Seminário Internacional de Estatística com R:
Inovação e Atuação do Profissional no Mercado
10 e 11 de Maio de 2016

ANAIS DO SER –SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R:
Inovação e Atuação do Profissional no Mercado

ORGANIZADORES
Luciane Ferreira Alcoforado
Orlando Celso Longo

NITERÓI
2016

ISBN 978-85-98026-63-3

UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL
Coordenador: Carlos Alberto Pereira Soares
www.poscivil.uff.br

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA
Chefe: Jony Arrais Pinto Junior
www.est.uff.br

DATAUFF – Núcleo de Pesquisas, Informações e Políticas Públicas
Diretora: Luciane Ferreira Alcoforado
Diretor-Adjunto: André Brandão
www.uff.br/datauff

RUA PASSO DA PÁTRIA 156
SÃO DOMINGOS
NITERÓI – RJ
CEP – 242210-240
TEL. +55 21 2629-5410

ISBN 978-85-98026-63-3

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

S471 Seminário Internacional de Estatística com R: inovação e atuação do profissional no mercado (1.: 2016: Niterói, RJ).
Anais ... / I Seminário Internacional de Estatística com R : inovação e atuação do profissional no mercado ; organizadores Luciane Ferreira Alcoforado, Orlando Celso Longo. – Juiz de Fora, MG : Templo, 2016.
113 p.
Evento realizado no período de 10 a 11 de maio de 2016.

1. Desenvolvimento de software. 2. Estatística. 3. Inovação tecnológica I. Alcoforado, Luciane Ferreira (org.). II. Longo, Orlando Celso (org.). III. Título.

CDD 005.1 (21. ed)

ISBN 978-85-98026-63-3

Templo Gráfica Editora

Endereço: Rua da Glória, 92, Juiz de Fora – MG, 36.035-150

ISBN 978-85-98026-63-3

ISBN 978-85-98026-63-3

ISBN 978-85-98026-63-3

Sumário

ISBN 978-85-98026-63-3

ISBN 978-85-98026-63-3

ISBN 978-85-98026-63-3

Cerimonial de Abertura

Dia 10/05/2016

Senhoras e senhores,

Boa tarde,

O Programa de Pós Graduação em Engenharia Civil da Escola de Engenharia da UFF, o Instituto de Matemática e Estatística da UFF, a Escola Nacional de Ciências Estatísticas do IBGE e o Laboratório de Eventos da Faculdade de Turismo e Hotelaria da Universidade Federal Fluminense tem um imenso prazer em recebê-los para participarem do SER – Seminário Internacional de Estatística com R, fruto da parceria com o Programa de Pós Graduação em Engenharia Civil, o Programa de Pós-Graduação em Administração e o Núcleo de Pesquisas Sociais Aplicadas Informações e Políticas Públicas (DATAUFF).

Solicitamos a todos os presentes que desliguem seus aparelhos celulares.

O SER – Seminário Internacional de Estatística com R visa promover o intercâmbio de conhecimento entre pesquisadores e usuários da linguagem R que, segundo a métrica da IEE Spectrum em 2015, obteve o sexto lugar no ranking das mais utilizadas, e a primeira quando se trata da análise de dados.

O R, projeto da comunidade mundial de colaboradores, vem sendo utilizado pelos alunos do curso de Estatística da UFF desde 2010 e vêm ao encontro das necessidades de reformulação do Programa de Pós-graduação em Engenharia Civil da UFF, promovendo a utilização prática pelos seus alunos e professores nas mais variadas abordagens de problemas que são desenvolvidos no âmbito das pesquisas, especialmente envolvendo modelagem, estatística espacial, simulação e análise de dados.

Este pioneiro evento em Niterói e no Brasil procura inserir-se no calendário anual global dos pesquisadores e práticos de Análise de Dados e se justifica pela enorme importância

ISBN 978-85-98026-63-3

que este software traz para a melhoria das pesquisas acadêmicas realizadas no país.

Chamamos para compor a mesa as seguintes autoridades:

O Magnífico Reitor da Universidade Federal Fluminense, Prof. Dr. Sidney Luiz Mello.
O Professor Vitor Francisco Cadorin representando o Pró-reitor de Extensão da Universidade Federal Fluminense, Prof. Dr. Cresus Vinícius Depes de Gouvêa.
O Pró-reitor de Pró-Reitoria de Graduação – Prof. José Rodrigues Farias Filho.
O Sr. Thiago Renault, representante do Pró-reitor de Inovação Prof. Roberto Kant de Lima.
O Diretor da Escola de Engenharia - Prof. Fábio Passos.
O Professor e Coordenador geral substituto da Escola Nacional de Ciências Estatísticas, José André de Moura Brito, representando a Sra. Wasmália Bivar, Presidente do IBGE e a Coordenadora Geral, Profa. Maysa Magalhães
O Coordenador do mestrado profissional em finanças e do Laboratório de Análise e Modelagem em Ciências aplicadas do Instituto Nacional de Matemática Pura Aplicada (IMPA), Prof. Jorge Passamani Zubelli.
A coordenadora do SER - Seminário Internacional de Estatística com R, Profa. Luciane Alcoforado.
Agradecemos a presença dos professores, alunos, profissionais do mercado e demais convidados.
Para darmos início à esta solenidade, passamos a palavra à coordenadora do SER - Seminário Internacional de Estatística com R, Profa. Luciane Alcoforado.

Palestra de abertura do SER 2016

Profa. Luciane Ferreira Alcoforado

Boa tarde a todos, sejam muito bem-vindos ao Seminário Internacional de Estatística com R, o SER.
Nossa vida profissional começa quando escolhemos a profissão, no meu caso Matemática.

ISBN 978-85-98026-63-3

Em 2006 procurei a ENCE para me atualizar e qualificar. Foi meu primeiro contato com o R. Vi ali muitas possibilidades de realização.

Em 2007 iniciei o doutorado na Pós-Graduação em Engenharia Civil da UFF e utilizei o R para fazer as análises estatísticas da minha tese. Quando terminei o doutorado fui convidada a ministrar a disciplina de Métodos Computacionais para o curso de graduação em Estatística e a disciplina de Matemática Aplicada à Engenharia no programa de Mestrado e Doutorado da Pós Civil. Isso foi no ano de 2010 e eu seria a pessoa que iria introduzir o aprendizado do R para os futuros Estatísticos formados pela UFF, tarefa de grande responsabilidade para com aqueles jovens, muitos deles hoje já atuam no mercado como profissionais. Da mesma forma procedi com a Pós-Graduação.

Iniciei um projeto de monitoria neste mesmo ano de 2010 que produz excelentes materiais didáticos gerando um livro publicado pela Eduff em 2014.

Esse projeto de monitoria impulsionou o desenvolvimento do portal Estatística é com R e a produção de vídeos no ano de 2015. Através de dois projetos de extensão, produzimos diversos materiais contanto apenas com o trabalho voluntário de toda a equipe, formada principalmente por alunos, utilizando espaço cedido pela Pós-Graduação em Engenharia Civil e demais parceiros, mas infelizmente o país passa por uma crise e sem apoio institucional adequado, seja na forma de bolsa para os alunos como de equipamentos para produção e edição dos vídeos, não conseguimos atingir a meta planejada.

Trabalhando junto aos alunos de graduação e pós-graduação percebi que há uma grande lacuna entre a formação acadêmica e a realidade do mercado de trabalho destes profissionais. Assim a demanda pelo SER era concreta e urgente.

Hoje, 10 de maio de 2016, passaram-se exatos 14 anos em que eu deixava o IBGE para tomar posse como professora da UFF. E tenho a felicidade de poder reunir estas duas grandes instituições na abertura deste evento.

Estamos aqui reunidos para testemunharmos o surgimento de um Seminário pioneiro no Brasil, o SER. Tudo isso aconteceu graças às parcerias institucionais que nos apoiaram e acreditaram no projeto SER, mas em especial gostaria de agradecer imensamente ao Prof. Orlando Longo que foi incansável na busca por soluções que tornaram esse evento possível. Seu apoio e incentivo nos levou a chegarmos mais longe do que poderíamos

ISBN 978-85-98026-63-3

imaginar no início dos trabalhos. Também não poderia deixar de agradecer ao LEVE, o Laboratório de Eventos da Faculdade de Turismo e Hotelaria, que esteve presente desde o início do projeto SER, ao prof. Ariel e demais membros da Comissão Organizadora e Científica e ao NAB por ceder este espaço, além de todos os demais envolvidos que se juntaram a nós posteriormente.

Agradeço às autoridades presentes e aos participantes que atenderam ao chamado deste evento e espero que todos possam sair daqui com novos conhecimentos e perspectivas quanto a aplicabilidade da linguagem R.

Um bom SER a todos! Muito Obrigada.

Falará agora o Coordenador do mestrado profissional em finanças e do Laboratório de Análise e Modelagem em Ciências aplicadas do Instituto Nacional de Matemática Pura Aplicada (IMPA), Prof. Jorge Passamani Zubelli.

Com a palavra o Professor e Coordenador geral substituto da Escola Nacional de Ciências Estatísticas, José André de Moura Brito, representando a Sra. Wasmália Bivar, Presidente do IBGE e a Coordenadora Geral, Profa. Maysa Magalhães.

Ouviremos agora o Diretor da Escola de Engenharia - Prof. Fábio Passos.

Passamos a palavra ao Sr. Thiago Renault, representante do Pró-reitor de Inovação Prof. Roberto Kant de Lima;

Passaremos a ouvir o Professor Vitor Francisco Cadorin representando o Pró-reitor de Extensão da Universidade Federal Fluminense, Prof. Dr. Cresus Vinícius Depes de Gouvêa;

Encerrando esta Mesa de abertura, tem a palavra o Magnífico Reitor da Universidade Federal Fluminense, Prof. Dr. Sidney Luiz Mello.

Neste momento encerramos esta mesa de abertura e solicitamos às autoridades presentes que ocupem os seus lugares na plateia para darmos prosseguimento ao SER - Seminário Internacional de Estatística com R, Profa. Luciane Alcoforado.

14h - Convidamos a Profa. Luciane Alcoforado, coordenadora geral do SER - Seminário

ISBN 978-85-98026-63-3

Internacional de Estatística com R, para apresentar sua palestra intitulada Estatística é com R: ações para o aprendizado do R.

Formada em Matemática pela Universidade Federal de Santa Maria (UFSM), possui doutorado em Engenharia Civil pela UFF, mestrado em Engenharia de Sistemas pela (COPPE/UFRJ). Atua no departamento de Estatística da UFF desde 2002 e no Programa de Pós-Graduação em Engenharia Civil da UFF desde 2010. Atualmente é diretora do Núcleo de Pesquisas Sociais, Informações e Políticas Públicas (DATAUFF), coordenadora do portal Estatística é com R! e coordenadora de monitoria do departamento de Estatística.

Agradecemos a profa. Luciane Alcoforado pela excelente apresentação, agora abriremos 10 min para 3 perguntas da plateia

14h30 - Seguindo a programação, convidamos o Prof.Dr Manuel Febrero para fazer apresentar a sua palestra intitulada Dynamic reports in R (with LateX or HTML)

O Professor Manuel Febrero é professor de Estatística e Pesquisa Operacional da Universidade de Santiago de Compostela, na Espanha. Ele recebeu o seu B.S. degree em Matemática em 1990 e o título de Ph.D. em Estatística em 1995 pela Universidade de Santiago de Compostela, A Coruña, Espanha. Ele publicou em séries temporais, de inicialização, os dados funcionais e em métodos estatísticos aplicados ao meio ambiente. É também o coordenador acadêmico do Programa Interuniversitário PhD em Estatística e Op. Res. organizado em conjunto pelas universidades de Santiago de Compostela, A Coruña e Vigo.

Agradecemos ao Prof.Dr Manuel Febrero pela sua valiosa colaboração e abrimos para 3 perguntas da plateia.

15h40 - Neste momento convidamos a todos para degustar o lanche oferecido pela Escola Nacional de Ciências Estatísticas – ENCE. Retomaremos os trabalhos em 20 minutos.

16h10 – Neste momento, o Prof. Dr. Pedro Guilherme Costa Ferreira, da Fundação Getúlio Vargas apresentará o TED - Pesquisa aplicada com o R: BETS package e

ISBN 978-85-98026-63-3

Indicador de Incerteza da Economia (IIE-Br)

O Prof. Pedro Costa Ferreira é Doutor em Engenharia Elétrica - (Decision Support Methods) e Mestre em Economia. É o primeiro pesquisador da América Latina a ser recomendado pela empresa RStudio Inc. Atuou em projetos de Pesquisa e Desenvolvimento (P&D) no setor elétrico nas empresas Light S.A. (e.g. estudo de contingências judiciais), Cemig S.A, Duke Energy S.A, entre outras. Ministrou cursos de estatística e séries temporais na PUC-Rio e IBMEC e em empresas como a Operador Nacional do Setor Elétrico (ONS), Petrobras e CPFL S.A. Atualmente é professor de Econometria de Séries Temporais e Estatística e coordenador do Núcleo de Métodos Estatísticos e Computacionais na Fundação Getúlio Vargas (FGV|IBRE).

Agradecemos ao prof. Pedro Costa pela excelente apresentação e abrimos para 3 perguntas da plateia.

16h40 – Convidamos o Prof. Dr. Djalma Galvão Carneiro Pessoa para apresentar a palestra intitulada: Análise de dados amostrais utilizando a library survey do R

O Prof. Dr. Djalma Galvão Carneiro Pessoa é PhD em Estatística pela Universidade da Califórnia – Berkeley e Pós-doutor pela Universidade de Stanford, na Califórnia. Foi o primeiro Presidente da Associação Brasileira de Estatística, Superintendente da Escola Nacional de Ciências Estatísticas – ENCE, Diretor Executivo do IBGE e Consultor em Estatística do IBGE. Atualmente é colaborador do site asdfree.com nos blogs contendo análise de dados de pesquisas domiciliares do IBGE.

Agradecemos pela excelente apresentação, agora abriremos 10 minutos para 3 perguntas da plateia.

17h40 - Encerrando nossas atividades de hoje, teremos agora a Mesa Redonda " Os avanços da linguagem R: das pesquisas acadêmicas às grandes empresas.

Para fazer a mediação chamamos o Prof.Dr Ariel Levy. Chamamos agora os professores:

Jorge Passamani Zubelli
Manuel Febrero (ES),
Djalma Pessoa (ASDFREE/IBGE),
Steven Ross(UNIRIO),
Pedro Guilherme (FGV)

ISBN 978-85-98026-63-3

Agradecemos à todos os professores pelo ótimo debate e a todos presentes. Encerramos as atividades do primeiro dia do Seminário Internacional de Estatística com R.

Tenham todos uma boa noite.

Dia 11/05/2016

Boa tarde senhoras e senhores,
Daremos prosseguimento à programação do SER – Seminário Internacional de Estatística com R

13h30 – Convidamos para apresentar o blog "Paixão por Dados", Sr. Sillas Gonzaga.

13h40 – Agradecemos ao Sr. Sillas Gonzaga e assistiremos agora a apresentação do blog "GAE" pelos estudantes de gaduação na Universidade Federal do Estado do Rio de Janeiro (UNIRIO) Adriano Mourthé.

13h50 – Agradecemos pela apresentação e convidamos agora os graduandos em Estatística pela Faculdade Federal Fluminense, Leonardo Filgueira e Camila Simões para apresentar o blog " Estatística é com R"
Agradecemos ao Leonardo Filgueira e Camila Simões pela contribuição

14h00 Neste momento faremos a entrega do prêmio do 3 melhores pôsteres, um oferecimento do nosso parceiro SBBnet, fornecendo as medalhas dos respectivos ganhadores e descontos de 8% à todos para compras no site, utilizando a #SOUSER2016.

