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

Anais

IV SEMANÍSTICA

IV Semana Acadêmica do Departamento de Estatística

da UFRGS

http://www.ufrgs.br/semanistica

Porto Alegre - 20 a 22 de outubro de 2014

Organização:

Promoção:

Conteúdo

1 Cartaz da IV SEMANÍSTICA

2 Introdução

A IV Semana Acadêmica da Estatística (IV SEMANÍSTICA) será realizada de 20 a 22 de outubro de 2014, no Instituto de Matemática, Campus do Vale da UFRGS, Porto Alegre, RS. O evento engloba os mais variados temas dentro da área acadêmica e profissional. Entretanto, para esta edição, o assunto proposto para as palestras é BIG DATA.

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 IV SEMANÍSTICA (Semana Acadêmica da Estatística 2014) ver

i) http://www.worldofstatistics.org - The World of Statistics

ii) www.ufrgs.br/semanistica - IV SEMANÍSTICA (Semana Acadêmica da Estatística 2014)

3 Agradecimentos

A IV 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
- IM-UFRGS - Instituto de Matemática da UFRGS
- PROREXT-UFRGS - Pró-Reitoria de Extensão da UFRGS
- PROPESQ-UFRGS - Pró-Reitoria de Pesquisa da UFRGS
- StatSoft South America

A Comissão Organizadora da IV 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)

Álvaro Vigo (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)

Álvaro Vigo (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

Angélica Segala (Curso de Estatística-UFRGS)

Bruna Martini Dalmoro (Curso de Estatística-UFRGS)

Camila Thaís Weber (Curso de Estatística-UFRGS)

Gabriel da Cunha (Curso de Estatística-UFRGS)

Raiane Padilha Silveira (Curso de Estatística-UFRGS)

Renata Fragoso Máximo (Curso de Estatística-UFRGS)

7 Apresentação

O programa da IV 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;
- 1 Minicurso relacionado ao tema Data Mining;
- Comunicações orais apresentadas pelos participantes do evento;

8 Programação

Horário	20/10/2014	21/10/2014	22/10/2014
	Segunda-Feira	Terça-Feira	Quarta-Feira
08:30-09:15	C1	C3	Minicurso
09:15-10:00	C2		
10:00-10:30	Coffee - Break	Coffee - Break	Coffee - Break
10:30-11:15	Minicurso	C4	C6
11:15-12:00		C5	Apresentação Oral

Minicurso: *Data Mining*

Ministrantes:

Dra. Taiane Schaedler Prass

Conferências:

(C1) Conferência 1 - Profª. Drª. Linda Lee Ho - POLI - USP

Título: Monitoramento da média e variância de processo através de gráficos de controle por atributos

(C2) Conferência 2 - Profª. Drª. Gabriela Cybis – DEST - UFRGS

Título: Vigilância global da gripe e a filodinâmica Bayesiana

(C3) Conferência 3 - Josias Oliveira – StatSoft South America

Título: Big Data – Tutorial e Aplicações

(C4) Conferência 4 – Prof. Dr. Osvaldo Sergio Farhat de Carvalho – DCC - UFMG

Título: O Desenvolvimento do InfoSAS, um sistema para detecção de anomalias na produção do SUS

(C5) Conferência 5 - Prof. Dr. Lori Viali – DEST - UFRGS

Título: A estatística, a tecnologia e o mercado de trabalho

(C6) Conferência 6 - Prof. Dr. Flávio Augusto Ziegelmann - DEST - UFRGS

Título: Algumas abordagens para modelar séries temporais de altas dimensões

9 Minicurso

Data Mining

Drª. Taiane Schaedler Prass
StatSoft South America

Resumo

Data Mining é um processo analítico projetado para explorar grandes quantidades de dados em busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados. Neste minicurso abordaremos os principais conceitos relacionados a data mining. Apresentaremos uma discussão de quando devemos considerar a utilização de técnicas de data mining; quais as características de uma boa ferramenta de data mining; principais modelos para data mining; passos para a realização da técnica, entre outros. Dedicaremos especial atenção ao Text Mining, apresentando alguns exemplos práticos de aplicação da técnica.

10 Conferências

Conferência 1

Monitoramento da média e variância de processo através de gráficos de controle por atributos

Profª. Drª. Linda Lee Ho
Escola Politécnica da Universidade de São Paulo - USP

Resumo

Tradicionalmente para monitorar a média e a variância de um processo são utilizados, respectivamente os gráficos de controle $\overline{X}$ e S^2. Recentemente, gráficos de controle por atributos tem sido propostos para monitorar estas duas grandezas. Um dispositivo do tipo PASSA/NÃO PASSA é empregado e cada elemento da amostra é classificado como aprovado ou reprovado segundo um critério. Se o número de reprovações for superior a um limite de controle, é declarado que o processo está fora de controle. Várias vantagens em adotar estes gráficos: I) a inspeção é mais rápida; barata e evita desperdício de material em casos de ensaios destrutivos; II) não é necessário medir os elementos da amostra. É possível calibrar estes gráficos por atributo para ter um desempenho igual aos gráficos tradicionalmente empregados (em termos de ARL). Exemplos práticos ilustram a proposta apresentada.

Conferência 2

Vigilância global da gripe e a filodinâmica Bayesiana

Profª. Drª. Gabriela Cybis
Departamento de Estatística - UFRGS

Resumo

Gastos globais relacionados a gripe são da ordem de bilhões de dólares por ano. A vacina da gripe, produzida a cada ano, é uma importante ferramenta para mitigar o custo humano e financeiro. A produção da vacina inicia vários meses antes da temporada da gripe, e sua eficácia depende da adequação as variedades do vírus que irão circular nessa temporada. Eu apresentarei alguns métodos de inferência Bayesiana que integram dados geográficos, imunológicos e sequências de RNA viral para estudar a evolução da epidemia de gripe ano a ano, gerando informações relevantes para o design das vacinas. Alguns desafios da construção desses métodos incluem a incorporação eficiente de informações realistas sobre o processo biológico no modelo probabilístico, conjuntos de dados de alta dimensionalidade, e eficiência computacional.

Conferência 3

Big Data – Tutorial e Aplicações

Josias Oliveira
StatSoft South America

Resumo

a. O que é Big Data?

b. Tecnologias para Big Data

c. Técnicas para explorar o universo Big Data

d. Aplicações

No futuro, qualquer decisão de negócio poderá ser tomada apoiada 100% em informações, apenas usando técnicas, tecnologias e metodologias Big Data. O crescente uso de devices, o desenvolvimento de novas tecnologias de coleta e armazenamento, softwares capazes de observar um evento a partir de dezenas de milhares de dados e a capacidade humana de interpretar tudo isso mudará nossa experiência de gestão do tempo, da produtividade, dos lucros, de pessoas etc. O fato essencial nesse vertiginoso desenvolvimento tecnológico é que a Cultura Analítica do mundo dos negócios já está sofrendo a maior mudança da sua história.
Literatura:

i. Big Data Glossary: Warden, Pete, O'Reilly Ed.

ii. Understanding Big Data, Zikopoulos, Paul; Eaton, Chris; MacGraw-Hill.

iii. The little book of Big Data, Burlingame, Noreen.

Aplicações:

i. Case UOL-PagSeguro: Fraude no e-commerce;

ii. Case AVON: Estratégia de Venda Direta;

iii. Case VALE: Inovar no Processo de Pelotização

Conferência 4

O Desenvolvimento do InfoSAS, um sistema para detecção de anomalias na produção do SUS

Prof. Dr. Osvaldo Sergio Farhat de Carvalho
Departamento de Ciência da Computação - UFMG

Resumo

Nós apresentamos a experiência da UFMG com o desenvolvimento do InfoSAS, um sistema para detecção de anomalias em dados de produção do SUS. Serão abordados aspectos como o tratamento de um grande volume de dados, com problemas de qualidade, e com restrições severas de confidencialidade, a forma de trabalho de uma equipe multidisciplinar com desenvolvedores, estatísticos, cientistas da computação e médicos sanitaristas, a harmonização entre linguagens de programação e modelos de desenvolvimento de software da cultura destas comunidades, e o atendimento a restrições técnicas do Ministério da Saúde.

Conferência 5

A estatística, a tecnologia e o mercado de trabalho

Prof. Dr. Lori Viali
Departamento de Estatística - UFRGS

Resumo

A estatística moderna não pode mais prescindir da tecnologia e esta muda de rumos de acordo com o mercado. A palestra faz um levantamento sobre os principais cargos disponíveis no mercado para egressos das áreas de exatas, principalmente estatísticos e dos pré-requisitos que um egresso precisa ter para poder ocupar estes cargos e os recursos tecnológicos que ele precisa conhecer (dominar).

Conferência 6

Estatística na área de pesquisa médica

Prof. Dr. Flávio Augusto Ziegelmann
Departamento de Estatística - UFRGS

Resumo

Atualmente, com o rápido desenvolvimento computacional e a alta capacidade de armazenamento de grande quantidade de informações, tem havido uma demanda por novos modelos estatísticos capazes de descrever bem um grande número de variáveis. Em particular, no estudo de séries temporais acontece o mesmo. Aqui, serão motivados e apresentados alguns modelos e procedimentos de estimação capazes de descrever a dinâmica de várias séries de tempo simultaneamente. Abordaremos os seguintes tópicos: cópulas dinâmicas (especificamente "vine and factor copulas"), séries temporais funcionais, modelos MIDAS e método LASSO.

11 Seções de Comunicações

Comunicações Orais

O Processo Estocástico k-Factor GARMA, Um Estudo de Simulação e Aplicação em Tempestades Solares

Ian Meneghel Danilevicz[1] [3]
Cleber Bisognin[2] [3]

Resumo: Estudo teórico e aplicado do processo estocástico k-*factor* GARMA(p,**u**,**λ**,q). Considerações teóricas, apresentação de resultados de simulações e resultados de aplicação a dados reais. Proposição de diferentes estimadores para os parâmetros p,**u**,**λ**,q e análise do comportamento desses estimadores para diferentes cenários, nos quais o comprimento de k e a quantidade de *outliers* aditivos foram testados de diversas formas. A série real analisada foi a famosa sunspot.year (disponível no software estatístico R), nela foram aplicados os estimadores propostos, observado a qualidade dos resíduos e proposto um modelo candidato para previsões.

Palavras-chave: k-*factor* GARMA, longa dependência, *outliers*, processo estocástico.

1 Introdução

No estudo de séries temporais tem crescido o interesse por processos de longa dependência (Bisognin), assim como pela identificação de outliers (Tsay). Apresentamos o modelo k-*factor* GARMA, uma generalização dos processos SARFIMA, com a propriedade da longa dependência. Apresentamos algumas considerações teóricas, resultados de simulação de estimadores e uma aplicação em uma série temporal real.

Definição de k-*factor* GARMA

Seja Xt um processo estocástico que satisfaz a equação

$$\varphi(\beta)\prod_{j=1}^{k}\left(1-2u_j\beta+\beta^2\right)^{\lambda_j}(X_t-\mu)=\theta(\beta)\varepsilon_t$$

[1] UFRGS - Universidade Federal do Rio Grande do Sul. Email: iandanilevicz*@gmail.com*
[2] UFRGS - Universidade Federal do Rio Grande do Sul. Email: cbisognin*@ufrgs.br*
[3] Agradecimentos a bolsa de IC da PIC CAPES.

na qual k é um inteiro finito, $|u_j| \leq 1$ e λ_j é um número fracionário, para j=1, ...,k, μ é a média do processo, $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ é um processo de ruído branco e $\phi(.)$ e $\theta(.)$ são os polinômios de grau p e q, respectivamente.

