Brasil) para avaliar a prevalência de anticorpos contra BoHV-1 em amostras de leite do tanque por teste ELISA. Foram amostrados 388 rebanhos (unidade primária de interesse) estratificados por mesorregião do Estado. O cálculo da amostra considerou todas as propriedades leiteiras do RS, sua localização dentro das mesorregiões e foi utilizada uma precisão absoluta de 5%, prevalência esperada de 50% e nível de confiança de 95%. Amostras de leite do tanque de resfriamento foram coletadas para a determinação dos níveis de anticorpos anti-BoHV 1 através de teste ELISA comercial. Devido as características da infecção por herpesvírus, uma amostra positiva significa que o rebanho apresenta infecção ativa. O SaTScan 9.4 foi utilizado para detectar as áreas com maior probabilidade de ocorrência de infecção através de um modelo de Bernoulli. O método de estatística de varredura é utilizado para detectar e avaliar grupos de casos em qualquer ambiente puramente temporal, puramente espacial ou espaço-temporal. Para cada localidade e tamanho da janela de varredura, a hipótese alternativa é de que os casos não estão distribuídos aleatoriamente e o número de casos dentro da janela possui um valor maior que o número de casos esperados. Para o modelo de Bernoulli utilizado aqui, a função de verossimilhança de uma janela específica é dada por:

$$I = (c/n)^{c} \, (n\text{-}c/n)^{c} \, (C\text{-}c/N\text{-}n)^{C\text{-}c} \, (c/n)^{(N\text{-}n)\text{-}(C\text{-}c)} \qquad \text{(I)}$$

Onde *C* é o número total de casos, *c* é o número de casos em uma determinada janela de varredura, *n* é o número total de casos e controles dentro da janela, *N* é número total de casos e controles do banco de dados e *I* ( ) é a indicação de função (I).

Para avaliar a sensibilidade da varredura, a população em risco foi fixada em 50% e a validade (ou seja, valor de *p*) foi avaliada pela mudança do tamanho máximo do cluster espacial (janela de varredura) usando 5, 10, 20, 50, 100, 150 km e o padrão do programa (*default*).

Resultados e discussões: A prevalência encontrada no rebanho foi de 48% (IC 95% = 43-53; deff = 0,9) e observou-se uma forte diferença entre os clusters gerados nas análises de verificação. O tamanho dos clusters variou entre 8,71-147,3 km de raio (média de 53,55); os casos observados variaram 8-109 média de 30 casos e a proporção de casos observados e esperados (O / E) foram 1,32-2,09 com uma média de 1,89 (Tabela 01). O *default* e o cluster de maior tamanho de varredura (150 km) resultou em menor

detecção de áreas de alto risco (p <0,05) e tendo essas áreas com um grande raio (122,33 e 147,69Km), enquanto usando 20 e 50 km, as análises resultaram em uma maior detecção de áreas de risco com um menor raio de abrangência e estas apresentando uma menor granularidade (Figura 01). Os resultados mostraram uma alta variação na definição de áreas de risco. Assim, quando diminuir o tamanho máximo potencial de um cluster pode aumentar a variação na relação O/E (casos observados/esperados) de cada grupo e quando o tamanho do potencial de agregados é aumentado, estes tendem a ter menor quantidade de aglomerados com uma forma mais homogénea e inferior na relação O/E, o que torna difícil interpretar os resultados e escolher a melhor opção para tomada de decisão.

Conclusão: É de suma importância no atual cenário econômico e profissional, a aplicação de medidas preventivas direcionadas a área de maior probabilidade de ocorrência, visando de modo eficaz mobilizar as práticas de controle, efetivo para realizar tais medidas e recursos financeiros do sistema, agindo de maneira mais sábia nas tomadas de decisão. Desta forma os agentes que tomam decisões devem levar em conta qual a melhor estratégia para cada doença do ponto de vista dos objetivos dos programas sanitários, visto que as doenças possuem caraterísticas epidemiológicas distintas.

Referências:

Instituto Brasileiro de Geografia e Estatística. IBGE. Pesquisa Pecuária Municipal (PPM), 2013. Disponível em: <http://www.sidra.ibge.gov.br/bda/acervo/acervo9.asp?ti=1&tf=99999&e=c&p=PP&z=t&o=24>. Acesso: 22 set. 2015. 2013.

Instituto gaúcho do leite. IGL RS. 2015. Disponível em: <http://www.iglrs.com.br/>. Acesso: 16 set. 2015.

RAAPERI, K.; ORRO T.; VILTROP, A. *Epidemiology and control of bovine herpesvirus 1 infection in Europe*. The Veterinary Journal. v201, Issue 3, September 2014, Pages 249–256. 2014.

KULLDORFF, M. *A spatial scan statistic*. Communications in Statistics: Theory and Methods. 26:1481-1496. 1997

KULLDORFF, M.; NAGARWALLA, N. *Spatial disease clusters: Detection and inference.* Statistics in Medicine. 14:799-810. 1995.

SaTScan - User Guide v9.4. Disponível em: < http://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download>. Acesso: 08 out 2015.

Tabela 01 – Resultados estatístico da análise de sensibilidade alterando a janela de varredura do modelo

Tamanho máx. da janela	Sig. Clusters	Cluster	Pop	Raio (km)	Casos Obs.	Casos Esp.	O/E	Log Likelihood	% Casos na área	p-valor
5	0									
10	1	**1**	**8**	**8,71**	**8**	**3,83**	**2,09**	**5,97**	**100,0%**	**0,03**
20	3	**1**	**13**	**19,68**	**13**	**6,23**	**2,09**	**9,81**	**100,0%**	**0,0039**
		2	12	18,39	12	5,75	2,09	9,04	100,0%	0,015
		3	11	12,59	11	5,27	2,09	8,27	100%	0,033
50	4	**1**	**46**	**40,04**	**37**	**22,05**	**1,68**	**11,70**	**80,4%**	**0,0021**
		2	24	41,23	22	11,5	1,91	11,21	91,7%	0,0032
		3	14	29,05	14	6,71	2,09	10,58	100%	0,0057
		4	12	18,39	12	5,75	2,09	9,04	100,0%	0,036
100	4	**1**	**46**	**40,04**	**37**	**22,05**	**1,68**	**11,67**	**80,4%**	**0,0029**
		2	24	41,23	22	11,5	1,91	11,21	91,7%	0,0044
		3	14	29,05	14	6,71	2,09	10,58	100%	0,01
		4	12	18,39	12	5,75	2,09	9,04	100,0%	0,046
150	2	**1**	**173**	**122,33**	**109**	**82,91**	**1,31**	**14,45**	**63,0%**	**0,00026**
		2	27	147,69	24	12,94	1,85	10,87	88,9%	0,0064
default	2	**1**	**173**	**122,33**	**109**	**82,91**	**1,31**	**14,45**	**63,0%**	**0,00027**
		2	27	147,69	24	12,94	1,85	10,87	88,9%	0,0067

Figura 01 – Outputs da análise especial de varredura.