A comissão avaliadora foi composta pelos professores José Rodrigo de Moraes, Steven Ross, Maysa Magalhães e Orlando Longo. Convidamos os professores José Rodrigo de Moraes e Wenceslao González Manteiga para entregar as medalhas aos ganhadores.

O 3º melhor trabalho foi dos participantes Hugo M Rodrigues e Gustavo Marques, intitulado "Propriedades eletromagnéticas mostram potencial para mapear atributos do solo correlacionado em R"

O pôster classificado em 2° lugar tem como título "A tarifação de um plano de saúde autogestão aplicando os modelos lineares generalizados utilizando o R" dos participantes Helano Silva Eugênio de Souza e Luiz Carlos da Silva Leão.

Em 1º lugar, vencedor como melhor trabalho da sessão pôster do "Seminário Internacional de Estatística com R" vai para os participantes Júlio César de Azevedo Vieira e Jony Arrais Junior, autores do pôster "Métodos de Seleção de Variáveis via

ISBN 978-85-98026-63-3

Verossimilhança Penalizada"

Parabéns aos vencedores pelos ótimos trabalhos!

14h10 – Neste momento chamamos para apresentar sua palestra "Biologia Computacional usando a linguagem R" o Prof. Dr. Leonardo Soares Bastos, da Fiocruz. Estatístico bayesiano graduado em estatística pela UFMG, mestre em estatística pela UFRJ e doutor em estatística pela University of Sheffield, no Reino Unido. É pesquisador em saúde pública na Fundação Oswaldo Cruz atuando nos programas de pós-graduação de Biologia Computacional e Sistema (IOC/Fiocruz) e Epidemiologia em Saúde Pública (ENSP/Fiocruz). As principais linhas de pesquisa são modelagem de epidemias, inferência em populações de difícil acesso, e recentemente modelagem estrutural de proteínas.

Agradecemos pela excelente apresentação e abrimos 10 minutos para 3 perguntas da plateia.

14h30 - Agradecemos ao prof. Leonardo Soares Bastos e assistiremos agora a videoconferência do Prof. Dr. Sean Kross diretamente da Universidade de Maryland, nos Estados Unidos. O título de sua palestra é "The swirl R package: Learn R in R".

O Prof. Dr. Sean Kross possui BA em Biologia pela New York University, BS em Ciência da Computação pela Universidade de Maryland classe de 2015. Co-criador do Swirl R Package para a aprendizagem de R em R e Co-criador da Especialização Johns Hopkins Data Science no Coursera.org. Também foi Co-criador da Computer-Science-in-a-Box da Universidade de Maryland no Centro para Mulheres em Computação, assistente de Pesquisa da Universidade de Maryland no Centro de Bioinformática e Biologia Computacional em Metagenômico Taxonomia, Assistente de Pesquisa da Universidade de Maryland Smith School of Business e Assistente de Pesquisa na Universidade Central de Nova York para estudos de genômicos, biologia de sistemas e microbiologia

Agradecemos ao Prof. Dr. Sean Kross pela valiosa contribuição, abriremos para 3 perguntas

ISBN 978-85-98026-63-3

15h - Neste momento chamamos a Dra. Maria Luiza Guerra de Toledo da ENCE/IBGE que falará sobre os " Modelos de Confiabilidade no R"

A Dra. Maria Luiza Guerra de Toledo possui doutorado em Engenharia de Produção pela Universidade Federal de Minas Gerais (2014), Bacharelado (2005) e Mestrado (2007) em Estatística pela Universidade Federal de Minas Gerais. É pesquisadora e professora na Escola Nacional de Ciências Estatísticas (Ence) do IBGE, onde atua na Graduação em Estatística, e no Programa de Pós-Graduação em População, Território e Estatísticas Públicas. Suas áreas de interesse são Probabilidade e Estatística aplicada à Análise de Sobrevivência, Confiabilidade e Manutenção.

Agradecemos à Profª Maria Luiza Guerra de Toledo pela ótima apresentação e abrimos 10 min para 3 perguntas da plateia

15h30 - Faremos agora uma pausa de 20 minutos para o coffee breack oferecido pela ENCE, que está sendo servido na área externa do NAB.

16h - Dando continuidade à nossa programação, convidamos o Prof. Dr. Gustavo da Silva Ferreira da Escola Nacional de Ciências Estatísticas (ENCE-IBGE) para falar sobre "Planejamento amostral em Geoestatística com R".

O prof. Gustavo Ferreira possui graduação em Estatística pela Universidade Federal do Rio Grande do Sul - UFRGS (2002), mestrado em Estatística pela Universidade Federal do Rio de Janeiro - UFRJ (2004) e doutorado em Estatística pela Universidade Federal do Rio de Janeiro - UFRJ (2013). Além de ter atuado como estatístico na ELETROBRAS e PETROBRAS, atualmente é pesquisador da Escola Nacional de Ciências Estatísticas - ENCE onde desenvolve pesquisas nas áreas de Estatística Espacial, Planejamento Amostral Ótimo, Epidemiologia e Modelos para Redes Sociais.

Agradecemos ao Prof. Gustavo Ferreira pela valiosa apresentação e abrimos para 3 perguntas

16h30 - Neste momento a Palestra "Estatísticas Espaciais no R: limites e possibilidades"

ISBN 978-85-98026-63-3

será proferida pelo Prof. Dr. Alexandre Souza da Silva da UNIRIO.

Prof. Alexandre é - Professor Adjunto do Departamento de Matemática e Estatística da UNIRIO. Toda a sua formação é em Estatística, com Graduação na UNESP, Mestrado em Estatística e Experimentação Agronômica pela USP e Doutorado na UFRJ. Possui experiência na área de Probabilidade e Estatística, com ênfase em Análise Espacial, Modelos Espaço-Temporais Bayesianos e Educação Estatística.

Agradecemos pela valiosa contribuição do Prof. Alexandre Silva e abrimos 10min para 3 perguntas

17h - Ouviremos agora o Prof. Dr. Rodrigo Otávio de Araújo Ribeiro (diretor IBOPE-DTM/UERJ) que falará sobre “BIG DATA Analytics e o R”.

Prof. Dr. Rodrigo Otávio de Araújo Ribeiro possui graduação em Estatística pela Escola Nacional de Ciências Estatísticas (2003), mestrado (2007) e doutorado (2012) em Engenharia de Produção pela Universidade Federal Fluminense. Atualmente lidera a área de Inteligência de Marketing no IBOPE DTM, vencedor do Prêmio Alfredo Carmo em 2014, a premiação mais importante no mercado de pesquisa do Brasil, por ter sido autor principal do artigo "Análise de usuários que conversam sobre cerveja no Twitter", considerado o melhor trabalho apresentado no Sexto Congresso de Brasileiro de Pesquisa, organizado pela ABEP. Atua também como professor adjunto do Instituto de Matemática e Estatística da Universidade do Estado do Rio de Janeiro.

Agradecemos ao Prof. Rodrigo Otávio pela excelente apresentação e abrimos para 3 perguntas da plateia.

17h30 - Ouviremos agora os senhores Philipe Rabelo e Savano Pereira, representantes da empresa Mobi2by

Agradecemos pela valiosa contribuição.

Agora chamamos o prof. Orlando Celso Longo, para as considerações finais.

Em nome da Universidade Federal Fluminense, agradecemos a presença de todos e os convidamos a confraternizar conosco o sucesso deste evento que, esperamos, seja o

ISBN 978-85-98026-63-3

primeiro de uma série.

O Laboratório de Eventos da Faculdade de Turismo e Hotelaria deseja a todos uma boa noite.

ISBN 978-85-98026-63-3

Avaliação das externalidades das políticas públicas: o caso dos efeitos não previstos do programa bolsa família.

Roberta Daniela Costa Botelho (UNIRIO) / e-mail: roberta.cbotelho@gmail.com
Flavia dos Santos (UNIRIO) / e-mail: dossantosflavia1996@gmail.com
Steven Dutt-Ross (UNIRIO) / e-mail: steven.ross@uniriotec.br

1. INTRODUÇÃO

O programa Bolsa Família, de responsabilidade do governo federal brasileiro, tem como objetivo beneficiar famílias em situação de extrema pobreza, com a transferência direta de renda. Entretanto, desde a sua implementação, em 2003, diversas críticas vêm sido construídas ao seu redor, em que se questiona a sua veracidade e autonomia. Dentre as principais, está a atribuição errônea de que os mais pobres possuem um comportamento oportunista em relação à maternidade, segundo a ministra do Desenvolvimento Social e Combate à Fome, Tereza Campello. Esse pensamento comum, no entanto, desvirtua-se dos resultados da Pesquisa Nacional de Amostra por Municípios (Pnad), que demonstrou que as famílias que recebem o benefício do Bolsa Família tiveram menos filhos que a média brasileira entre os anos de 2003 e 2013. Além disso, também é questionado o avanço proveniente e a dependência para os 13.732.792 milhões de brasileiros que o usufruem (IBGE, 2015), já que seu uso pode acarretar efeitos não previstos e que não dispõem de políticas públicas de prevenção e controle, como o caso da taxa de fecundidade.

2. OBJETIVO

O principal objetivo deste trabalho é, com o auxilio do programa R (R CORE, 2016), analisar, mediante testes não-paramétricos, se há alguma relação linear entre duas variáveis em estudo: a taxa de fecundidade total - TFT- das famílias brasileiras e o número de beneficiários do programa Bolsa Família -PBF-, ambas no período analisado de 2007 a 2009.

3. MÉTODO

Todos os dados sendo coletados a partir de um período de três anos, correspondentes **aos anos de 2007, 2008 e 2009**. Foi escolhido esse período por ser considerado, tecnicamente, atualizado, recente e em virtude

ISBN 978-85-98026-63-3

do programa debatido ter sido implementado em 2005. Os dados coletados foram retirados do site do IBGE (Instituto Brasileiro de Geografia e Estatística) e do IPEADATA. Nesse estudo, foi criado um novo indicador: o número relativo de beneficiários do programa Bolsa Família. Esse indicador é a razão entre o total de beneficiários do PBF e o total da população. Em outras palavras, é o número de beneficiários do PBF *per capita*. Um fato importante a ser ressaltado é que a validade dos resultados obtidos através dos testes de hipótese paramétricos é fortemente dependente da normalidade dos dados analisados (RODRIGUES; IEMMA, 2005). Assim, foi realizado o teste ***Shapiro-Wilk*** para verificar a normalidade da amostra. Por meio deste, o p-valor encontrado foi **0,0006422,** isto é, menor que o nível de significância, portanto, trata-se de um caso em que a distribuição dos dados não é normal e que o estudo acerca da proporção de beneficiários em nível regional deve basear-se em testes não-paramétricos, como o ***Kruskal-Wallis*** e o ***Spearman***. O primeiro é usado para testar a hipótese de que todas as populações possuem funções de distribuição iguais contra a hipótese alternativa de que ao menos duas das populações possuem funções de distribuição diferentes (KRUSKAL; WALLIS, 1952). Para tal teste, tem-se as seguintes hipóteses: Existe diferenças regionais no PBF. Já o teste de correlação *Spearman* é utilizado para verificar a associação linear entre o gasto com o programa e a taxa de fecundidade. Neste sentido, foi realizada a análise entre o número de beneficiários relativos à população total e a TFT, para que fosse possível averiguar a existência da correlação entre ambas variáveis.

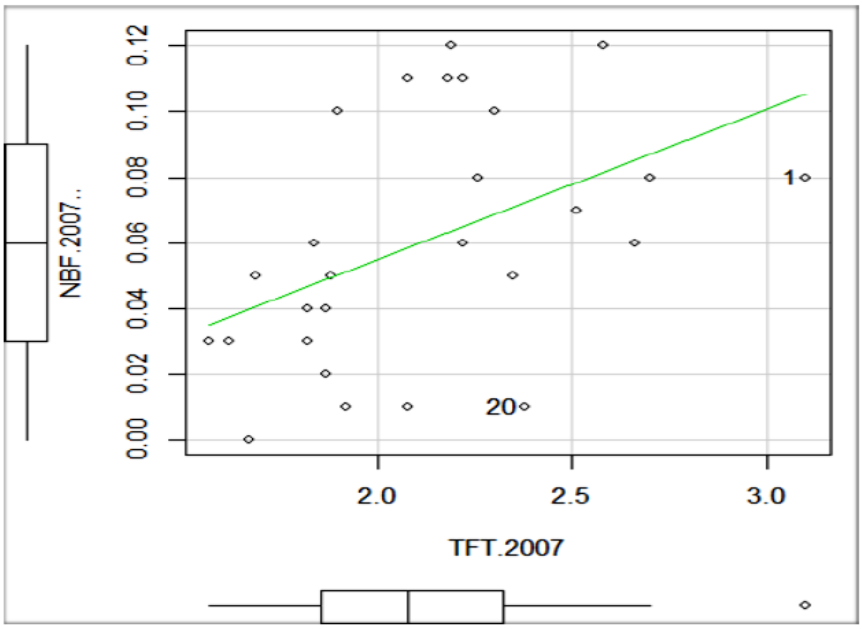
4. RESULTADOS

No teste Kruskal-Wallis, o p-valor de 0,04231 mostra que há uma região com maior proporção de beneficiários: o Nordeste. Para tornar mais visível geograficamente esta distribuição heterogênea, foram construídos mapas, através do uso conjunto do programa R com o API Google, acerca do número relativo de beneficiários por estado, nos anos de 2007 e de 2008. Assim, é perceptível que as regiões Norte e Nordeste, principalmente a segunda, são detentoras da maior proporção de famílias beneficiárias do Bolsa Família, dentre a população total. Este evento ocorre sobretudo por fatores econômicos.

Em 2013, a região nordestina detinha a maior porcentagem de pessoas abaixo da linha da extrema pobreza, cerca de 10%. No entanto, segundo o IPEA, em uma década o número de miseráveis no Nordeste caiu a quase um terço, e uma das principais causas foi o auxílio do Bolsa Família.

ISBN 978-85-98026-63-3

Gráfico 1- Diagrama de dispersão que relaciona o número relativo de beneficiários com a taxa de fecundidade total em 2007.