Definição de *outlier* Aditivo

Seja Z_t um processo estocástico que satisfaz a equação

$$Z_t = X_t + V_t$$

na qual X_t é o processo k-*factor* GARMA visto acima e V_t é constituido de variáveis aleatórias independentes e identicamente distribuídas com distribuição dada por

$$H_v = (1 - c)\delta_0 + cG$$

para a qual $0 \leq c \leq 1$, δ_0 é uma distribuição degenerada na origem e G é uma distribuição arbitrária. Além disso, os processos V_t e X_t são independentes. Portanto, Z_t é um processo contaminado por misturas de AO (*Additive Outliers*).

2 Metodologia

Num primeiro momento, realizamos um estudo de simulação de Monte Carlo, no qual submetemos os estimadores a diferentes cenários e comparamos os resultados alcançados por cada um deles. Como estatísticas para avaliar o desempenho de cada um dos estimadores temos a média, o vício, o erro quadrático médio (EQM) e a variância.

Os cenários gerados foram séries temporais contaminadas por *outliers* aditivos (AO), ou seja, séries nas quais há observações com valores discrepantes (erro de aferição por exemplo), mas que não interferem nos resultados seguintes. Também diversificamos o percentual e a grandeza da contaminação na série com o intuito de perceber a robustez dos estimadores, ou seja, até que ponto os estimadores produzem estimativas com baixos valores de vício.

Simulamos séries de tamanho mil com mil e duzentas repetições, p igual a zero, q igual a zero e k igual a dois, pois nosso foco é concentrar esforços no estudo dos estimadores para u e λ. Dessa forma temos processos k-*factor* GARMA(p,λ,u,q) com $\lambda=\{\lambda_1, \lambda_2\}$ e $u=\{u_1, u_2\}$.

Estimadores Propostos

Direcionamos nosso estudo para os estimadores paramétricos, pois os estimadores semi-paramétricos não conseguem estimar todos os parâmetros. Os estimadores analisados são os respectivos:

1. **FT** - extenção do modelo proposto por Fox e Taqqu (1995) para os processos ARFIMA, o qual utiliza uma aproximação para a matriz de autocovariância sugerida por Whittle (1954) e a função periodograma;

2. **FTmod** - segue o estimador FT, mas a função a ser minimizada é calculada para mais frequências do que as de Fourier;

3. **MCMC** - utiliza o método de Metropolis-Hasting no estimador FT;

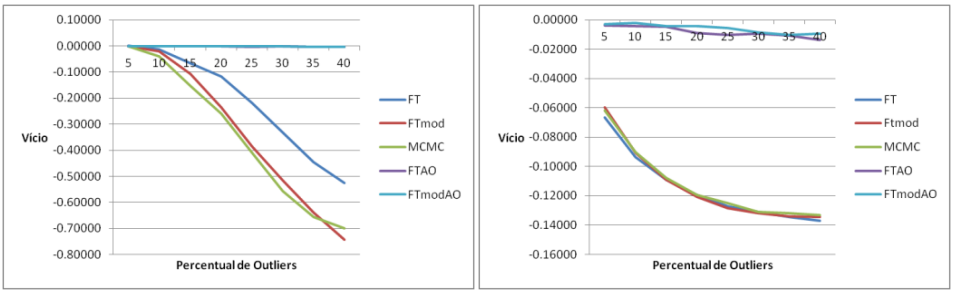
4. **FTAO** - análogo ao estimador FT, porém acrescenta-se uma função de densidade espectral capaz de estimar a magnitude e o percentual de *outliers* no processo;

5. **FTmodAO** - análogo ao estimador FTmod, porém acrescenta-se uma função espectral semelhante a executada no FTAO.

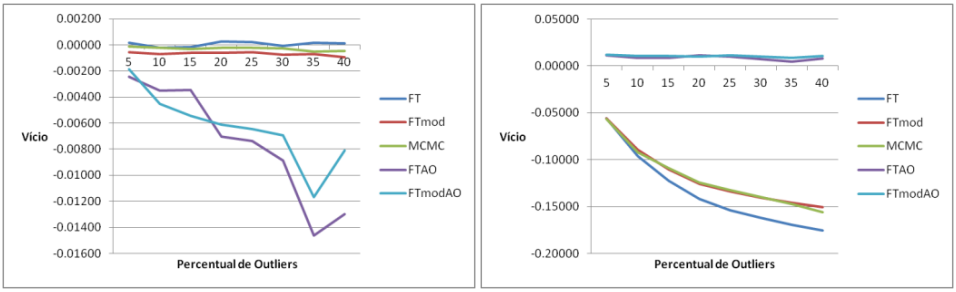
Em um segundo momento, passamos para o desafio de trabalhar com dados reais. Primeiramente analisamos a função periodograma da série temporal, afim de observar quantos picos a função apresenta no intervalo menor do que π. Esse procedimento nos indicará qual o provável melhor valor de k. Depois disso, fazemos um estudo de *outliers*, se há, se não há, qual a magnitude deles. Depois disso, propomos modelos e analisamos os resíduos, principalmente o gráfico de correlação e correlação parcial, para selecionar um modelo competitivo e se necessário acrescentar parâmetros φ e θ. Passado esse modelo estamos mais seguros para propor um modelo de previsão.

3 Resultados

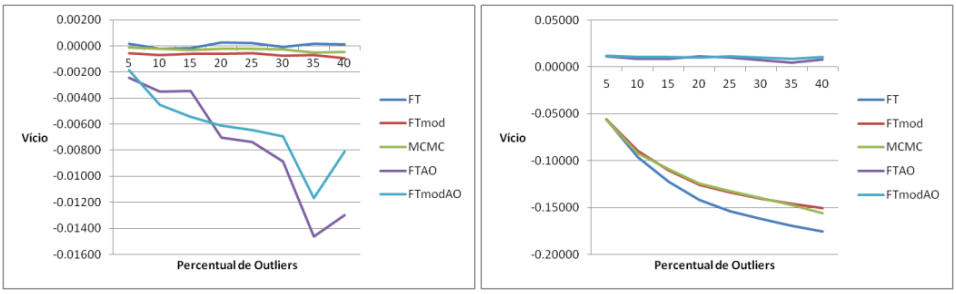
Simulamos séries com *outliers* com variância da variável de contaminação $\tau 2$ igual a 2, 5 e 10. e probabilidades de Contaminação c={0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40}. Para economizar espaço, mostramos apenas os resultados para τ^2 igual a cinco, mas o comportamento dos estimadores é semelhante. Os primeiros gráficos são dos vícios dos estimadores e na próxima página estão os respectivos EQMs.



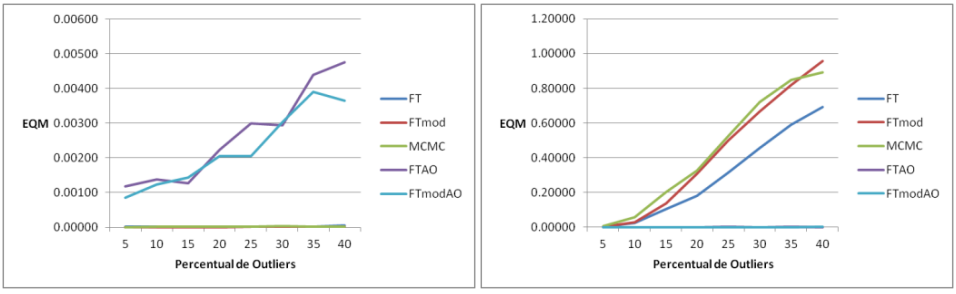
vício u_1 vício u_2



vício λ_1 vício λ_2



EQM u_1 EQM u_2



EQM λ_1 EQM λ_2

Figura 1: Resultados de Simulação

Além dos resultados de simulação (Figura 1) temos os resultados para séries reais. Primeiramente as questões mais genéricas da série temporal *sunspot* (Figura 2), ou seja, uma identificação de *outliers* e um periodograma suavizado.

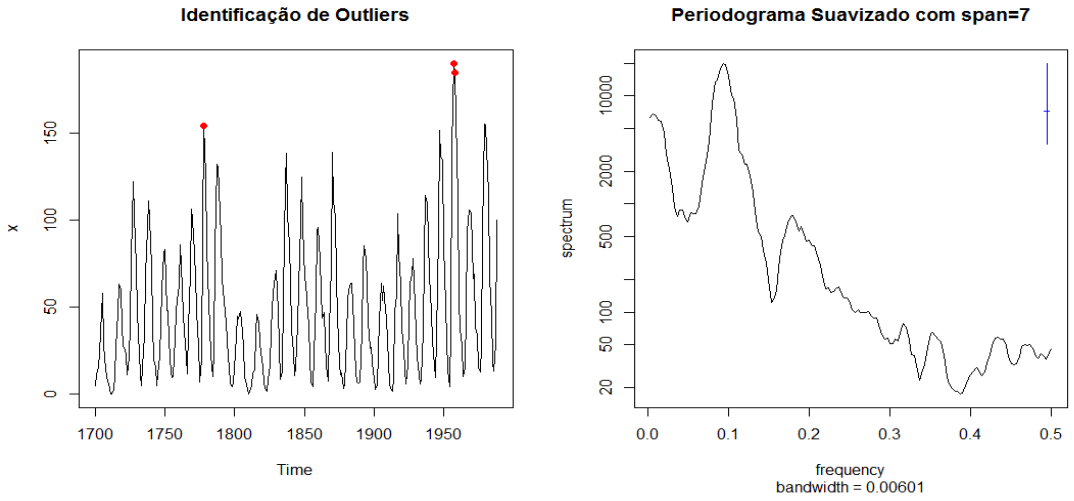


Figura 2: Série Real: *sunspot.*

Testamos os cinco tipos de estimadores para um modelo com k=3 e p=0=q. Uma vez que os resíduos (ver Figura 3) não estavam muito bons acrescentamos parâmetros p, chegando em p=3, o qual produziu resultados minimamente satisfatórios de resíduos. Embora nem todos os *lags* estejam dentro dos intervalos de confiança, o preço a ser pago para ajustá-los seria demasiado caro, ou seja, teríamos que colocar muitos φs e/ou θs no modelo. O que implica em aumentar de forma expressiva o erro padrão.

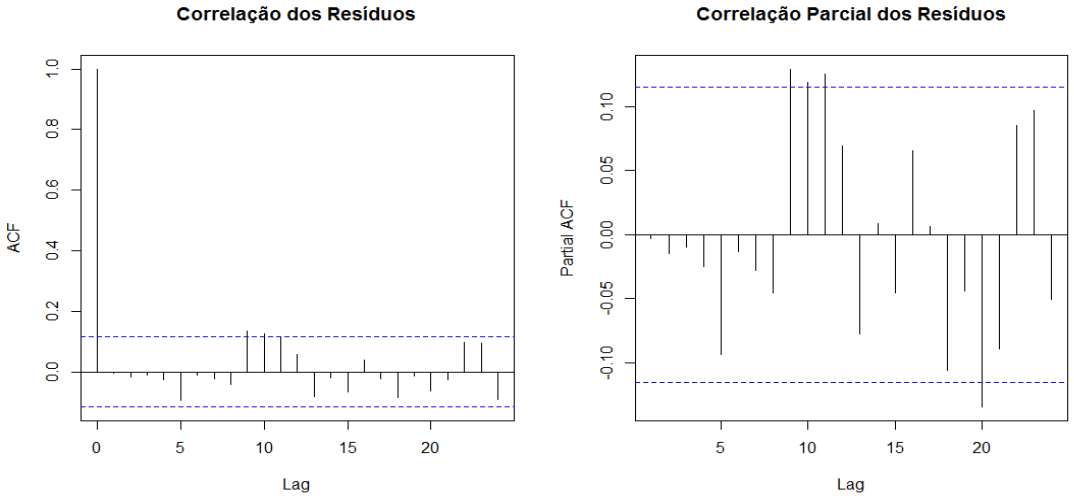


Figura 3: Autocorrelação e Autocorrelação Parcial dos Resíduos.