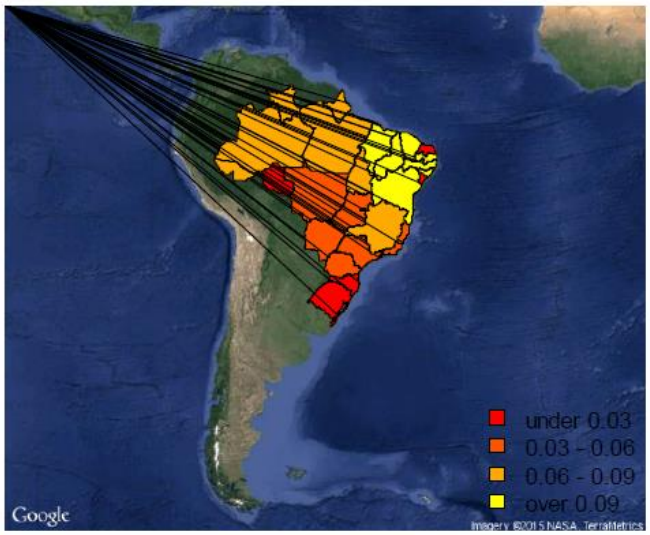


Estes dados sugerem que o programa atua de maneira eficaz quanto à retirada de indivíduos da linha de extrema pobreza. Por fim, para verificar a validade da correlação entre beneficiários do Bolsa Família e número de filhos por mulher, foi utilizado o **teste de correlação de *Spearman***. Tem-se como resultado: : **ρ = 0,516 e** o **p-valor = 0,005916.** Assim, o p-valor sugere a existência de uma correlação entre ambas variáveis, bem como o diagrama de dispersão expresso no gráfico 01. No entanto, por se tratar de um estudo exploratório com dados agregados no período de 2007/2009, esses resultados diferem da pesquisa do IBGE sobre o tema (IBGE, 2013). Desse modo, pesquisas futuras devem verificar de forma mais

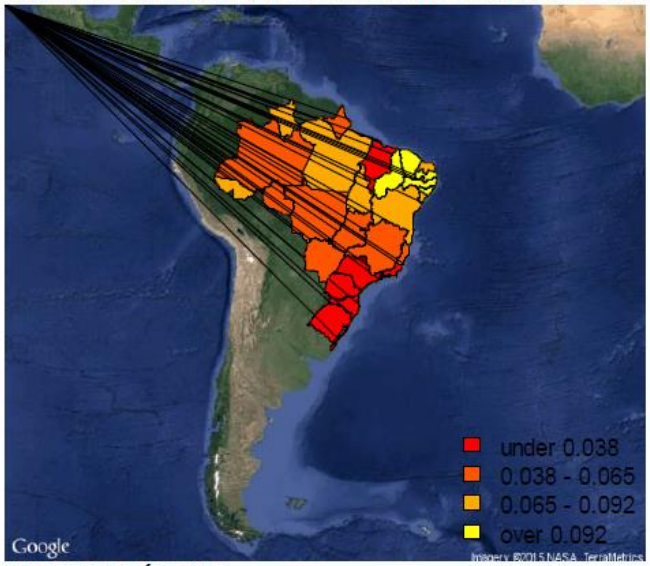
ISBN 978-85-98026-63-3

aprofundada e controlando por outras variáveis a relação entre o PBF e a TFT.

Mapa 01 - Número relativo de beneficiários do Bolsa Família em 2007



Mapa 02 - Número relativo de beneficiários do Bolsa Família em 2008



REFERÊNCIAS BIBLIOGRÁFICAS

IBGE, PNAD Pesquisa Nacional por Amostra de Domicílios. Rio Janeiro, 2013

IBGE, PNAD Pesquisa Nacional por Amostra de Domicílios. Rio Janeiro, 2015

R DEVELOPMENT CORE TEAM, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/

RODRIGUES, M. I.; IEMMA, A. F. Planejamento de Experimentos e Otimização de Processos: uma estratégia seqüêncial de planejamentos, Campinas, SP, Casa do Pão Editora, 2005.

ISBN 978-85-98026-63-3

Avaliação das unidades de atenção primária à saúde no município do Rio de Janeiro segundo os resultados do PMAQ 2012.

Langs de Arantes F. de Mello (UNIRIO) / email: langsmello@live.com
Alexandre Souza da Silva (UNIRIO) / email: alexandre.silva@uniriotec.br
Steven Dutt-Ross (UNIRIO) / email: steven.ross@uniriotec.br
Luciane de S. Velasque (UNIRIO) / email: luciane.velasque@uniriotec.br

INTRODUÇÃO: A partir de 2009 o município do Rio de Janeiro passou por uma marcante reforma na Atenção Primária à Saúde. A cobertura da Estratégia de Saúde da família saltou de 7% em 2009 para 40% em 2012. No intuito de avaliar a qualidade da rede APS, o ministério da saúde através da portaria nº 1.654, de 2011 instituiu o PMAQ, uma avaliação nacional das equipes de saúde da família.

O PMAQ é composto por quatro fases distintas e complementares entre si: Adesão e contratualização; desenvolvimento; avaliação externa; e recontratualização. Desta forma, se constitui como um processo cíclico promovendo a institucionalização da avaliação em saúde no contexto da atenção primária. Neste sentido, o presente trabalho tem como **objetivo** apresentar a distribuição espacial dos indicadores de qualidade do PMAQ no município do Rio de Janeiro.

METODOLOGIA: Estudo ecológico com dados secundários provenientes do Programa Nacional de Melhoria do Acesso e da Qualidade da Atenção Básica (PMAQ) do Município do Rio de Janeiro no ano de 2012, junto ao Ministério da Saúde via Lei de Acesso a Informação e dados públicos.

A análise dos dados foi realizada com software estatístico R versão 3.2.1, onde foi realizada a análise descritiva e espacial dos dados. Para verificar a existência associação entre os tipos de UBS e o resultado encontrado na avaliação do PMAQ foi utilizado o teste exato de Fisher.

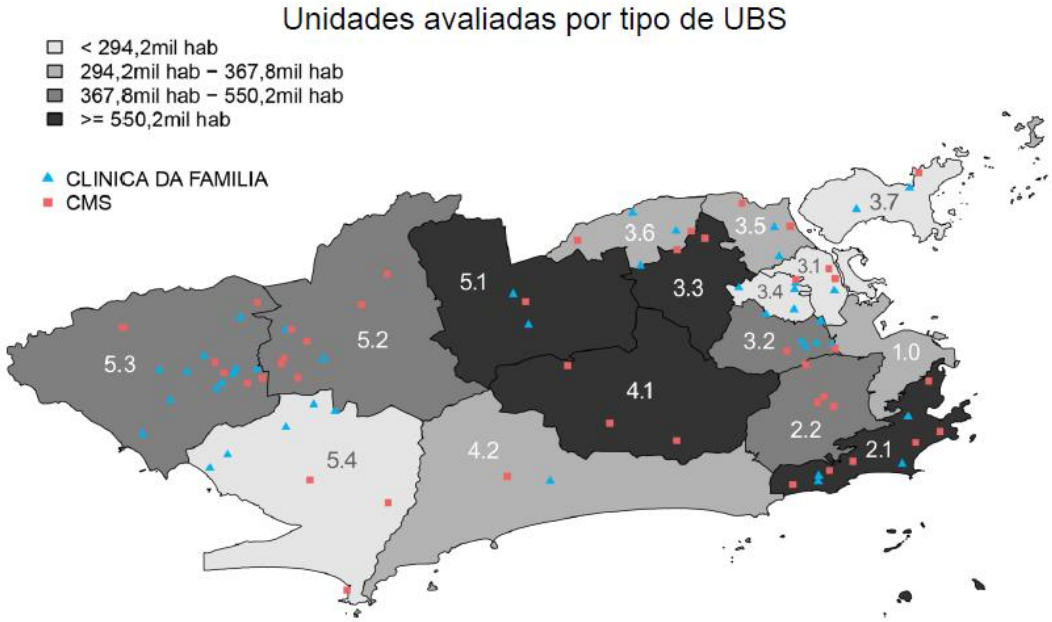
ISBN 978-85-98026-63-3

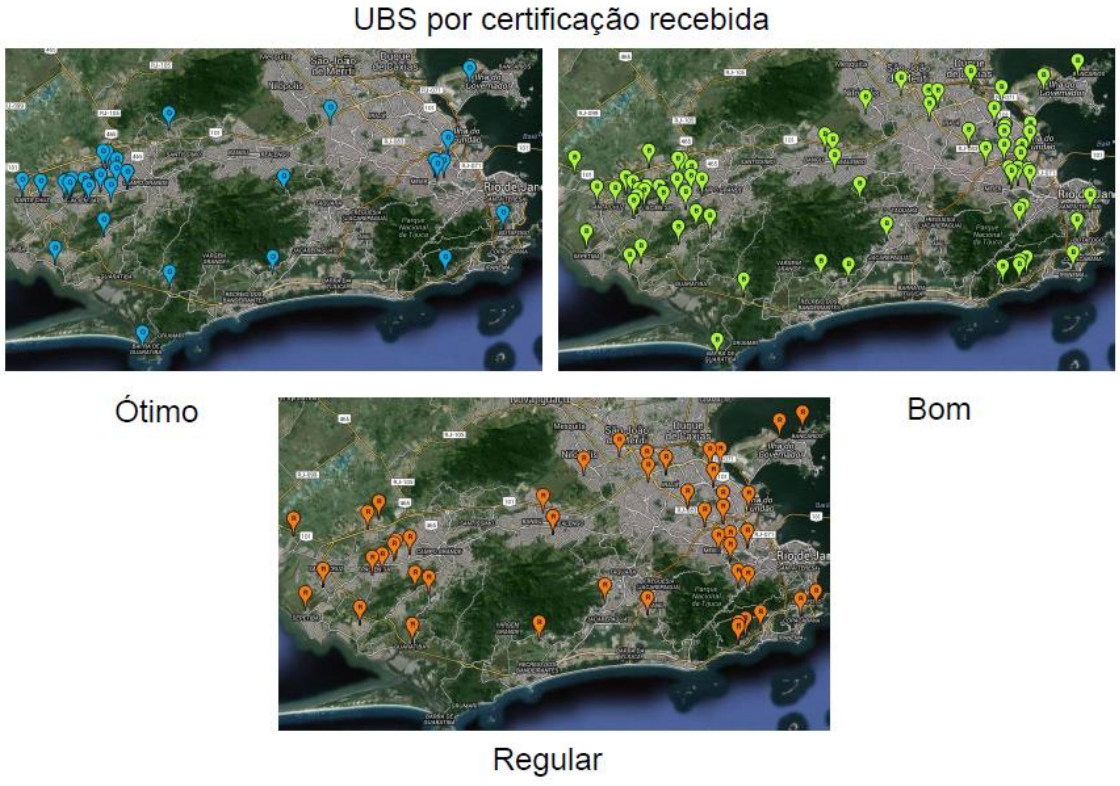
Os indicadores obtidos pelo PMAQ foram apresentados em mapa para permitir a visualização espacial da distribuição dos mesmos. Este projeto foi aprovado pelo Comitê de Ética e Pesquisa com Seres Humanos (CEP-UNIRIO) sob protocolo 952.274 em conformidade com a resolução CNS Nº 466, de 12 de dezembro de 2012.

Os mapas foram gerados utilizando o software estatístico R com pacotes específicos para cada tipo de mapa. Os mapas coropléticos foram criados a partir dos *shapefiles* disponíveis no portal do Sistema Municipal de Informações Urbanas, órgão da prefeitura do município do Rio de Janeiro. Os pacotes utilizados no R para gerar os mapas coropléticos foram o "maptools", "spdep", "stringr" e "rgdal". Os dados populacionais são do censo IBGE 2010. Para o mapa de satélite, foram utilizados os pacotes "RgoogleMaps", "googleVis", "plotGoogleMaps", "dismo", "ggmap" e "XML" pacotes estes que utilizam uma API do Google Maps para seu funcionamento, tendo o "Google Maps" como base para gerar os mapas.

RESULTADOS: No município do Rio de Janeiro, 65% das equipes obtiveram desempenho "bom" ou "ótimo", 34,7% "regular" e 0,3% "insatisfatório". Apesar de uma discreta tendência de desempenho superior das equipes vinculadas a clínicas da família, não houve associação entre os indicadores e o tipo de unidade a qual a equipe avaliada estava vinculada (p-valor = 0,119).

ISBN 978-85-98026-63-3





ISBN 978-85-98026-63-3

Nota-se uma concentração das equipes classificadas como “ótimas” na zona oeste, equipes classificadas como “boas” mais amplamente distribuídas no território e as classificadas como “regular” concentradas na zona norte e centro. A ausência de uma amostragem aleatória pode representar viés de seleção, pois dá margem para o gestor escolher equipes com melhor desempenho potencial, especialmente ao considerar que uma avaliação positiva no PMAQ aumenta o repasse do governo federal à equipe através do Piso da Atenção Básica variável.

Ponderamos que os resultados obtidos demonstram entraves inerentes ao contexto territorial e populacional de cada região, mas não afasta a necessidade da melhoria da qualidade do atendimento ofertado em especial nas zonas norte e oeste

Referências: BRASIL. **Portal do Departamento de Atenção Básica.** 2015. Disponível em: < http://dab.saude.gov.br/portaldab/historico_cobertura_sf.php >. Acesso em: 07/12/2015.
FAUSTO, M. C. R.; FONSECA, H. M. S. **Rotas da atenção básica no Brasil: experiências do trabalho de campo PMAQ AB.** Rio de Janeiro: Saberes: 32-59 p. 2014.
HARZHEIM, E.; LIMA, K. M.; HAUSER, L. **Reforma da atenção primária à saúde na cidade do Rio de Janeiro: avaliação dos três anos de Clínicas da Família**. Organização Pan-Americana da Saúde, 2013.
IBGE. **Sinopse do Censo Demográfico 2010**. Rio de Janeiro: 2011. 261 p Disponível em: < http://biblioteca.ibge.gov.br/visualizacao/livros/liv49230.pdf >. Acesso em: 14/07/2015.
LIMA, Maria Alice Dias da Silva et al. Acesso e acolhimento em unidades de saúde na visão dos usuários. **Acta paul enferm**, v. 20, n. 1, p. 12-7, 2007. Disponível em: <http://www.scielo.br/pdf/ape/v20n1/a03v20n1 >. Acesso em: 18 jan. 2016.
MARQUES, Giselda Quintana; LIMA, M. A. D. S. Demandas de usuários a um serviço de pronto atendimento e seu acolhimento ao sistema de saúde. **Rev Latino-am Enfermagem**, v. 15, n. 1, p. 13-9, 2007. Disponível em: < http://www.scielo.br/pdf/rlae/v15n1/pt_v15n1a03.pdf > Acesso em: 18 jan. 2016.
MINISTÉRIO DA SAÚDE. **Portaria Nº 1.654 de 19 de julho de 2011**. Brasil 2011.

ISBN 978-85-98026-63-3

Planejamento amostral ótimo em geoestatística usando o R.

Catarina Dall'Agnol Zidde (ENCE/IBGE) / e-mail: catarinazidde@gmail.com
Gustavo da Silva Ferreira (ENCE/IBGE) / e-mail: gustavo.ferreira@ibge.gov.br

Introdução

O processo de escolha ótima de novos locais para amostragem é bastante difundido na literatura de Estatística e sua aplicação no contexto de Geoestatística possui inúmeros artigos recentes, como os trabalhos de Zidek et al. (2000), Fernandez et al. (2005), Zhu e Stein (2005), Diggle e Lophaven (2006), Gumprecht et al. (2007), entre outros.

Dentre os avanços recentes, destacam-se aqueles baseados na maximização de funções utilidades, os quais são de grande utilidade no contexto da Geoestatística e que fazem uso intensivo de métodos de simulação estocástica.

Em virtude da complexidade envolvida no planejamento amostral, torna-se necessário criar procedimentos e rotinas que compatibilizem os múltiplos objetivos do pesquisador e, ao mesmo tempo, garantam a produção de uma solução de qualidade.

Neste trabalho objetiva-se desenvolver rotinas em linguagem R para automatizar as etapas envolvidas na realização de um planejamento amostral em uma região do plano bidimensional.

Metodologia

Podemos supor que estamos interessados em estudar as características de um processo estocástico $\{S(x): x \in D\}$, onde D representa uma região qualquer no $\mathbb{R}^2$. Adicionalmente, podemos supor que S(x) possui variância σ^2 e função de autocorrelação $\rho(S(x), S(x+h); \phi)$, $x \in D$, que pode depender de um ou mais parâmetros ϕ.

Para a elaboração das rotinas, assumiu-se que o processo do qual deriva a amostra original é gaussiano estacionário isotrópico. Assim, temos que a função de autocorrelação só depende da distância absoluta ($\|h\|$) entre dois pontos: $\rho(S(x), S(x+h); \phi) = \rho(\|h\|; \phi)$.

ISBN 978-85-98026-63-3

A partir desta amostra precisamos definir um modelo de covariância espacial e, com esse modelo definido, realizar a inferência e a predição espacial para todos os outros pontos da área de interesse (krigagem). A estimação dos parâmetros do modelo pode ser feita via Método da Máxima Verossimilhança, Mínimos Quadrados Ordinários, Mínimos Quadrados Ponderados, ou ainda utilizando-se a abordagem Bayesiana.

Podemos, então, usar as estimativas dos parâmetros e dos valores preditos via krigagem para calcular funções utilidades. Elas serão calculadas para cada ponto de uma malha de candidatos, e então serão maximizadas no processo de planejamento amostral.

A primeira função utilidade disponível no conjunto de rotinas desenvolvido é a de diminuição da variância preditiva amostral. Em Geoestatística, por muitas vezes o foco do pesquisador é em relação a previsões de valores para toda a área de interesse, e ele pode querer adicionar pontos à sua amostra que possibilitem previsões mais precisas. Em outras palavras, o pesquisador deseja maximizar:

$$u(x,\theta,y_d) = \int V(S(x)|\theta,\boldsymbol{y}) - V(S(x)|\theta,\boldsymbol{y},y_d)\, dx$$

A segunda função utilidade disponível calcula a probabilidade de que seja observado um valor extremo na nova localização de interesse. Neste caso, o pesquisador deseja maximizar

$$u(x,\theta,y_d) = (\, P[\, S(x) > v_{max} |\, \theta\,]\,)^2$$

onde v_{max} representa um quantil elevado da distribuição ou algum outro parâmetro definido pelo pesquisador.

Também é possível trabalhar com funções utilidade que combinam estes dois critérios, caso seja de interesse do usuário.

Estudos de Simulação

1. Aumentando uma amostra em 10 pontos

A fim de exemplificar o potencial do conjunto de rotinas desenvolvidas neste trabalho, apresentamos aqui um estudo simulado onde se deseja aumentar uma amostra em 10 pontos, visando diminuir a sua variância preditiva média — ou seja, assumindo que o pesquisador deseja aumentar o tamanho da amostra a fim de realizar previsões com maior precisão.

ISBN 978-85-98026-63-3

A amostra inicial, de 15 pontos, é proveniente de um Processo Gaussiano simulado com função de correlação exponencial de parâmetros $\sigma^2 = 20$, $\phi = 5$ e $\tau^2 = 1$. A região de interesse D é representada pelo quadrado unitário e a estimação dos parâmetros do modelo foi realizada pelo método de Máxima Verossimilhança.

Neste caso, a rotina avaliou 400 pontos candidatos e rodou em um tempo médio de 30 seg (em 10 realizações), usando paralelamente dois núcleos i5-5200U 2.20GHz.

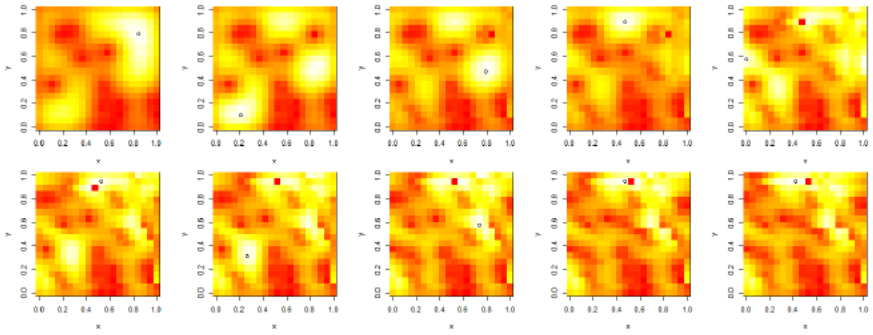


Figura 1: Evolução da função utilidade em 10 iterações do algoritmo.

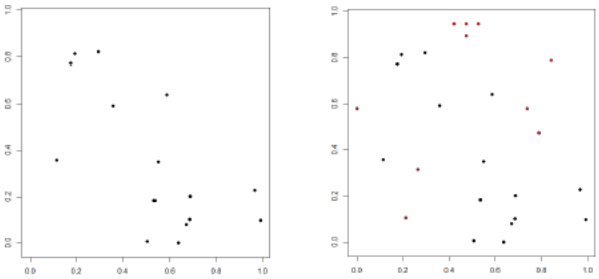
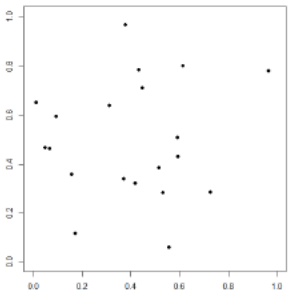


Figura 2: Amostra inicial (15 pontos) e final (25 pontos).

ISBN 978-85-98026-63-3

2. *Diferentes funções utilidades*

Esta simulação interativa tem como objetivo comparar a diferença entre as três funções utilidades: a de diminuição de variância preditiva, a de valores extremos e uma média entre as duas (com pesos iguais).



3. *Avaliando a saída de imagens e o tempo de processamento*

Nesta simulação, tinha-se como objetivo avaliar a qualidade da imagem relativa à superfície da função utilidade, que fica mais nítida conforme se aumenta o número de pontos na malha de candidatos (N), assim como o tempo necessário para que fosse criada cada imagem. A função utilidade escolhida foi a de diminuição da variância preditiva.

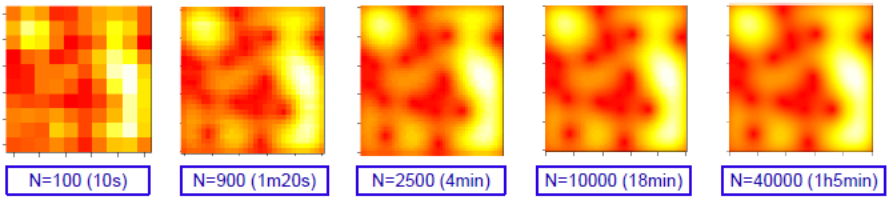


Figura 3: Imagens e tempos de processamento referentes a diferentes valores de N.

Considerações Finais

Como resultado deste trabalho, foi possível criar rotinas para o aumento sequencial de uma dada malha amostral podendo-se combinar duas diferentes funções utilidade. Além de fornecer uma solução de qualidade em um tempo factível, a rotina também fornece ao usuário visualizações gráficas dos resultados obtidos, facilitando a produção de relatórios e a realização de análises subsequentes.

ISBN 978-85-98026-63-3

Visualização de respostas dos gestores do setor público e privado sobre os atrasos em obras públicas usando o pacote sjplot do software R.

Alessandra Simão (Pós-Graduação Eng.Civil/UFF, Senac RJ) / e-mail: alessandra_simao@id.uff.br
Luciane Ferreira Alcoforado (Pós-Graduação Eng.Civil/UFF) / e-mail: lucianealcoforado@gmail.com
Orlando Celso Longo (Pós-Graduação Eng.Civil/UFF) / e-mail: orlandolongo@gmail.com

Introdução

As obras públicas apresentam, por diversas vezes, problemas na sua execução, seja por atraso do cronograma de execução, ou por serviços de má qualidade (ORANGI, PALANEESWARAN e WILSON, 2011).

Objetivo

Identificar e analisar a percepção dos gestores do setor público e privado da construção civil sobre o atraso das obras públicas na Região do Médio Paraíba.

Método

A pesquisa foi desenvolvida com a aplicação do questionário de Doloi, Sawhney, Iyer e Rentala (2012) nas Secretarias de obras dos municípios e de empresas construtoras da Região pesquisada. O mesmo foi elaborado, incorporando 45 fatores de atraso na construção descritos na literatura e agrupados em 6 categorias relacionadas: ao projeto, ao local, ao processo, aos recursos humanos, a autoridade e as questões técnicas. Utilizou-se a escala *Likert* com a pontuação entre 1 (menos importante) a 5 (muito importante), para verificar o grau de importância atribuída pelos entrevistados para os fatores que afetam o atraso nas obras. Após a coleta dos dados, estes foram tabulados em uma planilha Microsoft Excel®, sendo depois analisados pelo *software* estatístico R pacote sjPlot VERSÃO 3.2.1.

Script

```
require(sjPlot)
sjp.likert(mydf,
           cat.neutral = 3, # "3" is the new value for "neither"
           intercept.line.color = "White", # vertical middle line color
           axisLabels.y = items,
           legendLabels = labels,
           value.labels = "sum.outside",
           expand.grid = FALSE, # no inner margins in plot
           includeN = FALSE,    # hide N's in axis labels
           gridRange = 1.2,
           sort.frq = "neg.asc",
           geom.colors = "PRGn",         # purple to green
           title="Fatores relacionados ao processo")
```

Resultados

Elaborou-se os gráficos de comparação para os fatores de atraso em obras mais importantes e menos importantes, com o objetivo de identificar e analisá-los. As barras horizontais são divididas em três partes, neutra (cor cinza), positiva (cor verde) e negativa (cor lilás). Cada parte representa o percentual de respondentes para o fator em questão. Apresenta-se a seguir a análise das 6 categorias, procurando destacar os fatores identificados como mais e menos importantes nesta pesquisa.

ISBN 978-85-98026-63-3

Para os fatores relacionados ao projeto, conforme o Gráfico 1, pode-se verificar que os respondentes atribuem maior importância para o fator: *R4. Retrabalho devido à mudança de desenho ou desacordo de ordem* (53,8%), enquanto que *R1. Aumento do escopo do trabalho* (50%) é considerado como fator que afeta muito pouco o andamento de uma obra.

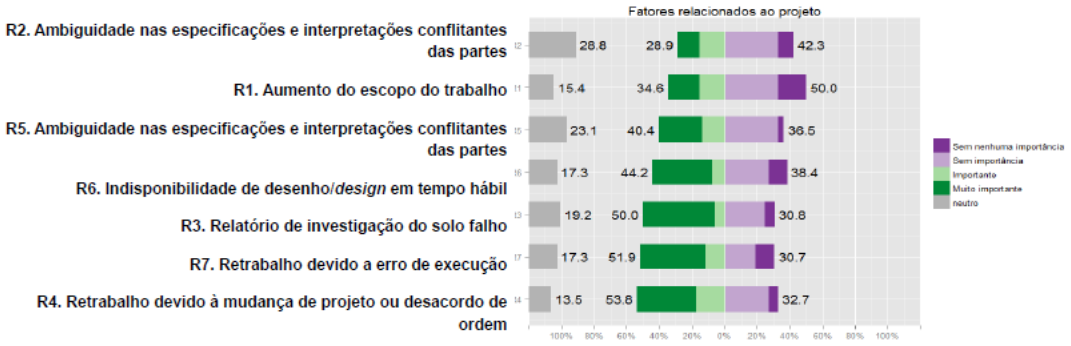


Gráfico 1 – Fatores relacionados ao projeto

Já de acordo com o Gráfico 2, no que se refere aos fatores relacionados ao local, o fator mais importante é *o R15. Condições políticas hostis* (59,6%), por outro lado, o fator considerado com pouca importância pelos respondentes é *o R12. Acidentes no local devido à negligencia* (41,2%).

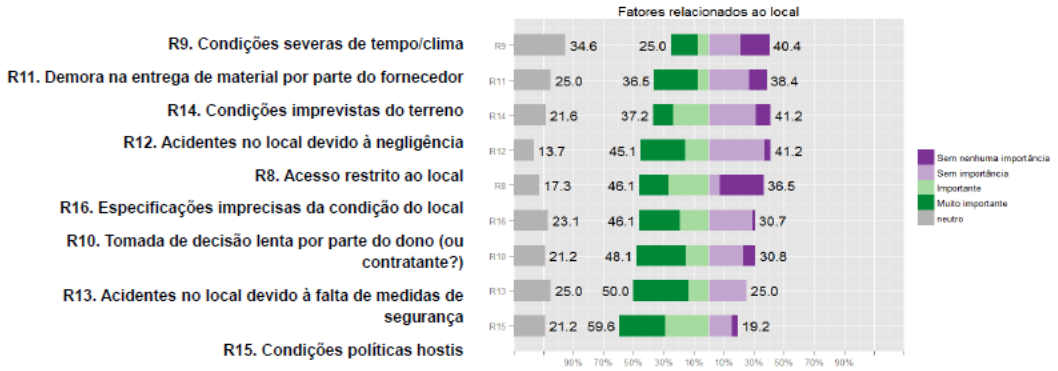


Gráfico 2 – Fatores relacionados ao local

Quanto aos fatores relacionados ao processo, segundo o Gráfico 3, o fator *R19. Demora na aquisição do material por parte da empreiteira* (59,6%) é considerado o mais importante, enquanto *R23. Demora em finalizar taxas para itens extras* (48,1%) é considerado com pouca importância.

ISBN 978-85-98026-63-3

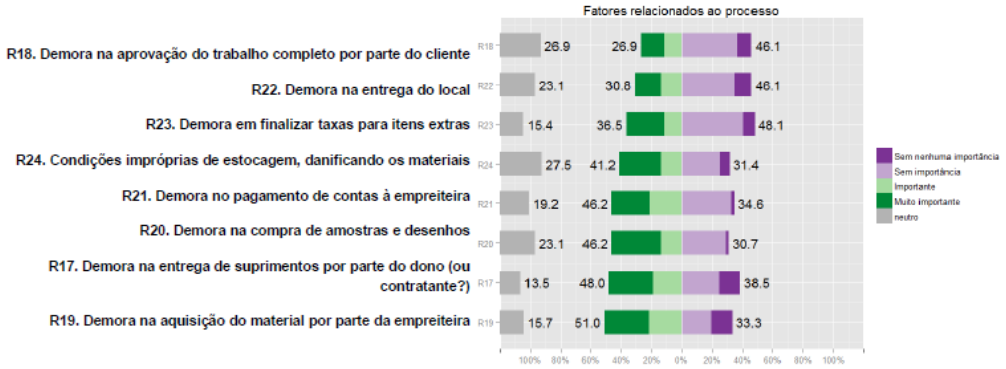


Gráfico 3 – Fatores relacionados ao processo

Para os Fatores relacionados aos recursos humanos, os entrevistados atribuíram maior importância para o fator: *R26. Má gestão/supervisão do local* (53,8%), e o fator *R27. Conflito entre donos e outras partes* (55,8%) é considerado como fator de pouca importância, conforme o Gráfico 4.

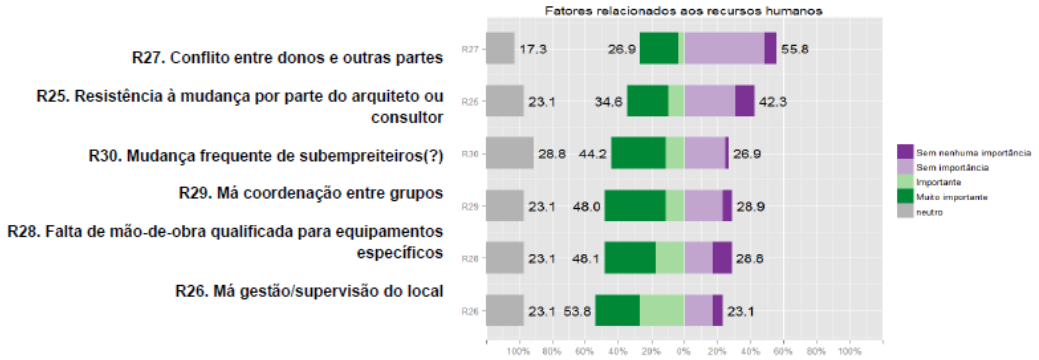


Gráfico 4 – Fatores relacionados aos recursos humanos

Já para os fatores relacionados as autoridades, os entrevistados consideram que o atributo que mais afeta o andamento das obras, e dessa forma os mais importantes: *R31. Obter permissão das autoridades locais* (53,8%), já o fator considerado sem importância, observa-se: *R35. Falta de controle em relação à subempreiteira* (43,1%).