Por fim temos o modelo k-*factor* GARMA com k=3 e p=3, u={0.80, 0.84, 0.99}, λ={0.19, 0.22, 0.23}, φ={0.01, -0.10, -0.20}. estimado pelo método de FTmod. O qual acerta as 25 observações futuras da série, noutras palavras, os valores reais da série estão nos intervalos de confiança, isso foi possível pois buscamos os valores de tempestades na superfície do sol, para anos subsequêntes ao disponíveis no R. A Figura 4 a seguir apresenta a previsão da série real para 25 anos a frente.

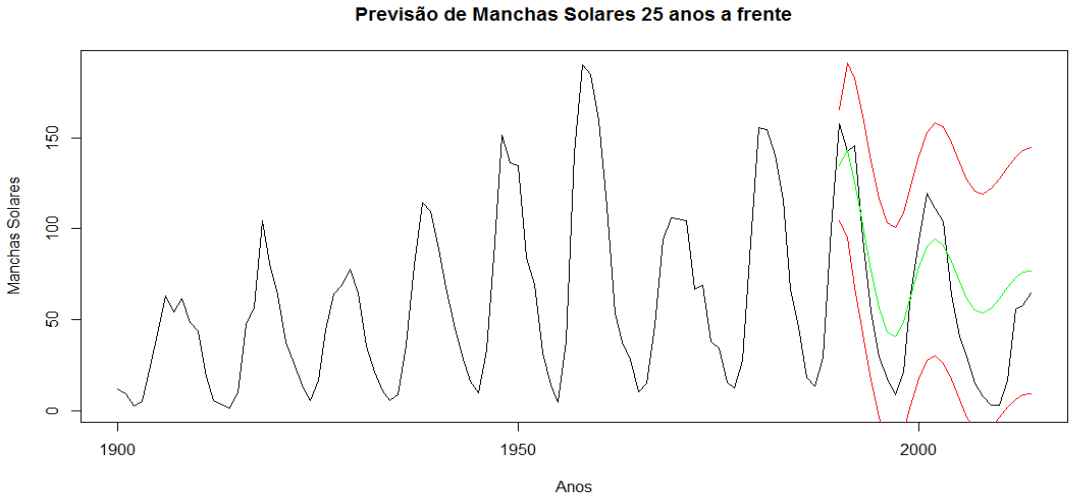


Figura 4: Previsão da Série Real.

4 Conclusões

Claramente temos dois padrões entre os estimadores, um primeiro grupo formado pelos FT, FTmod e MCMC, e um segundo grupo formado pelos FTAO e FTmodAO. Os vícios de FTAO e FTmodAO foram bem menores do que os de FT, FTmod e MCMC com exceção de u2, para o qual acontece o inverso. Já no tocante aos EQMs, temos novamente resultados análogos, pois os FTAO e FTmodAO foram melhores do que os estimadores não preparados para os outliers aditivos com exceção do EQM para de λ^2, no qual o resultado novamente foi o contrário.

Uma primeira aproximação com dados reais foi bastante satisfatória, uma vez que estimamos boas previsões para 25 anos de comportamento do sol. No entanto os ICs podem ter a sua amplitude reduzida uma vez que o desvio padrão dos resíduos diminua. Isso será possível se identificarmos um modelo com ainda melhor aderência do que o atual.

5 Referências

[1] BISOGNIN; C., LOPES, S.R.C., "Properties of Seasonal Long Memory Processes", Mathematical and Computer Modelling, 49, 1837-1851, 2009.

[2] Taqqu, M.S.; V. Teverovsky e W. Willinger, "Estimators for Long Range Dependence: an Empirical Study", Fractals, Vol. 3(4), pp. 785-798, 1995.

[3] TSAY, Ruey S., "Outliers, Level Shifts, and Variance Changes in Time Series", Journal of Forecasting, Vol. 7, I-20, 1988.

[4] Whittle, P., "Hypothesis Testing in Time Series Analysis", Journal of the American Statistical Association, Vol. 49, No. 265, pp. 197-200, 1954.

PROPOSTA DE ENSINO DE ESTATÍSTICA EM UMATURMA DE NONO ANO DO ENSINO FUNDAMENTAL COM USO DO PROGRAMA R-COMMANDER

Luís Henrique Pio de Almeida[1]

Aline Castello Branco Mancuso[2]

Luciana Neves Nunes[3]

Resumo: Muitas pesquisas estão sendo realizadas na área de ensino de Matemática que apontam a necessidade e os ganhos educacionais no estudo e no ensino de conceitos estatísticos. Neste contexto, este trabalho traz a análise e a proposta de uma atividade para estudantes do Ensino Fundamental. O principal objetivo deste trabalho foi planejar e aplicar uma sequência didática, envolvendo o uso do programa R-Commander, que estimulasse os alunos no interesse e compreensão dos conceitos estatísticos. Em termos metodológicos, a pesquisa empregada classifica-se como estudo de caso, realizada em uma escola estadual de Porto Alegre (RS) com uma turma do Nono ano do Ensino Fundamental. Esta proposta se enquadra no cenário de investigação descrito por Skovsmose (2003) e nos moldes da modelagem Matemática. A análise dos dados coletados foi baseada nos princípios da Educação Estatística. A partir dos resultados obtidos foi possível se observar que a modelagem matemática aliada ao uso do R-commander foi uma combinação favorável para uma boa abordagem do ensino de Estatística.

PALAVRA - CHAVE: Ambientes de Aprendizagem, Educação Estatística, Modelagem Matemática.

1. INTRODUÇÃO

A Estatística, como campo de estudo no currículo escolar, é recente. Apesar de presente nas orientações dos Parâmetros Curriculares Nacionais (PCN) (1998), ainda se apresenta pouco trabalhada nas escolas. A partir da experiência, como professor da rede estadual, em Porto Alegre (RS), percebe-se que o ensino de Estatística não faz parte dos planos de estudos.

A abordagem da Estatística nesta pesquisa é fundamentada nos PCNs , que aponta para a importância de seu estudo já no Ensino Fundamental, devido a crescente demanda social. Os parâmetros apontam que para um pleno exercício da cidadania os indivíduos devem saber calcular, medir, raciocinar, argumentar e tratar informações estatisticamente.

Mesmo a Estatística sendo parte integrante nos objetivos do ensino regular, esta é muito pouco trabalhada. Visto que a Estatística possui importante aplicabilidade nos meios sociais, seu ensino possibilita o desenvolvimento do pensamento crítico/social dos estudantes. Neste ideal, foi proposta e analisada uma sequência didática para uma turma de alunos do Nono Ano do Ensino Fundamental. Além disto, esta prática de ensino possibilita o trabalho com temas

1. UFRGS, apoio FAPERGS (lh.pioalmeida@gmail.com)
2. Hospital de Clínicas de Porto Alegre (aline.mancuso@gmail.com)
3. UFRGS (lununes@mat.ufrgs.br)

transversais. Desse modo, o ensino de Estatística se justificaria por promover o debate de assuntos no contexto social dos alunos.

A pesquisa teve por objetivo responder a questões consideradas norteadoras do trabalho: Como se desenvolve o ensino de Estatística no nível fundamental? Como podemos abordar este estudo a partir da modelagem matemática? A modelagem, aliada ao uso de um programa estatístico, é um caminho para uma boa abordagem da Estatística? A prática da pesquisa e coleta de dados por parte do corpo discente é capaz de ser motivadora para o entendimento da Estatística e formadora de um indivíduo mais critico socialmente?

Este trabalho está dividido em três tópicos. No primeiro é apresentada a metodologia utilizada na pesquisa e as referências para criação da sequência didática, embasadas nos Ambientes de Aprendizagem, na Educação Estatística e na Modelagem Matemática. No segundo, é apresentada a metodologia de Ação Docente. No terceiro tópico são apresentados os resultados e conclusões.

2. METODOLOGIA DE PESQUISA E EMBASAMENTO TEÓRICO

A metodologia adotada nesta pesquisa foi o estudo de caso com 24 alunos do Nono ano do Ensino Fundamental de uma escola de Porto Alegre da rede estadual do RS. O estudo de caso é um método qualitativo (Ludke e André, 1986) que busca descrever os ambientes e os indivíduos envolvidos na pesquisa. Segundo Fiorentini e Lorenzato (2007), o caso é o estudo que busca de forma profunda e o mais completa possível retratar a realidade. Por meio de uma sequência didática aplicada aos alunos, foram introduzidos os conceitos estatísticos propostos.

A atividade desenvolvida se enquadra nos princípios dos cenários de investigação. De acordo com Skovsmose (2000), a educação matemática tradicional se enquadra no que é denominado de paradigma do exercício. Neste enquadramento, o professor apresenta as ideias e técnicas e, em seguida, os alunos trabalham com exercícios selecionados. O autor ainda aponta que, geralmente, os livros didáticos, principal fonte de consulta dos professores, estão baseados no paradigma do exercício, visto que, os livros representam as condições da prática de sala de aula. Skovsmose (2000) coloca que no paradigma do exercício a premissa é que exista apenas uma resposta correta, o que contraria um cenário de investigação. Em um ambiente de investigação, não apenas os resultados, mas os meandros, os caminhos e as discussões devem ser considerados e avaliados. O autor define um cenário de investigação como sendo um ambiente que serve como suporte a um trabalho de investigação e que este cenário deve ser convidativo aos alunos a formularem questões e explicações.

Para desenvolver um ambiente de aprendizagem, descrito por Skovsmose como um cenário para investigação, foi utilizada a definição de Modelagem Matemática. Barbosa (2001) define Modelagem Matemática como um ambiente capaz de levar os alunos a indagarem e investigarem as diversas áreas da realidade por meio da Matemática. O autor também aponta distinções entre a Matemática pura e a Matemática aplicada. A Matemática aplicada teria por

objetivo explicar e dedicar-se a problemas oriundos de outras áreas do conhecimento, enquanto a matemática pura teria fim em si mesma. Assim sendo, a modelagem é capaz de oportunizar aos alunos, por meio da Matemática, questionamentos de situações sem procedimentos fixados, pois, neste ideal, os conceitos e ideias dependem dos encaminhamentos gerados pelos alunos (Barbosa, 2001).

Pensando no ensino da Estatística e em desenvolver, a partir de modelagem, os conceitos estatísticos, é abordada a Educação Estatística. Campos et al. (2011) colocam que, nos aspectos teóricos, são relevantes três competências no processo pedagógico de conteúdos estatísticos: a literacia, o pensamento e o raciocínio estatístico. Essas competências são relacionadas entre si. Os autores argumentam que as competências são voltadas a criação de uma cidadania crítica, visto que são baseadas na interpretação e compreensão crítica de informações, o que apontam para os princípios da Educação Crítica. É na perspectiva de uma sala de aula voltada para a Educação Crítica que podemos pensar o ensino e aprendizagem da Estatística.

3. METODOLOGIA DE AÇÃO DOCENTE

No primeiro encontro os alunos foram levados à sala de áudio visual, na qual foi apresentado o vídeo "O Prazer da Estatística". Este vídeo mostra aplicações e a importância da Estatística no contexto social, além de trazer um panorama do que é a Estatística. Ainda no primeiro encontro, foram entregues reportagens de jornais nas quais os alunos deveriam de encontrar indícios da aplicação da estatística. Neste momento, os alunos foram separados em grupos para que houvesse uma maior interação.

No segundo encontro, os alunos foram novamente organizados em grupos e foi proposta a realização de uma pesquisa com um questionário contendo cinco perguntas já elaboradas. Neste momento, os alunos vivenciaram uma experiência de organização de uma pesquisa. Tendo em mãos os dados coletados pelos alunos, ocorreu uma discussão sobre os diferentes tipos de variáveis.

Na terceira etapa, os alunos foram questionados sobre a melhor forma de organizar e resumir as informações obtidas na pesquisa realizada. Neste momento, foi abordada a frequência absoluta, relativa e medidas de tendência central (média, mediana e moda) a partir de exemplos levados pelo professor. De modo que os alunos percebessem que estas são medidas que resumem e caracterizam o comportamento das variáveis estudadas. Para a criação de um banco de dados foi apresentado aos grupos dados retirados do Campeonato Brasileiro de Futebol, no qual constavam informações tais como número de gols sofridos, número de vitórias, números de derrotas, etc. Em seguida, foi solicitado aos alunos que organizassem e resumissem as informações da pesquisa realizada no segundo encontro. Logo após a construção do banco de dados, os alunos resumiram as informações em tabelas, gráficos, média, moda e mediana.

Na quarta etapa, foi apresentado o software R-Commander. Esta apresentação ocorreu no laboratório de informática e os alunos estavam divididos em duplas. Foi proposto aos alunos que, com a utilização da ferramenta didática R-Commander, verificassem os resultados encontrados na primeira pesquisa realizada em sala de aula.

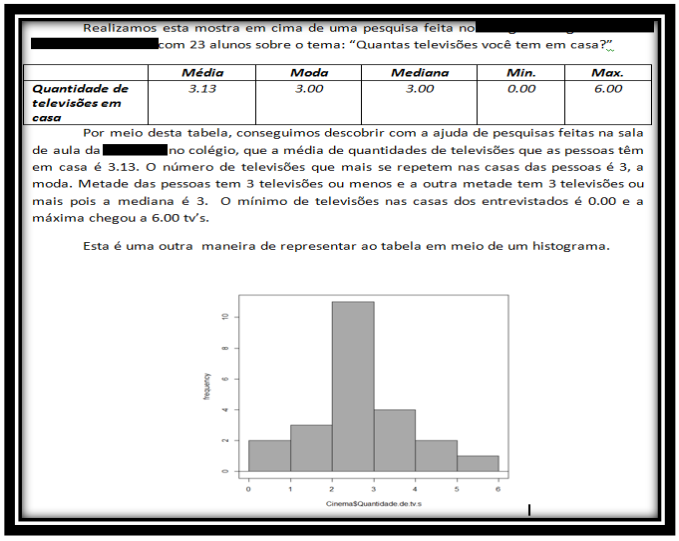
No quinto momento, foi proposta aos alunos a realização de uma segunda pesquisa, no âmbito escolar. A ideia inicial era entrevistar todos os alunos da escola. A intenção era proporcionar aos alunos a experiência de uma pesquisa, desde a escolha dos temas, organização, execução e análise dos dados obtidos. Os alunos escolheram os temas com base em discussões e elaboraram os questionários a partir do conhecimento adquirido. Os alunos se dividiram em cinco grupos e propuseram que as cinco pesquisas fossem realizadas em um único questionário. Logo após a coleta de dados, os grupos se reuniram para digitar e elaborar o banco de dados no software R-Commander. Criado o banco de dados e feita a análise estatística, os alunos avaliaram e interpretaram os resultados encontrados.

No sexto momento, visto que um dos objetivos da Estatística é a comunicação dos resultados encontrados, foi proposto aos alunos que elaborassem uma apresentação, em Power Point, que seria exposta na escola, a fim de comunicar aos demais colegas os resultados encontrados. Para finalizar a atividade, os alunos apresentaram os resultados da pesquisa. Esta apresentação ocorreria no pátio da escola (em área coberta), com a utilização de aparelhagem áudio visual.

4. RESULTADOS E CONCLUSÕES

Podemos perceber que os grupos de alunos, em geral, conseguiram utilizar com destreza os recursos do software, o que proporcionou uma maior concentração nas conclusões dos resultados, comparados com os trabalhos realizados sem o uso do software. Assim como em algumas partes da Matemática, sem o recurso computacional os alunos se preocuparam mais com as construções em detrimento do que estava sendo representado. Como mostra a Figura 1, o dinamismo do recurso possibilitou um maior entendimento dos conceitos e finalidades da Estatística.

Analisando as atividades e os resultados obtidos foi possível se observar que a modelagem matemática aliada ao uso de um programa estatístico, formando um cenário de investigação, é uma combinação favorável para uma boa abordagem de temas da Estatística. A proposta desenvolvida em sala de aula atendeu as expectativas dos alunos e foi motivadora para o entendimento da Estatística, visto que despertou o interesse destes alunos no assunto. Os alunos mostraram-se críticos sobre temas sociais, visto que traziam, para sala de aula, questionamentos e entendimentos sobre repoetagens visualizadas em casa.


Realizamos esta mostra em cima de uma pesquisa feita no [redacted] com 23 alunos sobre o tema: "Quantas televisões você tem em casa?".

	Média	Moda	Mediana	Min.	Max.
Quantidade de televisões em casa	3.13	3.00	3.00	0.00	6.00

Por meio desta tabela, conseguimos descobrir com a ajuda de pesquisas feitas na sala de aula da [redacted] no colégio, que a média de quantidades de televisões que as pessoas têm em casa é 3.13. O número de televisões que mais se repetem nas casas das pessoas é 3, a moda. Metade das pessoas tem 3 televisões ou menos e a outra metade tem 3 televisões ou mais pois a mediana é 3. O mínimo de televisões nas casas dos entrevistados é 0.00 e a máxima chegou a 6.00 tv's.

Esta é uma outra maneira de representar ao tabela em meio de um histograma.

Figura 1 – Resultados relatados pelos alunos.

Também se pôde perceber que o ensino de Estatística, em conjunto a proposta de ensino, é capaz de ser motivadora também no estudo da própria Matemática, visto as aplicações da mesma nos conceitos estatísticos. A proposta foi capaz de colocar os alunos em movimento. O movimento gerado, em busca do conhecimento, tirou o aluno da situação de repouso sobre as aulas de Matemática e os colocou em ação na construção de seu conhecimento.

Referências Bibliográficas:

BARBOSA, J. C. **Modelagem na Educação Matemática: contribuições para o debate teórico.** Reunião anual da ANPED, v. 24, p. 1-15, 2001.

BRASIL. Ministério da Educação e Desporto. Secretaria de ensino Fundamental. **Parâmetros Curriculares Nacionais** – Terceiro e quarto ciclos do Ensino Fundamental. Matemática. Brasília. DF: MEC, DEF, 1998.

FIORENTINI, Dario; LORENZATO, Sergio; **Investigação em Educação Matemática:** Percursos teóricos e Metodológicos. São Paulo: Autores Associados, ago. 2007.

LUDKE, Menga; ANDRÉ, Marli E. D. A. **Pesquisa em educação: Abordagens Qualitativas**. São Paulo: EPU, 1986.

SKOVSMOSE, Ole. **Cenários para investigação**. Bolema – Boletim de Educação Matemática. Rio Claro, nº 14, p. 66 a 91, 2000.

SILVA, E. L.; MENEZES, E. M. **Metodologia da Pesquisa e Elaboração de Dissertação**. Universidade Federal de Santa Catarina. Florianópolis, 2005.

Vídeo "**O Prazer da Estatística**" (Disponível em http://www.youtube.com/watch?v=xLr68J2yDJ8).

Análise multivariada de parâmetros físicos, químicos e biológicos da água da Baía de Guanabara-RJ

Vitor de Borba[1]

Prof. Dr. Fernando H. Pulgati[2]

Cristiano Sulzbach[3]

Prof. Dr. Rodolfo Paranhos[4]

Resumo: A pesquisa foi realizada na Baía de Guanabara-RJ entre Junho de 2005 e Julho de 2007. Foi observada uma amostra para a caracterização do ambiente e o planejamento do programa de monitoramento contínuo da Baía de Guanabara. As perguntas e hipóteses da pesquisa foram debatidas em reuniões com membros da Petrobras e com professores das universidades envolvidas no estudo de caracterização. O presente trabalho apresenta o estudo de variáveis bióticas e abióticas observadas no grupo da Hidrobiologia. A pesquisa foi conduzida para buscar estabelecer relações entre estas componentes. Foram consideras 10 estações amostrais, designadas como BGs, repetidas em 48 Campanhas.

Palavras-chave: *Baía de Guanabara, Bióticas e abióticas, IV SEMANÍSTICA, STATISTICS2014.*

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

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

[3] UFRGS - Universidade Federal do Rio Grande do Sul. Email: cristiano.sulzbach@hotmail.com

[4] UFRGS - Universidade Federal do Rio de Janeiro. Email: ufrj.rodolfo@gmail.com

1 Introdução

A pesquisa foi realizada a partir dos dados retirados do estudo de caracterização da Baía de Guanabara entre Junho de 2005 e Julho de 2007. Ao observar as informações iniciou-se um estudo que possibilitou a investigação dos diferentes cenários considerando o tamanho da amostra, modelos estatísticos e os testes de hipóteses, que foram propostos para o andamento do projeto que consiste no monitoramento contínuo da Baía de Guanabara.

2 Metodologia

As perguntas e hipóteses sugeridas para a pesquisa foram definidas em reunião técnicas da Petrobras e com professores das universidades envolvidas no estudo de caracterização. As análises foram feitas buscando estabelecer relações entre as variáveis bióticas e abióticas. E para a coleta dos dados foram consideras 10 estações amostrais, designadas como BGs, repetidas em 48 Campanhas.

3 Resultados

A Figura 1 mostra de forma nítida a relação inversa entre o indicador biológico e o Nitrito na água. O gráfico (a) registra a evolução do efeito ou coeficiente de regressão para o nitrito ao longo das 48 campanhas. É possível concluir, com 95% de confiança, que ele foi significativo durante todo o período observado. Já, os gráficos b e c da Figura 1 descrevem nitidamente o comportamento sazonal inverso das variáveis independente e dependente, Nitrito e Abundância Bacteriana respectivamente.

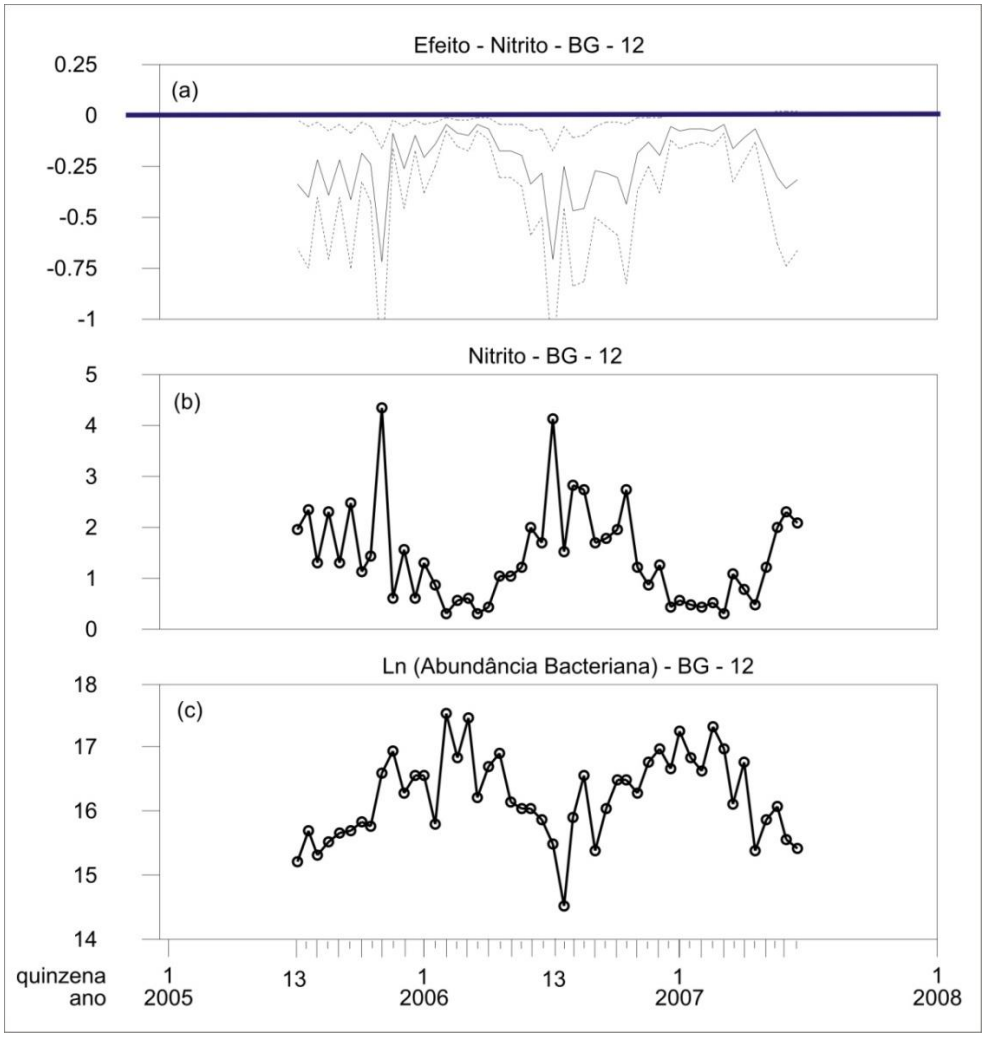


Figura 1. Série temporal da Abundância bacteriana e Nitrito na BG-12. Efeito de Nitrito sobre a Abundância bacteriana (a), Série temporal do Nitrito (b) e Série Temporal da Abundância bacteriana (c).