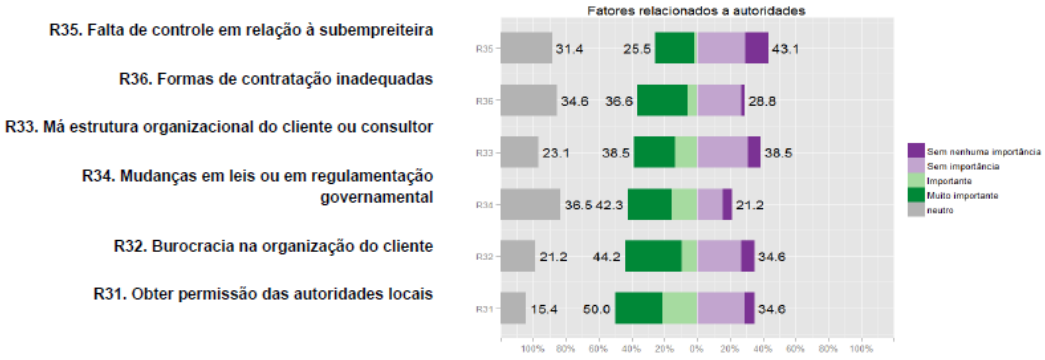


Gráfico 5 – Fatores relacionados a autoridades

ISBN 978-85-98026-63-3

Conforme o Gráfico 6, o fator *R40. Má produtividade do trabalho* é o mais relevante (61,5%) enquanto *R42. Mudanças nos preços dos materiais ou no levantamento de preços* (46%) é considerado o fator menos importante no andamento de obras.

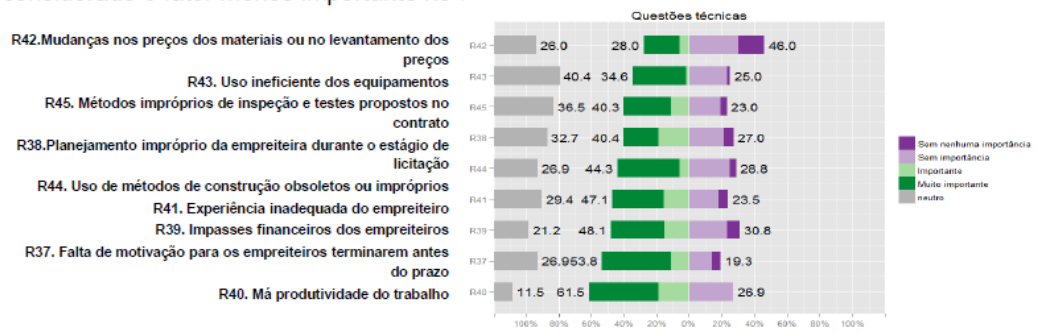


Gráfico 6 – Fatores Relacionados a questões técnicas

Considerações Finais

O objetivo geral deste trabalho foi identificar e analisar a percepção dos gestores dos setores público e privado da construção civil sobre o atraso das obras públicas na Região do Médio Paraíba e como principais fatores de atrasos apurados por meio desta pesquisa podem-se destacar: *R40. Má produtividade do trabalho (61,5%)*, *R15. Condições políticas hostis (59,6%)*, *R4. Retrabalho devido à mudança de projeto ou desacordo de ordem (53,8%)*, *R26. Má gestão/supervisão do local (53,8%)*, *R37. Falta de motivação para os empreiteiros terminarem antes do prazo* (53,8%).

Pode-se verificar de forma geral que os fatores de atraso mais relevantes para o público pesquisado estão relacionados às questões internas e de organização do canteiro de obras, do que a questões externas, o que reforça a importância do investimento em treinamento e qualificação de funcionários, processos, planejamento e controle de obras.

Principais Referências

DOLOI, Hemanta; SAWHNEY, Anil; IYER, K. C.; RENTALA, Sameer. *Analysing factors affecting delays in Indian construction projects*. International Journal of Project Management, vol. 30, issue 4, May 2012, Pages 479–489

LÜDECKE, D. (2015). sjPlot: Data Visualization for Statistics in Social Science_. R package version 1.8.4, <URL: http://CRAN.R-project.org/package=sjPlot>

ORANGI, A.; PALANEESWARAN, E.; WILSON, J. *Exploring Delays in Victoria-Based Australian Pipeline Projects*. Procedia Engineering 14 (2011) 874–881

ISBN 978-85-98026-63-3

Downside and Upside Risk Spillovers between Exchange Rates and Stock Prices

Andrea Ugolini (University of Florence) / e-mail: andreaugolini@me.com

Abstract

We examine downside and upside risk spillovers from exchange rates to stock prices and vice versa for a set of emerging economies. We characterized the dependence structure between currency and stock returns using copulas and computed the downside and upside value-at-risk and conditional value-at-risk. We document a positive relationship in emerging economies between stock prices and currency values with respect to the US dollar and the euro, with downside and upside spillover risk effects transmitted both ways. Finally, we also documented asymmetries in upside and downside risk spillovers and asymmetric differences in the size of risk spillovers when the domestic currency values against the US dollar and the euro. Our results, consistent with flight-to-quality phenomena, have implications for downside and upside risk management of international investor portfolios in emerging markets.

Introduction

Exchange rates and stock prices are two crucial macro finance variables that are intrinsically linked. Exchange rate movements have effects on stock prices given that an appreciation (depreciation) in a domestic currency reduces (increases) the international competitiveness of domestic firms and their cash flows, thereby reducing (increasing) domestic stock prices. Similarly, stock price changes impact exchange rates, since an increase in domestic stock prices triggers currency adjustment to accommodate variations in demand and supply for domestic and foreign assets included in internationally diversified portfolios.

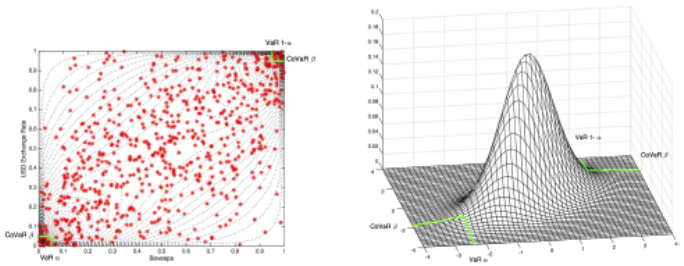
Main Objectives

1. We studied co-movement between stock and exchange rate markets using static and dynamic copula functions, which enable us to assess both average movements across marginals and joint extreme upward and downward movements.On the basis of copula information, we then evaluated the impact of downside and upside risk spillover from one market to the other by computing the downside and upside conditional value-at-risk (CoVaR) (Adrian and Brunnermeier, 2011; Girardi and Ergn, 2013) in the stock and exchange rate markets.
2. We studied downside and upside spillover effects between stock and exchange rate returns by examining dependence for a broad set of currency-equity pairs for emerging economies ? given that these financial markets are sensitive to speculative attacks, to changes in policies with the aim of managing exchange rates and to capital inflows and outflows as a result of currency rate and economic development uncertainties. As trading and capital flows in these economies are mainly denominated in dollars (USD) and euros (EUR).

ISBN 978-85-98026-63-3

Methodology

We quantify downside and upside risk using downside and upside VaR for currency and stock returns, given that VaR quantifies the maximum loss that an investor may experience in a specific time horizon and confidence level by holding a long position (downside risk) or a short position (upside risk). Hence, both risk measures are relevant for safety-first investors who want to minimize the chances of extreme losses that may drive them out of business. They are also essential in terms of pricing, as investors should be compensated for assuming potential extreme market losses (see Poon et al., 2004). To analyse the spillover risk from stock prices to exchange rates and vice versa, we considered the impact of financial distress in one market (measured by its VaR) on the VaR of another market. Spillover risk is closely related to the propagation of failures from one market to another, as confirmed in the systemic risk literature (see e.g., Billio et al., 2012; Bisias et al, 2012). To quantify downside or upside spillover risk, we employed the CoVaR measure as proposed by Adrian and Brunnermeier (2011) and generalized by Girardi and Ergn (2013). Namely, the CoVaR for the stock market is the VaR for stock market returns conditional on the fact that exchange rates experience an extreme movement. Let r_t^s be the returns for stocks and r_t^e be the returns for exchange rates. The (downside) CoVaR for stock returns and confidence level $1-\beta$ can be formally defined as the β-quantile of the conditional distribution of r_t^s as:



(a) Bivariate t-Copula . (b) Bivariate PDF t-student

$$Pr(r_t^s \leqslant CoVaR_{\beta,t}^s | r_t^e \leqslant VaR_{\alpha,t}^e) = \beta \tag{1}$$

where $VaR_{\alpha,t}^e$ is the α-quantile of the exchange rate return distribution: $Pr(r_t^e \leqslant VaR_{\alpha,t}^e) = \alpha$ measures the maximum loss that exchange rate returns may experience for a confidence level $1-\alpha$ and a specific time horizon. Similarly, we can measure (upside) CoVaR for a given extreme upward movement in exchange rate returns as:

$$Pr(r_t^s \geqslant CoVaR_{\beta,t}^s | r_t^e \geqslant VaR_{1-\alpha,t}^e) = \beta \tag{2}$$

where $VaR_{1-\alpha,t}^e$ now measures the maximum loss by considering a short position for a confidence level $1-\alpha$ and for a specific time horizon. On the other hand, we can measure the systemic impact of stock prices on exchange rates by considering the CoVaR for the exchange rate market instead of for the stock market, as in eqs. (1) and (2). CoVaR in eqs. (1) and (2) can be represented in terms of copulas, since the conditional probabilities can be re-written, respectively, as:

$$C(F_{r_t^s}(CoVaR_{\beta,t}^s), F_{r_t^e}(VaR_{\alpha,t}^e)) = \alpha\beta \tag{3}$$

$$1 - F_{r_t^s}(CoVaR_{\beta,t}^s) - F_{r_t^e}(VaR_{\alpha,t}^e) + C(F_{r_t^s}(CoVaR_{\beta,t}^s), F_{r_t^e}(VaR_{\alpha,t}^e)) = \alpha\beta \tag{4}$$

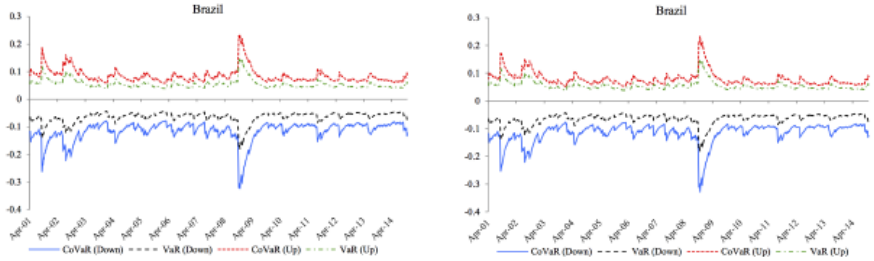
where $F_{r_t^s}$ and $F_{r_t^e}$ are the joint and marginal distributions of the stock and exchange rate returns, respectively. We can thus compute the CoVaR following a two-step procedure (see Reboredo and Ugolini, 2015): first, given the significance levels for the VaR and CoVaR (α and β, respectively) and for specific forms of the copula function we can solve (3) or (4) in order to obtain the value of $F_{r_t^s}(CoVaR_{\beta,t}^s)$; then, in a second step, using the distribution function for stock market returns as given by the marginal model, we can compute the CoVaR as $F_{r_t^s}^{-1}(F_{r_t^s}(CoVaR_{\beta,t}^s))$

ISBN 978-85-98026-63-3

Results

Exchange rate spillover effects to stock returns

Using the best copula fit and following the two-step procedures described above, for each time period we obtained the CoVaR value for stock returns at the 95% confidence level ($\beta = 0.05$) conditional on the VaR value for the exchange rate returns at the 95% confidence level ($\alpha = 0.05$). Considering spillover effects from exchange rates to stock returns, Figure 1 shows the temporal dynamics of results for the downside and upside VaR and CoVaR values for stock returns.



(c) USD exchange rate against the local currency. (d) EUR exchange rate against the local currency.

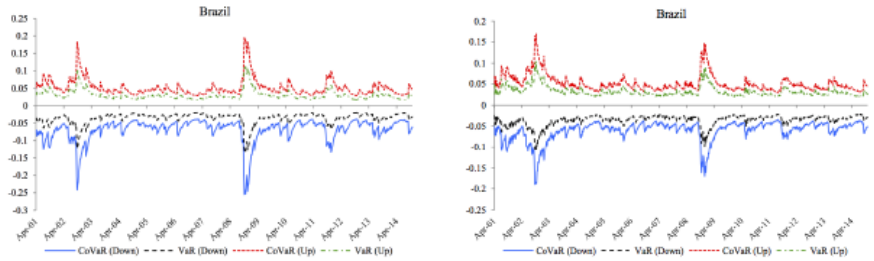
Figure 1: Upside and downside value-at-risk (VaR) and conditional value-at-risk (CoVaR) for stock market returns.

USD exchange rate				EUR exchange rate			
Downside		Upside		Downside		Upside	
VaR	CoVaR	VaR	CoVaR	VaR	CoVaR	VaR	CoVaR
-0.065	-0.116	0.055	0.085	-0.065	-0.116	0.055	0.079
(0.02)	(0.04	(0.02)	(0.03)	(0.02)	(0.04)	(0.02)	(0.03)

Table 1: Brazil Upside and downside value-at-risk (VaR) and conditional value-at-risk (CoVaR) for exchange rate returns

Stock return spillover effects to exchange rates

Figure 2 depicts the temporal dynamics of our results for the downside and upside VaR and CoVaR values for exchange rates by considering spillover effects from stock prices to exchange rates.



(a) USD exchange rate against the local currency. (b) EUR exchange rate against the local currency.

Figure 2: Upside and downside value-at-risk (VaR) and conditional value-at-risk (CoVaR) for exchange rate returns.

ISBN 978-85-98026-63-3

USD exchange rate				EUR exchange rate			
Downside		Upside		Downside		Upside	
VaR	CoVaR	VaR	CoVaR	VaR	CoVaR	VaR	CoVaR
-0.036	-0.069	0.031	0.052	-0.037	-0.062	0.032	0.052
(0.02)	(0.03)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)

Table 2: Brazil Upside and downside value-at-risk (VaR) and conditional value-at-risk (CoVaR) for exchange rate returns

Asymmetric spillover effects

Although the impact of extreme movements on stock prices on exchange rates and vice versa are symmetric at the theoretical level, the reaction of real and financial flows to stock and currency prices could not be symmetric due to different easons, e.g., the flight-to-quality effect. Hence, upside and downside risk spillover effects may be asymmetric and different across currencies. Asymmetric spillovers have crucial implications for hedging decisions in international investor portfolios.