A análise da Figura 2, gráficos (b) e (c), mostram as relações positivas entre o indicador biológico e temperatura na água. Já o gráfico (a) descreve a evolução do efeito ou coeficiente de regressão da temperatura ao longo das 48 campanhas. Então novamente, é possível constatar que ele foi significativo durante todo o período observado.

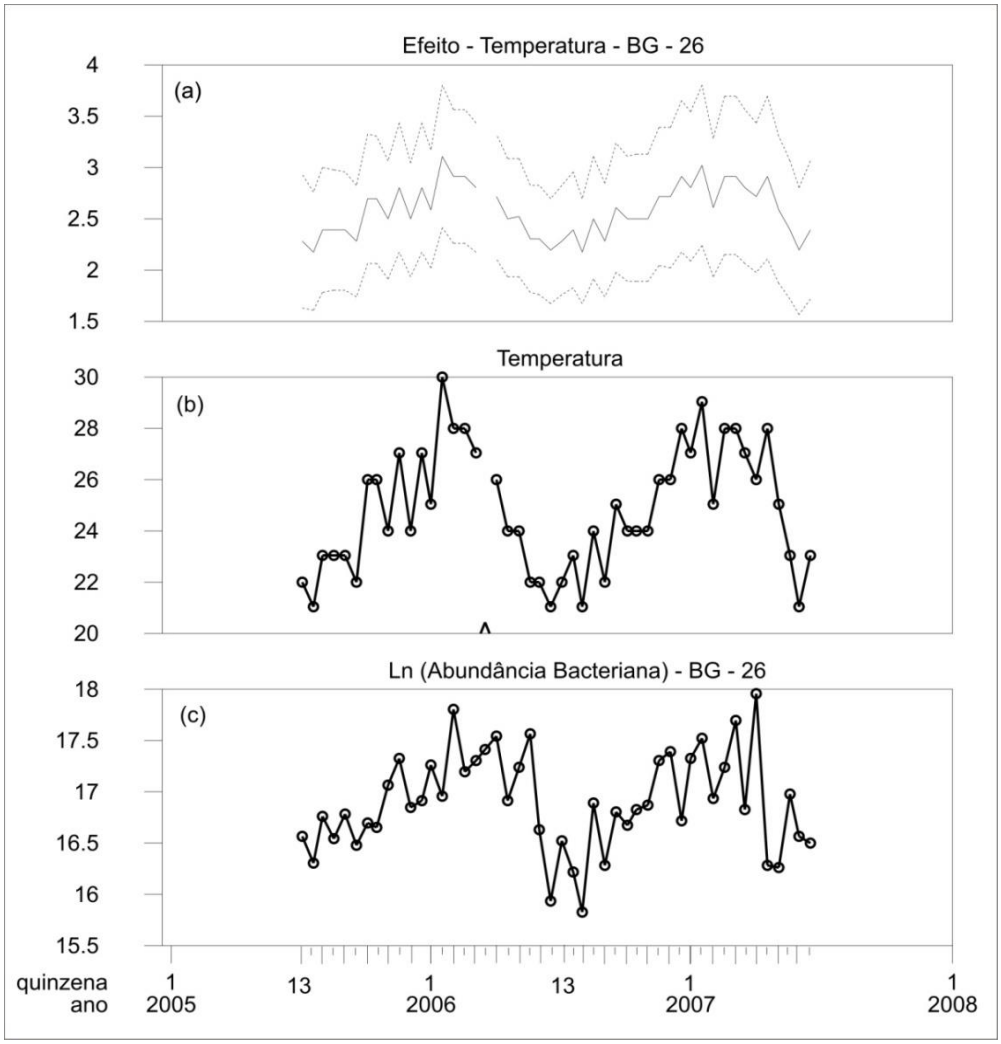


Figura 2 Série temporal da Abundância bacteriana e Temperatura da água na BG-26. Efeito de temperatura sobre a Abundância bacteriana (a), Série temporal da temperatura (b) e Série Temporal da Abundância bacteriana (c).

4 Conclusão

A pesquisa revela a relação inversa entre os componentes nas águas da Baía de Guanabara considerando o local e o tempo em que foram coletados. Foi possível também concluir sobre a evolução do efeito ou coeficiente de regressão da temperatura ao longo das 48 campanhas, no qual se pode constatar que ele foi significativo durante o período observado. Ocorreram, em geral, relações entre as componentes bióticas e abióticas ao longo do estudo sobre a caracterização da Baía de Guanabara.

Referências

[1] BROWN, M. B.; FORSYTHE, A. B. The small sample behaviour of some statistics which test the equality of severals means. *Technometrics*, vol. 16, p. 385-389, 1974. HANSEN, L.D.;

[2] CHOU, Y. M.; MASON, R. L.; YOUNG, J. C. Power comparisons for a Hotelling's 2 *T* statistic. *Communications in Statistics – Simulation and Computation*, vol, 24, nº 4, p. 1031-1050, 1999.

[3] COHEN, J. Statistical Power Analysis for the Behavioral Sciences.Lawrence Eribaum, Inc., Publishers, 1977.

[4] ITO, P. K. Robustness of ANOVA and MANOVA test procedures. *Handbook of Statistics*, vol. 1, p. 199-236, 1980.

[5] JAMES, G. S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of population variances are unknown. *Biometrika*, vol. 41, p. 19-43, 1954.

Análise multivariada de parâmetros físicos, químicos e biológicos da água da Baía de Guanabara-RJ

Cristiano Sulzbach[1]
Prof. Dr. Fernando H. Pulgati[2]
Vitor de Borba [3]
Prof. Dr. Rodolfo Paranhos [4]

Resumo: Através de dados provenientes do estudo de caracterização da Baía de Guanabara realizado entre junho de 2005 e julho de 2007, análises foram conduzidas buscando estabelecer relações entre as componentes bióticas e abióticas. A dimensão original do banco de variáveis abióticas foi reduzida através da Análise de Componentes Principais e, posteriormente, foram regredidos os indicadores biológicos sobre os vetores multivariados. As variáveis abióticas foram reduzidas para três (3) componentes. Estas três (3) explicam 76,23 da variabilidade total dos dados abióticos. Assim, permitiu-se fazer um estudo de caracterização, investigando diferentes cenários através de modelos estatísticos.

Palavras-chave: *Baía de Guanabara, Componentes Principais, variáveis.*

1 Introdução

[1] UFRGS - Universidade Federal do Rio Grande do Sul. Email: cristiano.sulzbach@*hotmail.com*
[2] UFRGS - Universidade Federal do Rio Grande do Sul. Email: pulgati@ufrgs.br
[3] UFRGS - Universidade Federal do Rio Grande do Sul. Email: vitorborbavtr@*gmail.com*
[4] UFRJ - Universidade Federal do Rio de Janeiro. Email: rodolforpparanhos@*gmail.com*

Através de técnicas conhecidas como *Power Analysis*, foram realizados estudos em bases de dados provenientes da caracterização da Baía de Guanabara, realizado entre junho de 2005 e julho de 2007. Inicialmente levantaram-se as informações do estudo de caracterização. A partir disso, a metodologia permitiu investigar diferentes cenários considerando o tamanho da amostra e modelos estatísticos para o projeto de monitoramento da Baía de Guanabara. O resultado do estudo tem por objetivo detectar diferenças importantes, com alta probabilidade, ressaltando que as análises foram conduzidas buscando estabelecer relações entre as componentes bióticas e abióticas.

2 Metodologia

As hipóteses de pesquisa foram discutidas em reuniões técnicas com membros de professores das universidades envolvidas no estudo de caracterização. A partir dessas reuniões definiu-se que o banco de dados do presente estudo constituir-se-ia por variáveis bióticas e abióticas observadas pelo grupo da Hidrobiologia (IB/UFRJ). Desse modo, considerou-se 10 estações amostrais, designadas como BGs, repetidas em 48 Campanhas, tanto na superfície, quanto no fundo da Baía de Guanabara.

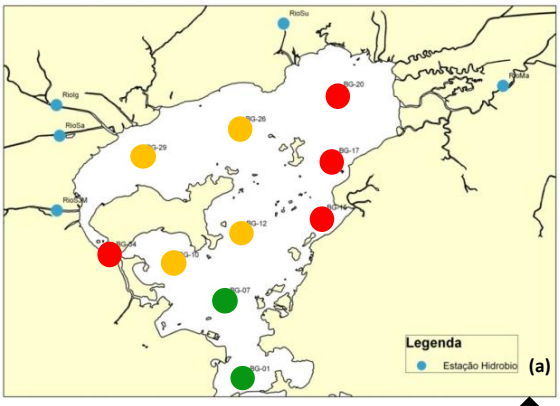
Para realizar a análise de dados, a dimensão original do banco de variáveis abióticas do Grupo da Hidrobiologia foi reduzida através da técnica de componentes principais. Regrediu-se o indicador biológico "Abundância Bacteriana" sobre a estrutura abiótica, representada pelas componentes principais. Por este viés, a estratégia consistiu em reduzir a dimensão original da estrutura de dados das variáveis físicas e químicas, considerando as particularidades de cada ambiente, superfície e fundo.

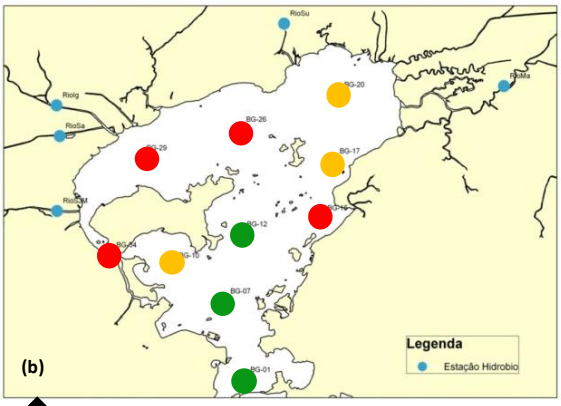
3 Resultados

A Figura 1 descreve as componentes e suas cargas fatoriais, representantes das variáveis abióticas do banco de dados da hidrobiologia. A cor verde indica a correlação significativa entre a variável e a componente. Por exemplo, na interpretação da Componente 1, têm-se as variáveis Silicato e Fósforo Total correlacionadas positivamente com esta componente, enquanto Salinidade, Nitrato e OD (Oxigênio Dissolvido) correlacionadas inversamente com esta mesma componente.