Currency to Stock		Stock to Currency		Currency to Stock		Stock to Currency	
USD	EUR	USD	EUR	Downside	Upside	Downside	Upside
$H_0 : \frac{CoVaR}{VaR}(Down) = \frac{CoVaR}{VaR}(Up)$				$H_0 : \frac{CoVaR}{VaR}(\$) = \frac{CoVaR}{VaR}(e)$			
$H_1 : \frac{CoVaR}{VaR}(Down) > \frac{CoVaR}{VaR}(Up)$				$H_1 : \frac{CoVaR}{VaR}(\$) > \frac{CoVaR}{VaR}(e)$			
1.000	1.000	1.000	0.422	0.192	0.958	0.994	0.825
[0.000]	[0.000]	[0.000]	[0.000]	[0.167]	[0.000]	[0.000]	[0.000]

Table 3: Asymmetric spillover effect

Conclusions

- We found evidence of a positive relationship between stock prices and currency values with respect to the USD and the EUR; thus, the home currencies appreciated (depreciated) when stock prices moved up (down).
- We found evidence of downside and upside spillover risk effects from currencies to stock returns and from stock returns to currency returns; this is consistent with the fact that bullish (bearish) stock markets attract capital inflows as demand for local assets by foreign investors increases (decreases), thus increasing (reducing) the value of the domestic currency.
- Our evidence is also consistent with the fact that the increase in international trade in emerging economies has strengthened financial integration in spite of capital movement restrictions.
- Our analysis reveals that downside and upside spillovers are asymmetric, with greater downside rather than upside risk spillover effects.
- We also found that spillovers from and to the USD were greater than from and to the EUR. This evidence is consistent with the fact that the USD plays a more crucial role than the EUR in trade and financial transactions in emerging economies.
- Our downside risk results are consistent with flight-to-quality, and our downside and upside risk analysis has practical implications for downside and upside risk management of international investor portfolios for emerging markets.

ISBN 978-85-98026-63-3

Modelagem do tipo de violência contra o idoso cometida por pessoas desconhecidas usando uma pesquisa amostral complexa

Fernando de Oliveira Alencar Júnior (UFF) / e-mail: fernandoalencar@id.uff.br
José Rodrigo de Moraes (GET-UFF) / e-mail: jrodrigo78@est.uff.br

1. Introdução

O rápido envelhecimento da população brasileira pode representar um grave problema para a sociedade, caso não sejam formuladas e implementadas políticas e ações preventivas específicas para os idosos, visando promover melhorias na sua qualidade de vida, assegurando maiores oportunidades de saúde, maior participação social e segurança. Em 2003, a fim de assegurar os direitos das pessoas com idade igual ou superior a 60 anos, foi sancionada a Lei 10.741, que dispõe sobre o Estatuto do Idoso. Para os efeitos dessa lei, a violência contra o idoso é definida como qualquer ação ou omissão praticada em local público ou privado que lhe cause morte, dano ou sofrimento físico ou psicológico. Com base nesta definição, verifica-se que a violência física não é o único tipo de violência praticada contra o idoso. Existem outros tipos de violência, muitas vezes ocultas na sociedade, que ocorrem tanto no âmbito privado quanto no âmbito público, que podem causar morte, dano ou sofrimento.

2. Objetivo

Este trabalho tem como objetivo estabelecer a associação entre um conjunto de características sociodemográficas, comportamentais e de saúde dos idosos (60 anos ou mais de idade) e a chance do idoso sofrer violência física por pessoas desconhecidas, relativamente a outros tipos de violência.

3. Material e Método

Neste trabalho foi ajustado o modelo de regressão logística binária através do método de Máxima Pseudo-Verossimilhança (MPV), considerando as principais características do plano amostral da Pesquisa Nacional de Saúde (PNS 2013), onde a variável resposta do modelo é uma variável binária que identifica o tipo de violência (física, não física) contra o idoso cometida por pessoa desconhecida. Com relação as variáveis explicativas foram consideradas: *sexo, faixa etária, cor/raça, região de residência, área de localização de domicílio, nível educacional, estado civil, qualidade da construção da moradia, autoavaliação de saúde geral, frequência da visita da equipe da saúde da família (ESF), alguma deficiência, consumo de bebida alcoólica e prática de exercício físico nos últimos 3 meses.*

Para o ajuste do modelo utilizou-se a biblioteca *survey do software R, versão 3.2.4.* Com relação à estratégia de modelagem, considerou-se na análise multivariada somente as variáveis que na análise bruta apresentaram pelo menos uma categoria com efeito significativo ao nível de significância de 20%. Na análise multivariada, foram excluídas, uma a uma, na ordem decrescente do p-valor, as variáveis que não apresentaram associação significativa com a prevalência de sucesso (sofrer violência física) considerando o nível de significância de 5%. Este procedimento foi repetido até que se obtivesse um modelo onde todas as variáveis explicativas tivessem pelo menos uma categoria com significância estatística ao nível de 5%. O modelo selecionado foi composto por seis variáveis explicativas, e pode ser representado da seguinte forma:

ISBN 978-85-98026-63-3

$$\ln\left(\frac{p_{abcdef}}{1-p_{abcdef}}\right) = \mu + \alpha_a + \beta_b + \gamma_c + \delta_d + \phi_e + \epsilon_f$$

onde,

$\mu \rightarrow$ intercepto do modelo.

$\alpha_a \rightarrow$ efeito principal do a-ésimo nível da variável região de residência; a=1,..., 5.

$\beta_b \rightarrow$ efeito principal do b-ésimo nível da variável nível educacional; b=1, 2, 3.

$\gamma_c \rightarrow$ efeito principal do c-ésimo nível da variável qualidade da construção da moradia; c=1, 2.

$\delta_d \rightarrow$ efeito principal do d-ésimo nível da variável prática de exercício físico nos últimos 3 meses; d=1, 2.

$\phi_e \rightarrow$ efeito principal do e-ésimo nível da variável autoavaliação de saúde geral; e=1, 2, 3.

$\epsilon_f \rightarrow$ efeito principal do f-ésimo nível da variável frequência da visita da ESF; f=1, ..., 4.

4. Resultados e Conclusões

Pode se observar na tabela 1 maior chance de sofrer violência física entre idosos: sem instrução, residentes em domicílios com construção inadequada, não praticantes de exercício físico e que reportaram saúde regular. Também se observou maior chance de violência física entre idosos residentes na região Sudeste (OR=3,080; p-valor=0,021) e menor chance entre idosos da região Centro-Oeste (OR=0,351; p-valor=0,012), comparativamente aos idosos da região Norte. Com relação a frequência da visita da equipe da saúde da família, verificou-se que idosos residentes em domicílios cadastrados na estratégia da saúde da família, mas que nunca receberam visita da equipe, têm uma chance de sofrer violência física 2,4 vezes maior que os idosos residentes em domicílios não cadastrados (OR=2,430; p-valor=0,012).

ISBN 978-85-98026-63-3

Tabela 1: Ajuste do modelo de regressão logística binária explicativo da ocorrência de violência física cometida por pessoa desconhecida contra idosos violentados, segundo os dados da PNS 2013.

Variáveis	Modelo selecionado	
	Razão de chance (OR)	P-valor
Região de residência		
Norte	1	-
Nordeste	1,640	0,257
Sudeste	3,080	0,021
Sul	2,006	0,213
Centro-Oeste	0,351	0,012
Nível educacional		
Sem instrução	1	-
Sem nível superior	0,116	<0,001
Com nível superior	0,066	<0,001
Qualidade da construção da moradia		
Inadequado	1	-
Adequado	0,135	<0,001
Prática de exercício físico nos últimos 3 meses		
Não	1	-
Sim	0,344	0,005
Autoavaliação de saúde geral		
Boa	1	-
Regular	1,973	0,035
Ruim	0,253	0,001
Frequência da visita da ESF		
Não cadastrado	1	-
Não sabe	0,287	0,004
Cadastrado (nunca recebeu visita)	2,430	0,012
Cadastrado (pelo menos 1 visita)	0,530	0,136

ISBN 978-85-98026-63-3

Como ainda são escassos estudos sobre violência contra idosos no Brasil, sobretudo no contexto de violência fora do âmbito familiar, este estudo buscou contribuir para aumentar o conhecimento sobre o tema, ressaltar a necessidade de formulação e aplicação de políticas específicas para a população idosa com a participação do Estado, sociedade e família para enfrentamento do problema da violência contra os idosos.

Para combater a violência contra idosos, destaca-se a importância do papel dos serviços de saúde na identificação dos casos de violência, a necessidade de maior conscientização e educação da população em geral para este problema, e a criação de leis mais severas para punir os agressores e condenar os atos violentos contra idosos nas suas diferentes formas.

5. Referências

BRASIL, Lei n. 10.741 de 1º de outubro de 2003. Dispõe sobre o Estatuto do Idoso e dá outras providências. Disponível em http://www.planalto.gov.br/.

INTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Pesquisa Nacional de Saúde 2013: Percepção do estado de saúde, estilos de vida e doenças crônicas – Brasil, Grandes Regiões e Unidade da Federação, 2014.

SALES, D. S.; FREITAS, C. A.; BRITO, M.C.; OLIVEIRA, E.; DIAS, F.; PARENTE, F.; SILVA, M. J. *A violência contra o idoso na visão do agente comunitário de saúde*. Estud. interdiscip. envelhec;19(1):63-77, 2014.

ISBN 978-85-98026-63-3

Uma análise dos dados a partir da pesquisa realizada sobre segurança no entorno do CEFET/RJ.

Caroline Ponce de Moraes (CEFET)/ poncecefet@gmail.com
Bianca Sampaio Monteiro (CEFET)/ biancasampaiomonteiro@gmail.com
Christiane Webster Carneiro (CEFET)/ webster.christiane@gmail.com
Mariana Sento Sé Costa (CEFET)/ marisentose@gmail.com

Introdução

A segurança pública é sempre um tema de muitas discussões no cotidiano da população do Rio de Janeiro. Pensando nisso, foi proposto para a turma de estatística aplicada de 2016.1 do CEFET/RJ a elaboração de uma pesquisa para que fosse possível analisar o quanto os frequentadores da instituição se sentem seguros em seu entorno. Após a aplicação do questionário, foi utilizado o software estatístico R para a análise dos dados coletados, tornando viável a transformação da opinião geral de estudantes, professores e servidores em números e gráficos.

Objetivo

Este trabalho foi desenvolvido com o objetivo de divulgar, através da análise de dados, a opinião dos estudantes e funcionários do Campus Maracanã acerca da segurança pública oferecida no entorno do CEFET/RJ. A partir desta análise, espera-se fazer uma reflexão sobre esta situação.

Metodologia

A construção do questionário foi realizada no Google Forms e a sua divulgação foi feita pelo Facebook e por outros meios de comunicação como, por exemplo, o Whatsapp. Assim, após a aplicação e coleta das informações, foi utilizado o software RStudio juntamente com sua linguagem de programação R, o que nos permitiu explorar com rapidez e eficiência uma variedade de análises estatísticas.

Análise dos Resultados

Uma das perguntas do questionário diz respeito a qualificação do entrevistado, ou seja, se ele é aluno, professor ou funcionário do CEFET/RJ.

Figura 1: Qualificação dos entrevitados: estudande, funcionário ou professor.

ISBN 978-85-98026-63-3

Como visto no gráfico anterior, a maior parte dos entrevistados é estudante, por essa razão a faixa etária predominante foi entre 18 e 24 anos (mais de 250, dos 356 entrevistados) conforme visto na Figura 2.

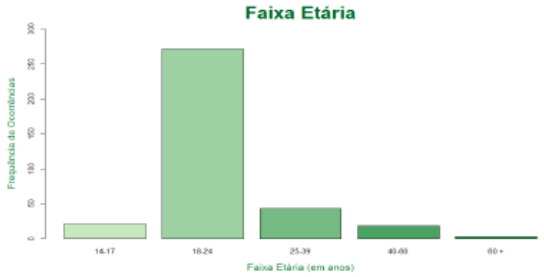


Figura 2: Ilustração da faixa etária dos respondentes.

Essa discrepância numérica na quantidade de estudantes entrevistados deve-se ao fato da amostra ter que ser representativa para condizer com a realidade da população. Como há no CEFET/RJ muito mais alunos do que funcionários e professores, é indicado que mais alunos respondam ao questionário para que a amostra apresente dados condizentes com a realidade.

Na Figura 3, observa-se que a proporção do sexo masculino teve uma representatividade maior na nossa pesquisa.

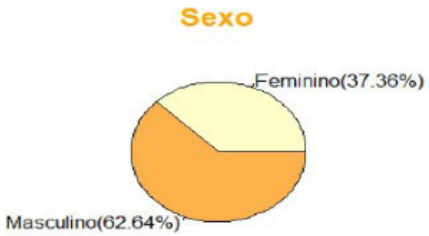


Figura 3: Sexo dos entrevistados.

Em seguida, na Figura 4, podemos perceber que tanto na hora da entrada como na da saída os respondentes optam pela Rua General Canabarro.

Figura 4: Portaria de entrada x Portaria de saída.

ISBN 978-85-98026-63-3

Para tentar compreender a análise anterior, foi questionado aos respondentes qual era o fator determinante para sua escolha.

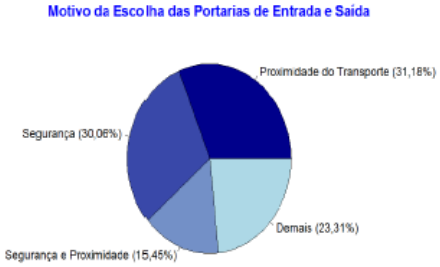


Figura 5: Motivo da escolha das portarias de entrada e saída do CEFET/RJ.

Assim, de acordo com as Figuras 4 e 5, observa-se que há uma grande relação entre as escolhas das portarias com a sensação de segurança e com a proximidade do transporte, pois foram esses os motivos com o maior percentual.

A partir da pesquisa, foi possível perceber que os respondentes, em sua maioria, acham a segurança no entorno do CEFET/RJ insuficiente.

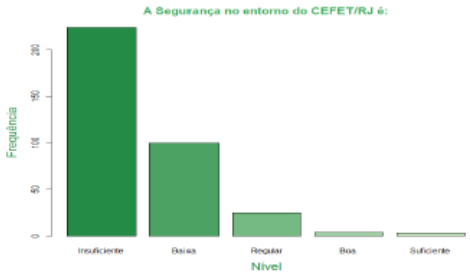


Figura 6: Ilustração dos respondentes quanto a segurança no entorno do CEFET/RJ.

Acredita-se que esse sentimento de insegurança é decorrente do alto nível de relatos de assaltos. Dos 356 entrevistados, mais de 90% conhecem alguém que já foi assaltado.

Análises Cruzadas

A análise abaixo relaciona o sexo dos entrevistados com o quanto se sentem seguros. Verifica-se que 85,71% das mulheres se sentem pouco ou pouquíssimo seguras, enquanto esse número se reduz para 58,75% para os homens entrevistados.

ISBN 978-85-98026-63-3

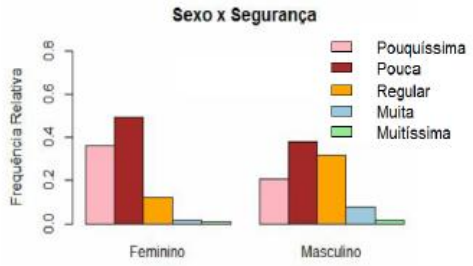


Figura 7: Relação do sexo x segurança dos entrevistados

A seguir na Figura 8, percebe-se que o nível de insegurança é maior dentre os entrevistados que conhecem alguém que já foi assaltado.

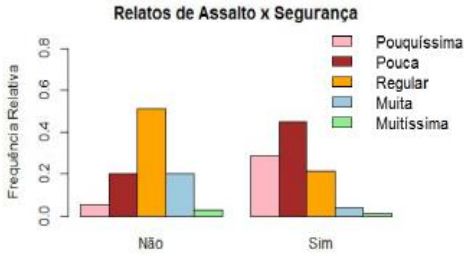


Figura 8: Relatos de assaltos x segurança

Por fim, como pode ser visto na Figura 9, analisamos as perguntas Presença de Policiamento e Turno Frequentado.

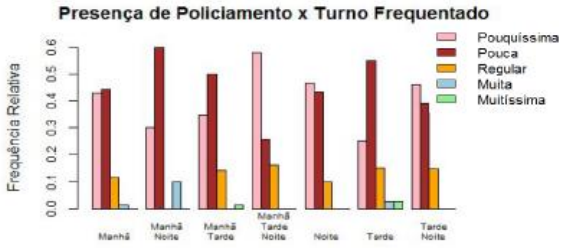


Figura 9: Relação entre a presença de policiamento e o turno frequentado

Considerações Finais

A partir da análise das respostas dos entrevistados foi possível concluir que a segurança no entorno do CEFET/RJ ainda é bastante precária. Mais de 90% dos respondentes disseram conhecer alguém que já foi assaltado nas redondezas da instituição e não se sentem seguros ao entrar e sair de lá. O público que mais demonstrou sentir essa insegurança foram as mulheres. Além disso, foi possível observar uma maior insegurança entre os respondentes que frequentam o CEFET/RJ no turno da noite. Essa insegurança é proveniente da raríssima presença, de acordo com a opinião dos entrevistados, de policiamento no entorno do CEFET/RJ. Outro questionamento levantado na pesquisa foi quanto ao acesso à instituição, visto que é possível entrar e sair por lugares distintos. A partir da análise, conclui-se que os fatores decisivos para essa escolha são: Proximidade do Transporte e Segurança. Algumas sugestões de melhorias apontadas pelos entrevistados foram: aumentar o policiamento, seguranças mais bem preparados no CEFET/RJ e em seu entorno, mais iluminação e câmeras.

ISBN 978-85-98026-63-3

Modelo semiautomático de normalização radiométrica de série temporal de imagens de satélite implementado em linguagem R

Pedro José Farias Fernandes (UFF) / e-mail: pj_fernandes@id.uff.br
Luiz Furtado (UFRJ) / e-mail: chefechefe@gmail.com
Raúl Sanchez Vicens (UFF) / e-mail: rsvicens@gmail.com

INTRODUÇÃO

Sensores que operam em baixa e média resolução espacial fornecem um conjunto de imagens sucessivas de uma mesma área, e formam uma série temporal de imagens. Essa resolução temporal permite entender as trajetórias evolutivas da superfície terrestre. Importantes fontes de dados multitemporais são a série de imagens dos satélites Landsat, com mais de 30 anos de imageamento.

Dentro dessa perspectiva, há as técnicas de detecção de mudanças para dados multitemporais de sensores remotos. Para o uso dessas técnicas, é necessário aplicar rigorosos pré-processamentos para minimizar variações radiométricas entre as imagens. Uma maneira de diminuir essas variações é a aplicação da normalização radiométrica.

A normalização radiométrica ajusta as propriedades radiométricas de uma imagem para corresponder com as de uma imagem de referência, por um modelo de regressão por mínimos quadrados, a partir de pontos pseudo invariantes (PIF), colocando as duas imagens em uma escala comum sem o uso de parâmetros extras. Portanto, o objetivo deste trabalho é desenvolver um código em linguagem R para a normalização radiométrica automática de série temporal Landsat 5 a partir de PIF coletados manualmente.

Especificamente, busca-se analisar os erros de procedimentos de normalização radiométrica a partir do uso da função em linguagem R desenvolvida, e verificar a eficácia dos modelos de regressão utilizados para cada banda

MATERIAIS E MÉTODOS

Foram utilizadas 11 imagens TM/Landsat 5, cena 217-76 (Figura 1), dos anos de 1984 até 2011 (quando possível, tentou-se adquirir imagens da mesma época para minimizar a influência da iluminação, geometria de visada e de fenologia vegetal. As imagens do sensor TM possuem resolução espacial de 30 metros, resolução temporal de 16 dias, e resolução radiométrica de 8 bits.

Foram coletados 122 PIF usando interpretação visual (Figura 1), distribuídos na imagem segundo um desenho amostral aleatoriamente estratificado para as classes: água (limpa e longe de plumas de sedimentos), antrópico (áreas urbanas, encontro de estradas e pistas de aeroportos), floresta e sombra.

ISBN 978-85-98026-63-3

Para realizar a normalização radiométrica, foi gerado um modelo de regressão ordinária por mínimos quadrados entre a imagem a ser normalizada e a imagem referência (2011), banda por banda. Assim, estimaram-se valores de reflectância a partir dos parâmetros do modelo, que corresponderiam às mesmas condições de aquisição na data de referência (ECKHARDT et al., 1990). Todas as normalizações foram feitas utilizando um *script* desenvolvido em linguagem R. Em um segundo momento, toda a série temporal foi normalizada, em que as médias de RMSE foram analisados para cada banda espectral (excetuando a banda 6 do infravermelho termal, que não foi normalizado) e para cada ano da série temporal.

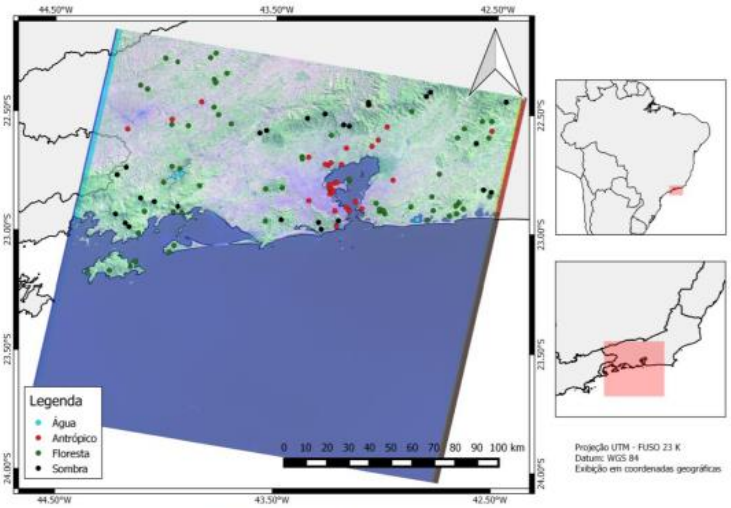


Figura 1 – Área de estudo

RESULTADOS E CONCLUSÃO

A função programada para normalização através de regressão ordinária de mínimos quadrados é chamada através do comando *norm()* no *R*. Assim, como resultado, a função *norm( )* retorna um arquivo na estrutura *layer stacking* com as bandas normalizadas; um arquivo com os gráficos de dispersão no formato BMP, equações e coeficientes de 3); e tabelas no formato XLS (uma para cada banda) com as estatísticas. Os gráficos para o ano de 1991 (Figura 2) mostram bom grau de ajuste à reta. Todos os valores dos coeficientes de determinação estão acima de 0,94.

ISBN 978-85-98026-63-3

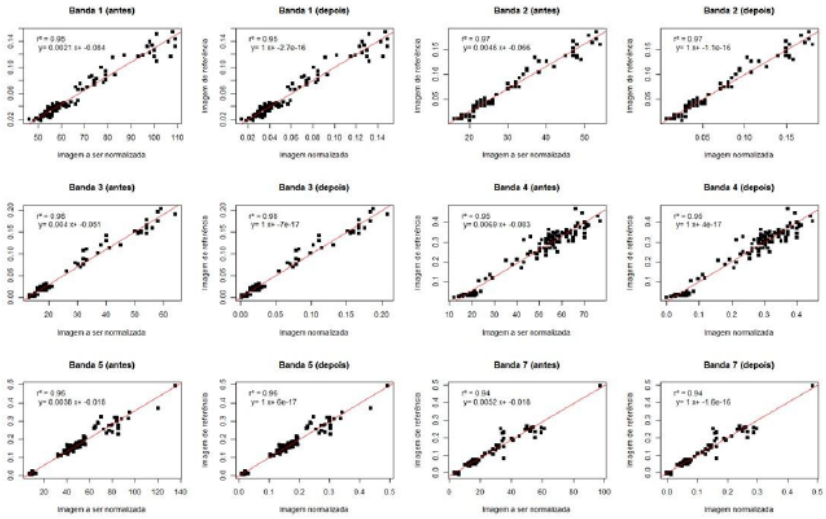


Fig. 2 - Exemplo de gráficos gerados por banda. A imagem de 1991, no caso, foi normalizada a partir da de 2011. São feitas duas regressões para a comparação gráfica: a primeira entre a imagem a ser normalizada e a referência (antes); a segunda entre a imagem normalizada e a imagem referência (depois).

Toda série temporal apresentou equações do modelo de regressão semelhantes (valores de x bem próximos de y). O menor valor de r^2 foi para a banda 1 de 1984 (r^2=0,79) e o maior valor foi para a banda 3 de 1991 (r^2=0,99). Valores menores que 0,9 foram encontrados apenas para os anos de 1984 e 1986.

As médias de RMSE por banda para todos os anos são mostradas na Figura 3, calculadas tanto para os PIF utilizados na normalização, quanto para os PIF de avaliação. Para os dois tipos de PIF, há o mesmo padrão: o erro aumenta pouco da banda 1 a 3, tem um pico na banda 4, e cai um pouco na banda 5 e banda 7, esta última com valor ligeiramente menor que a banda 5.

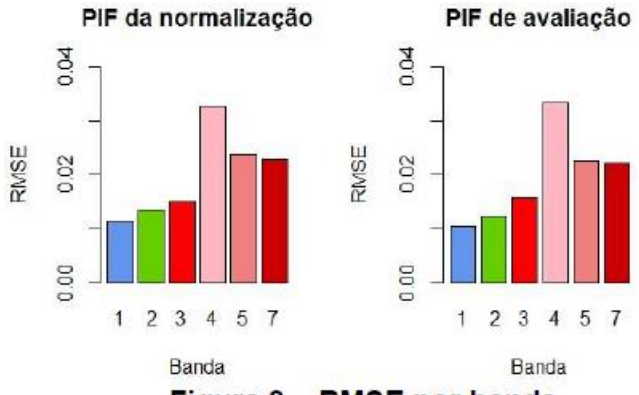


Figura 3 – RMSE por banda

ISBN 978-85-98026-63-3

Os erros calculados por data foram baixos (todos abaixo de 0,035), com os maiores erros para 1984 e 1986 (este último apresentou o maior). Também, como visto, foram os anos com menores r^2. Conclui-se que a função programada em R é eficaz para colocar a série temporal em escala de valores comum à imagem de referência.

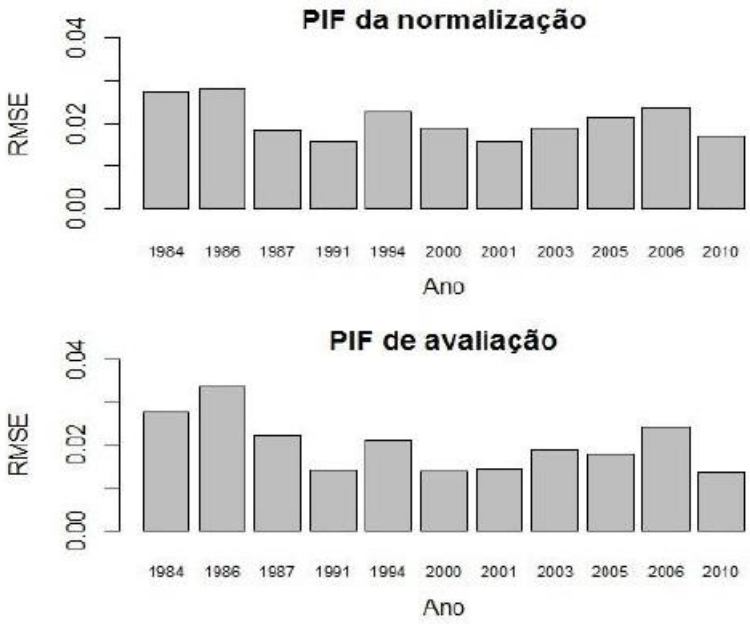


Figura 4 – RMSE por data

Anais do Seminário Internacional de Estatística com R: Inovação e Atuação do Profissional no Mercado

ISBN 978-85-98026-63-3

Métodos de Seleção de Variáveis via Verossimilhança Penalizada

Julio Cesar de Azevedo Vieira (EMAp/FGV e UFF) / e-mail: julio_vieira@globo.com
Jony Arrais Pinto Junior (UFF) / e-mail: jarrais@id.uff.br

Resumo: Realizar seleção de variáveis é uma parte importante no processo encontrar a relação entre um conjunto de covariáveis e uma variável de interesse para modelos lineares encaixados. Existem diversos métodos e critérios que realizam a seleção de variáveis em duas etapas, sendo elas, a estimação dos coeficientes e depois a seleção. Mais recentemente métodos baseados em verossimilhança penalizada se propõe a realizar simultaneamente as duas etapas da seleção. Este trabalho busca estudar a estrutura verossimilhança penalizada, entender os métodos LASSO e Adaptive LASSO e comparar as abordagens, por meio de um estudo de simulação, dos métodos usuais, AIC e BIC, e dos métodos baseados em verossimilhança penalizada, LASSO e Adaptive LASSO sob a ótica de modelos lineares. Foi possível observar que todos os métodos comparados são bons ao identificar as variáveis significativas sobre a ótica de modelos lineares, que os métodos usuais realizam essa identificação com probabilidades maiores do que os baseados em verossimilhança penalizada quando há grande variabilidade. Verificou-se também que o AIC apresenta dificuldade ao identificar as variáveis não significativas, diferentemente dos métodos baseados em verossimilhança penalizada que conseguem estimar com precisão ideal mesmo em cenários de alta variabilidade para modelos lineares e que o BIC só consegue essa estimação ideal com tamanhos de amostras grandes.

Introdução: Encontrar a relação entre um conjunto de covariáveis e uma variável de interesse pode ser trabalhoso, portanto é importante dispor de um conjunto de métodos e critérios que visem selecionar entre todas as possíveis covariáveis aquelas que realmente tenham relação com a sua variável de interesse. Os critérios de informação, como AIC e BIC, visam encontrar entre todos os possíveis modelos àquele que forneça o maior ganho de informação. Mais recentemente a literatura dispõe de métodos de estimação e seleção simultânea de variáveis, como LASSO e Adaptive LASSO.

Objetivo: Analisar as melhorias propostas na literatura para os métodos baseados em verossimilhança penalizada e, conhecidas as melhorias, realizar um estudo de simulação que visa comparar os métodos de seleção de variáveis, AIC, BIC, LASSO e Adaptive LASSO, no contexto de modelos lineares.