No Gráfico (a) da Figura 1, as estações amostrais marcadas pela cor vermelha registraram valores acima da média nas concentrações de Silicato e Fósforo Total. As demais de cor alaranjada, concentrações em torno da média, e as restantes, de cor verde, valores abaixo da média. Inversamente, as estações amostrais indicadas em vermelho registraram concentrações abaixo da média nas variáveis Salinidade, Nitrato e OD. Com base nas cargas fatoriais da Componente 2, interpretação análoga pode ser estendida para o Gráfico (b).

Os resultados descritos na Figura 3.1 também foram cruzados com as decorrências obtidas dentro de cada campanha, visto que a regressão da Abundância Bacteriana implementou-se em cada um dos tempos, enquanto a estrutura abiótica multivariada representada pelas componentes foi estabelecida independente de tempo. O objetivo deste procedimento foi estabelecer um modelo que representasse a estrutura abiótica de todo o estudo (48 campanhas) e pudesse ser avaliado isoladamente dentro de cada campanha.





Fundo	Componente		
	1	2	3
Rq(Silicato+1)	.877	.125	
Salinidade	-.866	.112	-.184
Log(Fosforo Total+1)	.714	.546	
Log(Nitrato+1)	-.608	-.327	-.368
%satOD	-.498	-.393	.376
Log(Amônia+1)	.155	.877	
Log(Nitrito+1)		.858	.119
Log(N_Total+1)	.457	.698	
logmpsmu	.166		.878

Figura 3.1. Distribuição espacial dos escores fatoriais. Gráfico (a) Componente 1 e Gráfico (b) Componente 2. Estações representadas pela cor vermelha sinalizam valores altos, laranja valores médios e verdes valores baixos.

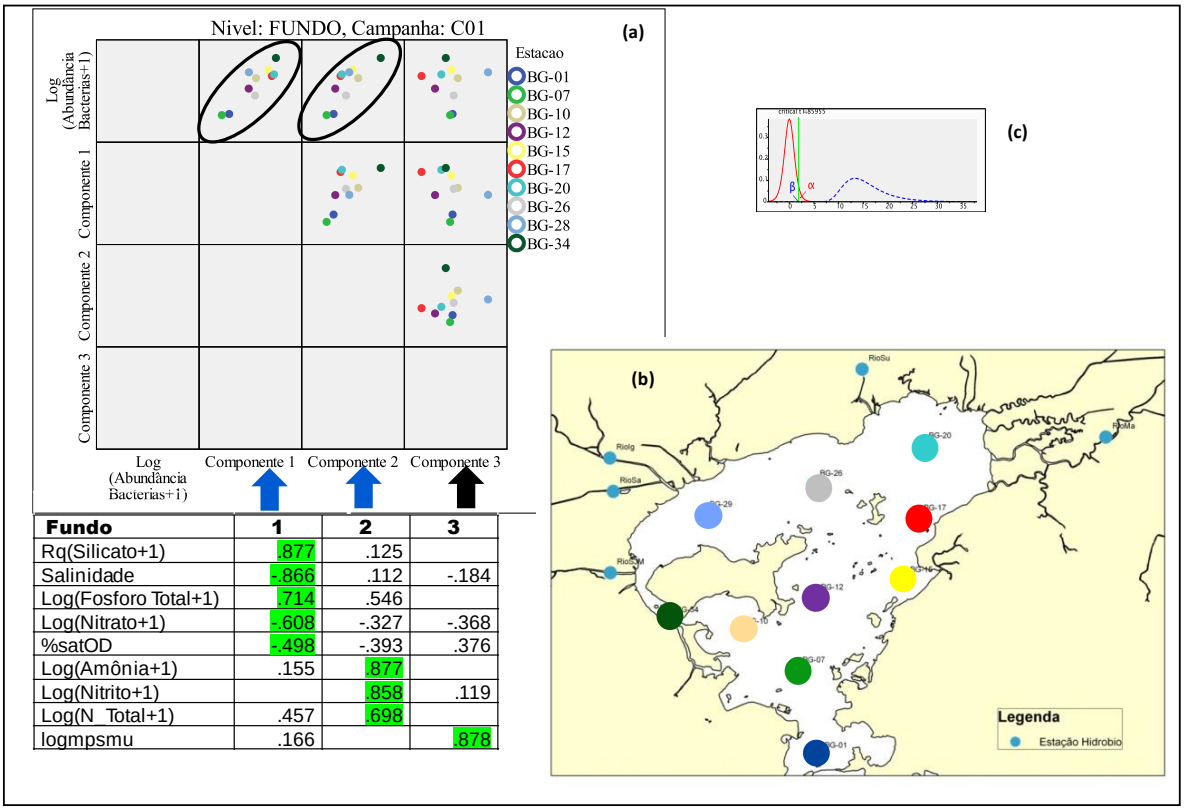
A Figura 2. descreve o resultado da aplicação do modelo dentro da Campanha 1, Fundo. Em (a) é mostrada a dispersão entre o indicador biológico e as componentes abióticas. Em (b) é apresentada a localização das estações amostrais com o respectivo *lable* utilizado em (a).

A análise integrada das figuras 1 e 2 (Campanha 1-Fundo) permitem produzir as seguintes afirmações:

i) existe forte relação entre a Abundância Bacteriana e as componentes abióticas;

ii) existe padrão espacial muito bem definido nesta relação;

iii) a estrutura multivariada construída com as 48 campanhas funcionou adequadamente dentro da Campanha 1, Fundo.



Fundo	1	2	3
Rq(Silicato+1)	.877	.125	
Salinidade	-.866	.112	-.184
Log(Fosforo Total+1)	.714	.546	
Log(Nitrato+1)	-.608	-.327	-.368
%satOD	-.498	-.393	.376
Log(Amônia+1)	.155	.877	
Log(Nitrito+1)		.858	.119
Log(N_Total+1)	.457	.698	
logmpsmu	.166		.878

Figura 2. (a) Matriz de dispersão entre as variáveis Abundância Bacteriana e componentes abióticas 1,2 e 3 na BG 1 - Fundo. (b) Distribuição espacial das campanhas com respectivo *lable*. Gráfico das distribuições especificadas sob as hipóteses, nula e alternativa.

4 Conclusão

O presente estudo possibilitou identificar padrões na relação entre os fatores bióticos e abióticos da Baia de Guanabara. Destarte, as análises desenvolveram-se dentro de cada campanha considerando, desta forma, a heterogeneidade espacial presente na Baía de Guanabara. As correlações esperadas entre as componentes bióticas e abióticas não se confirmaram em todas as campanhas, embora tenha sido possível observar forte correlação e, por conseqüência, relação de causa e efeito em alguns casos.

Referências

[1]BROWN, M. B.; FORSYTHE, A. B. The small sample behaviour of some statistics whichtest the equality of severals means. *Technometrics*, vol. 16, p. 385-389, 1974. HANSEN, L.D.;

[2]CHOU, Y. M.; MASON, R. L.; YOUNG, J. C. Power comparisons for a Hotelling's 2 *T* statistic. *Communications in Statistics – Simulation and Computation*, vol, 24, nº 4, p. 1031-1050, 1999.

[3]COHEN, J. Statistical Power Analysis for the Behavioral Sciences.Lawrence Eribaum, Inc., Publishers, 1977.

[4]ITO, P. K. Robustness of ANOVA and MANOVA test procedures. *Handbook of Statistics*, vol. 1, p.199-236, 1980.

[5]JAMES, G. S. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of population variances are unknown. *Biometrika*, vol. 41, p. 19-43, 1954.

Modelagem do preço pago ao produtor de erva-mate do estado do Paraná

Afonso Valau de Lima Junior [1]
Bianca Reichert [2]
Viviane de Senna[3]
Adriano Mendonça Souza[4]

Resumo: Na região sul do Brasil são produzidas anualmente cerca de 500.000 toneladas de erva-mate ao ano e movimenta a economia regional e nacional. O objetivo desta pesquisa é analisar a série do preço pago ao produtor de erva-mate por meio dos modelos ARIMA – ARCH. A série temporal utilizada é a do preço médio mensal em reais da arroba de erva-mate recebido pelos produtores no estado do Paraná. Dentre os modelos analisados o mais adequado para a volatilidade apresentada pela série é descrito como um modelo conjunto SARIMA $(2,1,1)\ (1,0,0)_8$ – ARCH (1).
Palavras-chave: *Modelos ARCH, erva-mate, volatilidade.*

1 Introdução

Os estados do sul do país, Rio Grande do Sul, Santa Catarina e Paraná são responsáveis por praticamente toda a produção nacional. No Rio Grande do Sul, considerado o maior produtor brasileiro de folha verde de erva-mate, são colhidos cerca de 50% da produção nacional com uma média de 247.583 toneladas anual no período entre 2001 e 2012, seguido pelo Paraná com 171.325 583 toneladas anual, representando cerca de 37% da produção, o estado de Santa Catarina produz em média 45.117 toneladas por ano, representando cerca de 10 % da produção.

O objetivo da pesquisa é analisar a série do preço pago ao produtor de erva-mate por meio dos modelos autoregressivos integrados e de médias móveis (ARIMA) e dos modelos autoregressivos condicionais a heteroscedasticidade (ARCH), com o intuito de tratar a autocorrelação presente nos dados e interpretar o comportamento da série em nível e a variabilidade da mesma.

2 Metodologia

Os dados utilizados para a execução deste estudo foram obtidos no Departamento de Economia Rural (DERAL), pertencente a Secretaria de Estado da Agricultura e do Abastecimento (SEAB), e referem-se ao preço médio mensal (R$/arroba) recebido pelos produtores de erva-mate verde no

[1] UFSM - Pós-graduando de Especialização em Estatística. Email: avljunior@yahoo.com.br
[2] UFSM - Graduanda em Engenharia de Produção. Email: bianca.reichert@hotmail.com
[3] UFSM - Mestranda de Engenharia da Produção. Email: vivianedsenna@hotmail.com
[4] UFSM - Orientador. Professor do departamento de Estatística. Email: amsouza@smail.ufsm.br

Estado do Paraná, entre janeiro de 2006 a dezembro de 2013, composto por 96 mensais, optou-se pelo preço do estado do PR, pois não havia disponível uma série completa no estado do RS.

Sabendo-se da condição de uma série temporal (MORETTIN, TOLOI, 1986), espera-se que a mesma seja estacionaria. Superada 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. Pretende-se com esta modelagem encontrar 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 seleção dos modelos será por meio dos critérios AIC (*Akaike Information Criteria*) e BIC (*Bayesian Information Criteria*) através das equações AIC= lnσe2 + (2(p + q))/n e BIC = lnσe2 + ((p + q)lnn)/n; p e q são os parâmetros conhecidos, n é o tamanho da amostra, ln é o logaritmo neperiano e σe2 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).

Na etapa de análise dos resíduos investiga-se também em relação a presença de volatilidade, Engle (1982) SÁFADI e ANDRADE FILHO (2007), se houver dependência nos resíduos quadráticos do modelo ARIMA, há a possibilidade de se encontra um modelo autoregressivo condicional a heteroscedasticidade (ARCH) (BOLLERSLEV, 1986 e GUJARATI, 2000).

O termo de erro e_t, condicionado à informação disponível no período *(t-1)* seria distribuído conforme a seguinte notação (CAMPOS, 2007):

$e_t \sim N[0,(\alpha_0 + \alpha_1 e_{t-1}^2)]$, em que α_0 e α_1 são parâmetros explicativos da variância do termo de erro e_t.

Um modelo ARCH(m), em que *m* denota a ordem do modelo, expressa a variância condicional do modelo para a média condicional como uma função das inovações quadráticas passadas (SILVA, SÁFADI e CASTRO JÚNIOR, 2005).

Para assegurar que a variância condicional seja positiva e fracamente estacionária, as seguintes restrições paramétricas são necessárias: $\alpha_0 > 0$, $\alpha_t \geq 0$ para todo $t = 2...m$ e $\sum \alpha_t < 1$. E, sob a condição de estacionariedade BUENO (2008), ENDERS (1995) e HAMILTON (1994).

3 Resultados e Discussões

A série original do preço médio mensal (R$/arroba) recebido pelos produtores de erva-mate verde no Estado do Paraná é apresentada na Figura 1, observa-se que está série possui uma tendência ascendente e um pico no ano de 2013 e também é constatado a necessidade da utilização da diferenciação para torná-la estacionária.

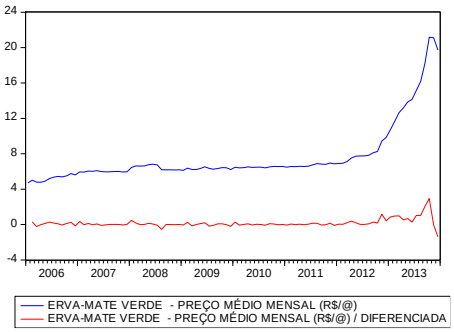


Figura 1 – Série original e Série diferenciada.

Ao efetuar o próximo procedimento metodológico, que consiste na identificação de modelos significativos que representem o comportamento da série, foi identificado um SARIMA (2,1,1) $(1,0,0)_8$ conforme mostra a Tabela 1 que apresenta o modelo SARIMA-ARCH para a variável preço recebido pelo produtor de erva mate no PR.

TABELA 1 – Estimação dos coeficientes, erro-padrão, estatística z e *p-value* do modelo SARIMA-ARCH preço recebido pelo produtor de erva mate no PR.

Método: ML – ARCH (Marquardt) - Distribuição Normal				
Equação para a média condicional				
	Coeficiente	Erro-Padrão	Estatística z	p-valor
AR(1)	0.380613	0.135534	2.808252	0.0050
AR(2)	0.328537	0.122409	2.683929	0.0073
SAR(8)	0.752266	0.261821	2.873213	0.0041
MA(1)	-0.464776	0.273471	-1.699541	0.0892
Equação para a variância condicional				
Constante	0.036399	0.003599	10.11398	0.0000
ARCH(1)	0.700097	0.236902	2.955219	0.0031

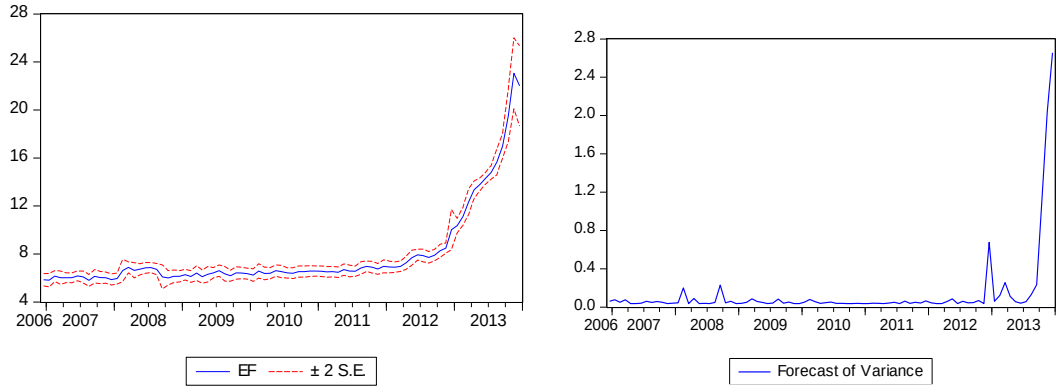
O modelo descrito é um modelo conjunto SARIMA (2,1,1) $(1,0,0)_8$ – ARCH (1) apresentando parâmetros estatisticamente significativos e resíduos com características de *ruído branco*. Constata-se que a soma dos parâmetros é menor que um, assegurando que a variância condicional seja positiva e estacionária. As estatísticas de ajuste para o modelo são de um R^2 = 0.162388 e os critérios AIC = 0.191494, BIC = 2.154094 e um valor da estatística de DW =1.228720. O valor do parâmetro ARCH é um valor elevado e representa a persistência desta volatilidade, o que mostra que esta variável custará a retornar ao seu patamar normal de variabilidade no curto prazo.

Os resíduos originados do modelo SARIMA são não-correlacionados, com distribuição aleatória em torno de zero e variância aproximadamente constante, caracterizando como sendo *ruído branco*, mas os resíduos quadráticos apresentam heteroscedasticidade.

Os resíduos originados do modelo SARIMA-ARCH são não-correlacionados, com distribuição aleatória em torno de zero e variância aproximadamente constante, caracterizando como sendoso ruído branco.

Os resíduos originados do modelo ARIMA – ARCH, os quais apresentam-se não-correlacionados, com distribuição Normal em torno de zero e variância aproximadamente constante, caracterizando como sendo *ruído branco*.

Na Figura 2 a apresenta-se a previsão do preço pago ao produtor de erva mate e na Figura 2 b, o comportamento da volatilidade.



(a) (b)

Figura 2 - FAC e FACP do modelo SARIMA – ARCH, (a) resíduos, (b) resíduos quadráticos.

Observa-se que o modelo ajustado consegue captar o movimento da séire representando as oscilações que ocorreram no período e a variância estimada conseguiu detectar o período de maior volatilidade iniciando em fins de 2102.

4 Conclusão

Ao realizar o estudo referente a série do preço médio mensal (R$/arroba) recebido pelos produtores de erva-mate verde no Estado do Paraná, entre janeiro de 2006 a dezembro de 2013. Dentre os modelos analisados o mais adequado para a volatilidade apresentada pela série é descrito como um modelo conjunto SARIMA $(2,1,1)\ (1,0,0)_8$ – ARCH (1) e que conseguiu captar os movimentos da série.

Referências

[1] BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, v. 31, p. 307-327, 1986.

[2] BOX, G.E.; JENKINS, G.M.; REINSEL, G.C. Time series analysis: Forecasting and control. 3 ed. New Jersey: Printice Hall, 1994.

[3] BUENO, R. L. S., Econometria de séries temporais. São Paulo: Cengage Learning, 2008.

[4] CAMPOS, K. C., Análise da volatilidade de preços de produtos agropecuários no Brasil. Revista de Economia de Agronegócio, v. 5, n. 3, p. 303-328, 2007.

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

[6] ENGLE, R. F., Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. Econometria, v. 50, n. 4, p. 987-1008, 1982.

[7] GUJARATI, D. N. Econometria básica. São Paulo: Makron Books, 2000.

[8] HAMILTON, J. Time series analysis. New Jersey: Princeton University Press, 1994.

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

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

[11] SÁFADI, T.; ANDRADE FILHO, M. G., Abordagem bayesiana de modelos de séries temporais. In: 12ª ESCOLA DE SÉRIES TEMPORAIS E ECONOMETRIA, 2007, Gramado. Minicurso... Gramado: Associação Brasileira de Estatística, 2007.

[12] SECRETARIA DA AGRICULTURA E DO ABASTECIMENTO DO ESTADO DO PARANÁ – SEAB. Disponível em: < http://www.agricultura.pr.gov.br/>. Acesso em: 01 out. 2014.

[13] SILVA, W. S.; SÁFADI, T.; CASTRO JÚNIOR, L. G., Uma análise empírica da volatilidade do retorno de commodities agrícolas utilizando modelos ARCH: os casos do café e da soja. Revista de Economia e Sociologia Rural, v. 43, n. 1, p. 120-134, 2005.

[14] 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.

Medidas de dissimilaridade para o método de classificação de séries temporais baseado em U-estatísticas

Augusto Marcolin [1] [3]
Marcio Valk[2] [3]

Resumo:O método de classificação e agrupamento de séries temporais baseado em U-estatísticas tem como característica a dependência de uma medida de dissimilaridade entre séries temporais. Essas medidas são utilizadas como núcleo das U-estatísticas e suas características influenciam diretamente o comportamento da estatística de teste. Na literatura, existem uma grande variedade dessas medidas e o objetivo deste trabalho é realizar um estudo comparativo, através de simulações de monte carlo, para identificar qual medida é mais adequada para o método, considerando-se diferentes tipos de processos estacionários na configuração dos grupos.
Palavras-chave: *Séries temporais, Classificação, U-estatística.*

1 Introdução

Atualmente existe uma demanda crescente pela utilização de métodos de classificação e agrupamento em séries temporais. Por esse motivo o assunto tem sido objeto de estudo em diversas áreas, tais como manutenção, medicina, biometria, química, astronomia,robótica, redes e indústria. Na medicina, por exemplo, a série temporal pode ser de valores da pressão sanguínea de um paciente a cada hora, ou a taxa de batimentos cardíacos por minuto. Um dos objetivos desses métodos de classificação é reconhecer a qual grupo pré-determinado o sinal pertence. Na aplicação anterior esses grupos podem corresponder, por exemplo, ao estado de saúde de um paciente (pressão normal, alta ou baixa no sangue, ritmo cardíaco regular ou irregular).

O método de agrupamento proposto por [1] consiste na suposição de homogeneidade dos grupos, ou seja, sob H_0 o processo gerador das séries temporais é o mesmo para todos os grupos. A suposição essencial é que dentro de cada grupo temos homogeneidade. A ideia então é utilizar as medidas de dissimilaridade [1] entre grupos e dentro dos grupos e mostrar que a estatística de teste composta pela diferença entre estas medidas é uma U-estatística e converge em distribuição para uma variável aleatória

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *augustomarcolin@gmail.com*
[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *marciovalk@gmail.com*
[3]

[1]O termo dissimilaridade é usado pois essas medidas não têm propriedade de uma distância

com distribuição normal. Essa convergência ocorre de duas formas: aumentando-se o tamanho das séries temporais e/ou aumentando-se o número de séries temporais (para mais detalhes ver [1]).

Na seção 2 apresentamos algumas das medidas de dissimilaridade mais comuns em séries temporais. Na seção 3 realizamos um estudo de simulação com o objetivo de verificar qual das medidas apresentadas na seção 2 têm melhor desempenho para uma determinada classe de processos estacionários, quando utilizadas como núcleo da estatística de teste no método proposto por [1].

2 Medidas de dissimilaridade entre séries temporais

Nesta seção apresentamos algumas das medidas de dissimilaridade mais comuns na literatura e que já estão implementadas no software R no pacote *"TSclust"*. No domínio da frequência, a medida conhecida como *logaritmo do periodograma normalizado* ($DNLP$) é definida como a distância euclidiana entre os coeficientes dos periodogramas das séries x e y,

$$DLNP(x,y) = \frac{1}{T}\sum_{\ell=1}^{\lfloor \frac{T}{2} \rfloor} \left(I^*{}_x(\omega_\ell) - I^*{}_y(\omega_\ell)\right)^2, \tag{1}$$

em que $I_x(\cdot)$ é a função periodograma da série x_t, $I^*{}_x(\omega) = \log\left[I_x(\omega)/\gamma_x(0)\right]$ é o logaritmo do periodograma normalizado e $\gamma_x(\cdot)$ é a função de autocovariância de x_t. Também no domínio da frequência, a medida de dissimilaridade entre duas séries temporais baseada na distância dos seus periodogramas integrados é definida por

$$INT.PER(x,y) = \int_{-\pi}^{\pi} |F_x(\lambda) - F_y(\lambda)| d\lambda \tag{2}$$

em que $F_x(\lambda_j) = C_x^{-1}\sum_{i=1}^{j} I_x(\lambda_i)$ e $F_y(\lambda_j) = C_y^{-1}\sum_{i=1}^{j} I_y(\lambda_i)$, com $C_x = \sum_{i=1} I_x(\lambda_i)$ e $C_y = \sum_{i=1} I_y(\lambda_i)$. Neste trabalho usamos a versão normalizada em que $C_x = C_y = 1$.

No domínio do tempo, uma das medidas é baseada na distância euclidiana ponderada entre os coeficientes de *autocorrelação*. O caso de ponderamento padrão será denotada por (DAC) e definida aqui por

$$DAC(x,y) = \sqrt{\sum_{h=1}^{L}\left(\widehat{\rho}_x(h) - \widehat{\rho}_y(h)\right)^2}, \tag{3}$$

em que $\widehat{\rho}_x(h) = \widehat{\gamma}_x(h)/\widehat{\gamma}_x(0)$, é a função de autocorrelação de x_t e L o número de autocorrelações, que deve ser determinado de alguma forma. Neste trabalho usamos $L = 50$. Igualmente relacionada a momentos amostrais, a medida de dissimilaridade baseada na correlação amostral (ou correlação de Pearson) é definida por

$$COR(x,y) = \sqrt{2(1-\rho)}, \tag{4}$$

em que ρ denota a correlação de Pearson entre as séries x e y. Uma medida adaptativa de dissimilaridade que cobre dissimilaridade no comportamento conjunto das séries e no comportamento dos coeficientes

de correlação temporal é definida por

$$CORT(x,y) = \Phi[crt(x,y)]\delta(x,y), \tag{5}$$

em que $crt(x,y) = \sum_t (x_{t+1} - x_t)(y_{t+1} - y_t)(\sum_t (x_{t+1} - x_t)^2 \sum_t (y_{t+1} - y_t)^2)^{-\frac{1}{2}}$ é o coeficiente de correlação temporal de ordem um e mede a proximidade entre o comportamento dinâmico de x e y. A função $\Phi[u] = 2/(1+e^{ku})$, com $k \geq 0$ é chamada de *"adaptative tuning function"* e $\delta(x,y)$ é a distância euclidiana entre x e y. Ainda no domínio do tempo, a medida que calcula a dissimilaridade baseada na distância euclidiana corrigida pela estimativa da complexidade da série é definida por

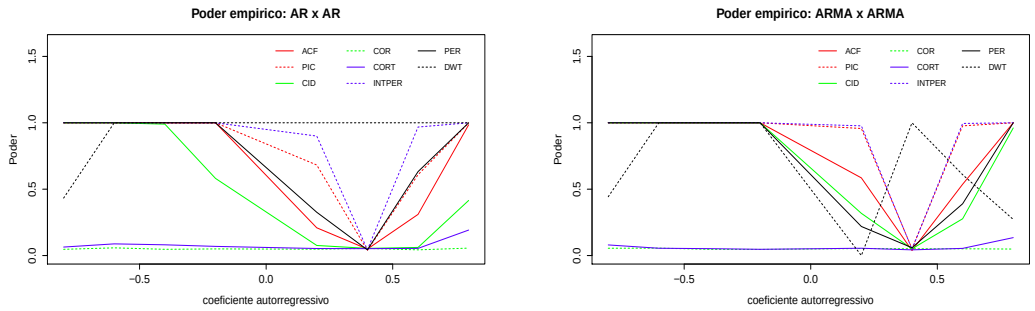
$$CID(x,y) = \delta(x,y) \times CF(x,y), \tag{6}$$

em que $CF(x,y)$ é o fator de correção de complexidade dado por $\max(CE(x), CE(y))/min(CE(x), CE(y))$, sendo $CE(x)$ a estimativa da complexidade de x definida por $CE(x) = \sqrt{\sum_{t=1}^{T}(x_{t+1} - x_t)^2}$. Outra maneira de medir dissimilaridade entre séries temporais é assumindo uma estrutura (modelo) para a série. Neste caso temos as chamadas *"model based distances"*. Assim, assumindo que a série pode ser representada através de um AR(∞), a medida baseada na distância euclidiana entre os coeficientes desta aproximação é chamada $AR.PIC(x,y)$. Ao substituir as séries temporais originais por seus coeficientes *"wavelets"* em uma escala apropriada e calcular a distância euclidiana entre esses coeficientes, temos a medida de dissimilaridade DWT.

3 Estudo de simulação

Realizamos um estudo de simulação para testar diferentes medidas de dissimilaridade para séries temporais, considerando primeiramente processos estacionários, em particular o processo autorregressivo (AR($\cdot$)). A primeira etapa consiste em gerar séries temporais artificiais a partir do processo AR(1). Para compor o primeiro grupo, foram geradas quatro séries a partir do processo AR(1), com coeficiente autorregressivo fixo $\phi_a = 0.4$, em que ϵ_t é aleatório com distribuição normal de média zero e variância um. Para compor o segundo grupo, foram geradas quatro séries a partir do mesmo processo X_t, mas com coeficiente autorregressivo tomando valores no conjunto $\phi_b \in \{-0.8, -0.6, -0.4, -0.2, 0.0, 0.2, 0.4, 0.6, 0.8\}$. Com essa configuração, quando $\phi_a = \phi_b = 0.4$ tem-se a situação de homogeneidade entre os grupos (H_0), e, ao nível de 5% de significância, espera-se que o teste rejeite H_0 aproximadamente 5% das vezes. Isso dá uma estimativa do tamanho do teste. Para as demais combinações de ϕ_a e ϕ_b, espera-se que o teste rejeite a hipótese de homogeneidade dos grupos (H_0) na maioria das vezes. A proporção de vezes em que rejeita-se H_0 é uma estimativa do poder do teste. O tamanho de cada série considerada foi $T = 512$ e foram realizadas 1000 replicações para cada ϕ_b. As medidas de distâncias consideradas no estudo foram ACF, PIC, CID, COR, $CORT$, $INTPER$, PER e DWT, as quais são descritas na seção anterior.

Os resultados deste estudo podem ser observados na figura 1a. A medida com melhor desempenho foi $INTPER$, seguida de PIC, PER, ACF e CID, nesta ordem. Para as medidas COR e $CORT$, o teste não apresentou aumento no poder, mesmo em situações em que os modelos são muito diferentes, ou seja, em situações em que a diferença entre ϕ_a e ϕ_b é grande. No caso de DWT, aparentemente sua utilização não é adequada para este método. Um estudo semelhante foi realizado considerando-se agora quatro séries no primeiro grupo geradas a partir de um processo ARMA(1,1), com coeficientes autorregressivo e de média móvel fixos, a saber, $\phi_a = 0.4$ e $\theta = 0.5$. As séries do segundo grupo são geradas também a partir de um processo ARMA(1,1), com o mesmo coeficiente de média móvel $\theta = 0.5$, mas com coeficiente autorregressivo ϕ_b variando no conjunto $\{-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$. Novamente, quando $\phi_a = \phi_b = 0.4$ tem-se a situação de homogeneidade entre os grupos (H_0), e, ao nível de 5% de significância, espera-se que o teste rejeite H_0 aproximadamente 5% das vezes. Os resultados exibidos na figura 1b mostram que neste caso as medidas $INTPER$ e PIC apresentam um desempenho equivalente e que a ACF é melhor que PER e CID. As medidas COR e $CORT$ continuam apresentando um fraco desempenho e DTW apresenta os mesmos problemas do caso anterior.



(a) AR x AR

(b) ARMA x ARMA

Figura 1: Estimativa do poder do teste para homogeneidade de dois grupos de séries temporais. No painel (a), o primeiro grupo contendo 4 séries é gerado a partir de um processo AR(1) com $\phi_a = 0.4$. As séries do segundo grupo são geradas também a partir de um processo AR(1), mas com coeficiente variando. O tamanho de cada série é $T = 512$. No painel (b), o primeiro grupo contendo 4 séries é gerado a partir de um processo ARMA(1,1) com coeficientes autorregressivo e de média móvel fixos em $\phi_a = 0.4$ e $\theta = 0.5$. As séries do segundo grupo são geradas também a partir de um processo ARMA(1,1), com $\theta = 0.5$ mas com coeficiente autorregressivo variando. Em ambos os casos, o tamanho de cada série é $T = 512$ e as métricas testadas foram ACF, PIC, CID, COR, $CORT$, $INTPER$, PER e DWT.

O método de classificação e agrupamento proposto por [1] depende diretamente da capacidade da medida de dissimilaridade diferenciar dois grupos distintos. Realizamos um estudo para testar a performance do método de classificação utilizando todas as medidas apresentadas na seção 2. O primeiro grupo de quatro séries é gerado a partir de um modelo AR(1) com coeficiente $\phi_a = 0.4$ e o segundo grupo com 4 séries é gerado a partir de um processo AR(1), mas com coeficiente autorregressivo $\phi_b \in \{-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$. Uma série extra é gerada a partir do primeiro modelo e então classificada utilizando o método baseado em U-estatística com as métricas apresentadas na seção 2. Podemos observar na tabela 1 que exibe o percentual de acerto para classificação de uma série

Tabela 1: Percentual de acerto para classificação de uma série temporal. O primeiro grupo de quatro séries é gerado a partir de um modelo AR(1) com coeficiente $\phi_a = 0.4$ e o segundo grupo com 4 séries é gerado a partir de um processo AR(1), mas com coeficiente autorregressivo $\phi_b \in \{-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8\}$. Uma série extra é gerada a partir do primeiro modelo e então classificada utilizando o método baseado em U-estatística com as métricas apresentadas na seção 2.

ϕ_b	ACF	PIC	CID	COR	$CORT$	$INTPER$	PER	DWT
-0.8	100	100	100	52.1	63.2	100	100	100
-0.6	100	100	100	51.5	53.1	100	100	100
-0.4	100	100	100	49.2	50.3	99.9	99.9	100
-0.2	100	100	100	47.9	49.4	100	97.3	100
0.2	88.4	97.4	76	52.7	52	97.2	62	100
0.4	48.2	48.6	50	54.4	51.8	50.7	46.8	99.9
0.6	91.9	98	74.9	49.3	52.1	98.1	69.6	99.9
0.8	99.8	99.8	96.7	49.9	67.8	99.9	99.5	99.4

temporal, que as medidas que apresentaram melhor desempenho relativamente ao poder do teste, é que possuem uma capacidade maior de classificar corretamente, que é o caso da $INTPER$. O resultado da DWT não está correto, pois para $\phi_b = 0.4$, o percentual de acerto deveria ser aproximadamente 50%.

4 Conclusão

Neste trabalho, estudamos o comportamento de algumas medidas de dissimilaridades entre séries temporais quando utilizadas como núcleo da estatística de teste no método de classificação baseado em U-estatísticas. Podemos observar que, no caso em que as séries temporais advindas de processos estacionários, as medidas $INTPER$ e PIC apresentam melhor desempenho, relativamente ao poder do teste, nas diferentes configurações testadas. Isso reflete diretamente na capacidade do método classificar uma série temporal em seu respectivo grupo. Os próximos esforços são direcionados para busca de propriedades assintóticas da U-estatística de teste, quando o núcleo for uma dessas medidas que apresentaram melhor desempenho.

Referências

[1] Valk, M. and A. Pinheiro (2012). Time-series clustering via quasi U-statistics. *Journal of Time Series Analysis*, vol.33(4), 608–619.

[2] Bagnall, A. and Janacek, G. (2005). Clustering Time Series with Clipped Data. *Machine Learning*. vol. 58, n. 2-3, pp. 151–178.

[3] Manso, P.M. (2013). *A package for stationary time series clustering*. Tese de Mestrado. Universidade da Coruña.