Métodos: Foi utilizado modelo linear normal em que a variável resposta tem distribuição normal e tem, por hipótese, os erros como variáveis aleatórias independentes e identicamente distribuídas com média 0 e variância σ^2. Para a estimação do vetor β foi utilizado o logaritmo da Máxima Verossimilhança Penalizada que pode ser escrito em função do logaritmo da máxima verossimilhança $l(\beta)$ e de uma função de penalização $p_\lambda(\beta)$.

$$lp(\beta) = l(\beta) - p_\lambda(\beta)$$

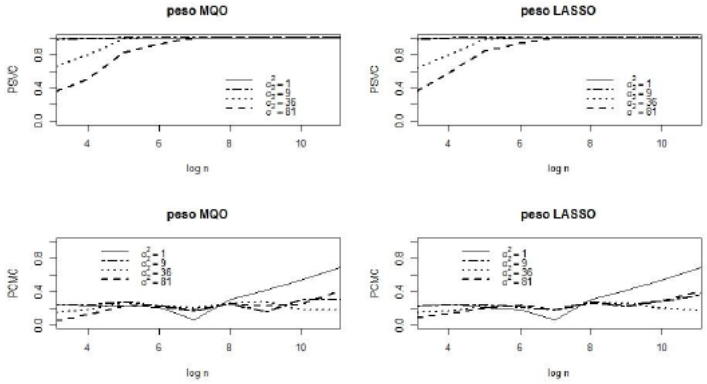
Os critérios AIC e BIC utilizam a informação da verossimilhança a fim de encontrar o modelo mais adequado sendo escolhido aquele modelo que possuir menor valor de AIC e BIC.
Dos métodos baseados verossimilhança penalizada, o LASSO possui função de penalização $p_\lambda(\beta) = \sum_{j=1}^{p} |\beta_j|$, já o Adaptive LASSO fornece pesos às covariáveis a fim de melhorar a seleção feita e tem função de penalização $p_\lambda(\beta) = \sum_{j=1}^{p} w_j |\beta_j|$ em que $\hat{w} = 1/|\beta|^\gamma$ com $\gamma > 0$.

ISBN 978-85-98026-63-3

Resultados: Foram geradas 100 repetições, cada uma com tamanhos de amostra de n = 30 até n = 50 mil, em que os erros são normalmente distribuídos, tal que a variável resposta (y_i) foi gerado a partir da combinação linear dos erros (ε_i) como a seguir:

$$y_i = x_i^T \beta + \varepsilon_i$$

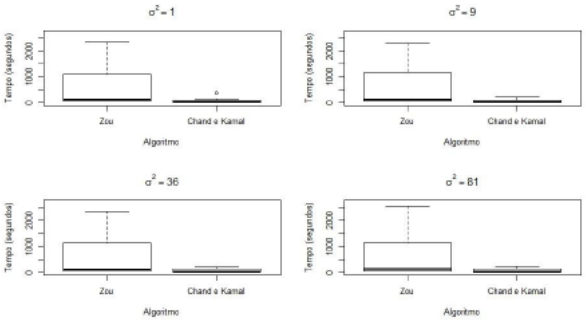
em que $\beta = (0;1,6;0,98;1,3;0)^T$ e $\varepsilon_i \sim N(0, \sigma^2)$ tal que σ^2 assuma valores 1, 9, 36, 81. O vetor de covariáveis foi gerado por meio de uma distribuição normal padrão tendo correlação entre x_i e x_j dadas por $0{,}5^{|i-j|}$. Para a estimação do parâmetro λ , é preciso definir valores iniciais de forma que os valores estimados para λsejam determinados pelos diferentes critérios que serão utilizados na estimação. Em todos os casos foi utilizado uma sequência de 100 valores de 0,01 a 4.

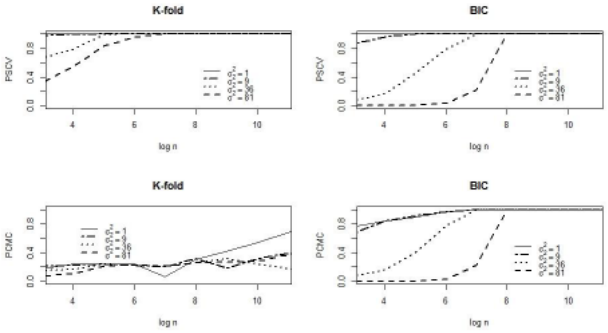


Peso: Mínimos Quadrados versus LASSO. Pela semelhança na estimação optou-se por utilizar o LASSO para se evitar Colinearidade, como proposto pela literatura.

ISBN 978-85-98026-63-3

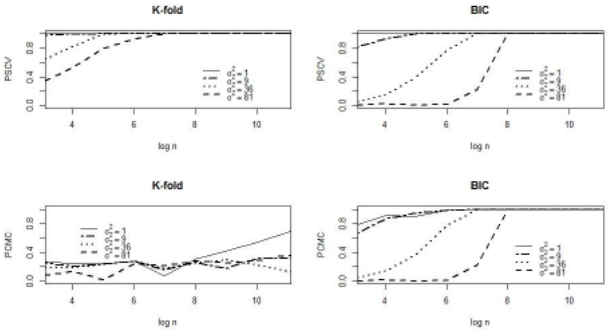
Algoritmo: Algoritmo do Zou versus Chand e Kamal não mostrou diferenças ao selecionar, mas sim no tempo, em segundos, que leva pra estimar o algoritmo. Por isso optou-se pelo algoritmo de Chand e Kamal.





Lambda: K-fold versus BIC evidenciou um ganho na seleção de variáveis ao se utilizar o BIC para estimar o parâmetro de penalização para o LASSO.

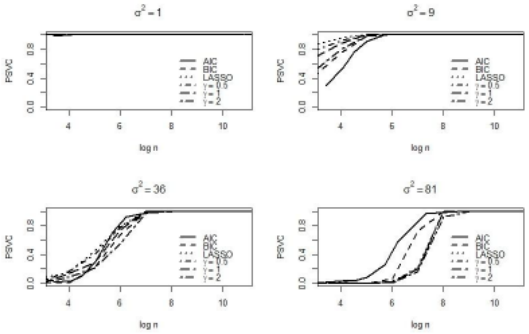
Lambda: K-fold versus BIC evidenciou um ganho ma seleção de variáveis ao se utilizar o BIC para estimar o parâmetro de penalização para o Adaptive LASSO.



Gama: Foram escolhidos os valores 0,5; 1 e 2 com base na literatura e por não apresentarem grandes diferenças entre os valores simulados.

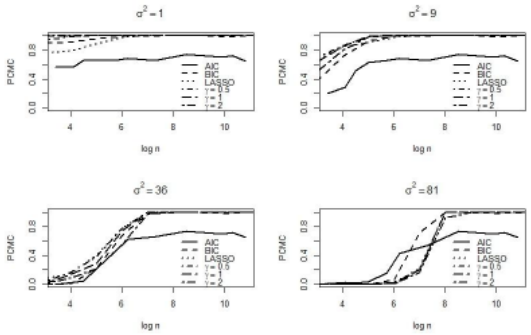
ISBN 978-85-98026-63-3

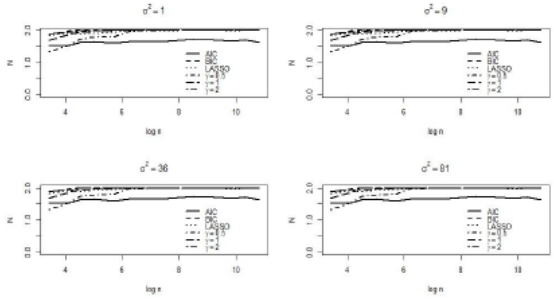
COMPARAÇÕES DOS MÉTODOS EM MODELOS LINEARES



PSVC: LASSO e Adaptive identificam as significativas com mais facilidade em pequenas amostras e baixa variabilidade do que os demais, já em grande variabilidade, o AIC e o BIC apresentam melhores resultados que os demais.

PCMC: Revela resultados semelhantes ao observado anteriormente, mas evidencia uma falha do AIC ao não conseguir estimar nem 80% dos casos independente do tamanho de amostra ou variabilidade.





O AIC não consegue estimar o número médio esperado independente o tamanho da amostra ou variabilidade.

O Adaptive consegue com amostras pequenas. O BIC só com amostras a partir de 5000.

Referências:

1) ZOU, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, Taylor & Francis, v. 101, n. 476, p. 1418{1429, 2006.

2) CHAND, S.; KAMAL, S. Variable selection by lasso-type methods. *Pakistan Journal of Statistics and Operation Research*, University of the Punjab, College of Statistical & Actuarial Science, v. 7, n. 2-Sp, 2011.

3) CHAND, S. On tuning parameter selection of lasso-type methods-a monte carlo study. In: IEEE. Applied Sciences and Technology (IBCAST), 2012 9th International Bhurban Conference on. [S.l.], 2012. p. 120-129.

ISBN 978-85-98026-63-3

Risco e tamanho da carteira de investimentos: uma análise gráfica acerca da diversificação.

Orlando Batista Damasceno (Faculdade de Natal / Estácio) / e-mail: orlandobatista.adm@hotmail.com
Luiz Carlos Santos Júnior (UNESP / UFPB) / e-mail: luiz.atuario@gmail.com

Introdução: A busca por investimentos mais atrativos e o avanço da tecnologia nos últimos anos vem popularizando o investimento em renda variável. Recentemente, sistemas que permitem a realização de negociações *online* dos ativos financeiros vêm agregando ao mercado um público que se encontrava pouco explorado. Após a popularização da internet as corretoras de valores, com o intuito de otimizar seus resultados – isto é, maximizar retorno e/ou minimizar riscos - começaram a investir nesta ferramenta. Uma das suas principais preocupações diz respeito à quantidade de *ações* que se deve manter em carteira e, apesar de a resposta ser controversa, é sabido que existem instrumentos que podem contribuir rumo à otimização.

Objetivos: Dada a importância da mensuração e do gerenciamento de risco relacionados a qualquer atividade econômica, em especial àquelas que fazem investimentos no mercado acionário, objetiva-se nesta pesquisa analisar, de forma gráfica e experimental, a relação Risco *versus* Tamanho da carteira de investimentos.

Método: Como o acesso a investidores reais é custoso (restrições orçamentárias e temporais), entrevistaram-se setenta alunos concluintes do curso de Administração e Ciências Contábeis da Faculdade de Natal (supondo-se que possuem algum conhecimento, mesmo que em nível acadêmico, acerca do mercado financeiro) e se aplicaram questionários para que esses "investidores fictícios" selecionassem ordenadamente trinta ações de suas preferências e montassem trinta carteiras (com uma, duas etc., até trinta ações). As ações foram selecionadas sem o suporte de uma análise financeira, ou seja, simplesmente com base numa lista que continha os nomes das setenta ações disponíveis (no mercado acionário brasileiro em dezembro de 2012). Com base nesses dados e nas cotações diárias das ações disponibilizadas no *Bloomberg* ao longo de cinco anos (de janeiro de 2008 a dezembro de 2012), foi possível realizar a mensuração do risco das trinta carteiras de ações montadas por cada um dos setenta entrevistados (um total de 2100 riscos calculados) utilizando-se da Teoria do Portfólio (MARKOWITZ, 1952) e dos *softwares R* e *Microsoft Excel*.

$$s_p^2(n) = \sum_{i=1}^{N} w_i^2 s_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j s_{ij}$$

s_p= desvio-padrão relativo à carteira p; w_i= peso do ativo i em relação ao total de ativos escolhidos; s_i= desvio-padrão da cotação do ativo i; s_{ij}= covariância entre as cotações dos ativos i e j.

Resultados: Dentre os principais resultados, tem-se que as curvas associadas à relação Risco x Tamanho da carteira de cada entrevistado foram classificadas de acordo com a sua disposição. A maioria dos casos individuais (mais de 95%) não apresentou o comportamento exponencial esperado, válido somente para o caso médio do grupo analisado. No entanto, os resultados também evidenciam que para alguns casos é possível diminuir o risco por meio da diversificação. Por fim, sugere-se a realização de uma análise mais robusta que considere efeitos marginais da diversificação sobre o risco, bem como uma abordagem de programação linear – ambos implementados numa linguagem em R.

Palavras-chaves: Mercado Acionário. Carteira de ações. Teoria do Portfólio. Diversificação.

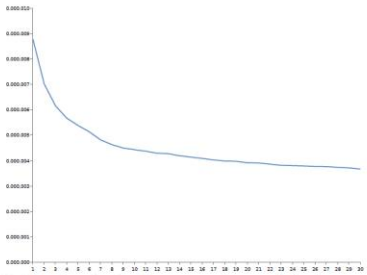


Gráfico 1 – A relação entre o risco médio da carteira e o número de ativos para todos os participantes

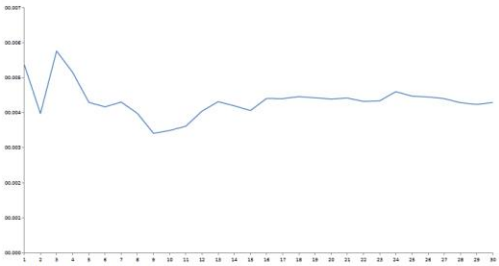


Gráfico 2 – Outros

Tabela 1 – Classificação dos formatos das curvas dos participantes individuais

Classificação	Quantidade de Participante	Porcentagens
Exponencial Errática	29	41,42%
Outros	18	25,71%
Sem Tendência	12	17,14%
Convexa	5	7,14%
Log Normal	3	4,28%
Declínio Predominantemente Exponencial	3	4,28%
Total	70	100,00%

ISBN 978-85-98026-63-3

Quantitative-Trading In R.

Daniel Karp (UFF) / e-mail: danielkarp@id.uff.br
Renato Lerípio (UFF) / e-mail: leripiorenato@gmail.com

Purpose of the paper: Show how R can help in

- Automate the decision process of quant-trading;
- Reduce the probability of operational errors.

How?

- Automatically choosing the pairs and displaying plots with the trade points;
- Calculating backtesting results.

Pairs Trading in 4-steps

1. Find a cointegrated pair of stocks and estimate the beta between them (hedge ratio);
2. Wait the spread (ê) to reach ±2 s.d.;
3. Trade the spread between them (being long on one and short on the other).
4. Chances are that the spread will come back to its mean, yielding a financial return.

Algorithm Structure

Download, import and transform prices series in returns

↓

Perform ADF tests on each stock's returns

↓

If ADF's p-value < 0.1 (stationary) then discard

If ADF's p-value > 0.1 (non-stationary) then keep

↓

Calculate two-by-two correlations

↓

If correlation >0.8 maintain

If correlation <0.8 discard

↓

Run regressions (Y=βX + e)

↓

Perform ADF test on the estimated residuals ê

↓

If ADF's p-value < 0.1 (stationary), keep the residual series => **there is a trade here!**

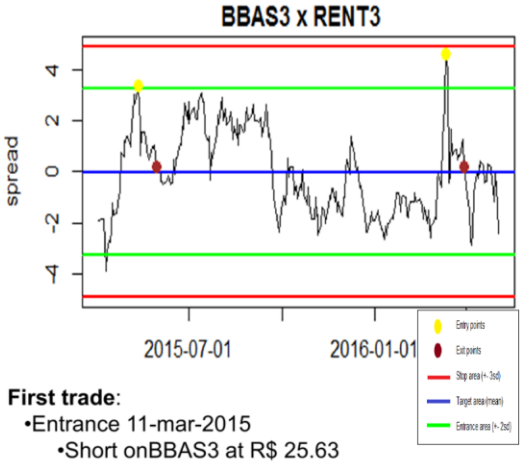
If ADF's p-value > 0.1 (non-stationary), discard the pair's residual

↓

Plot the spread series (pair's residuals) with:
- entrance lines (+- 2 s.d);
- stop lines (+- 3 s.d);
- and the target line (mean).

→ Calculate the past returns of the strategies and compare with Ibovespa (backtesting).

Trade Example



First trade:
- Entrance 11-mar-2015
  - Short onBBAS3 at R$ 25.63
  - Long on 0.6723 ($\hat{\beta}$) RENT3 at R$ 33.20
  - Spread sold at R$ 3.31.
- Exit 31-mar-15
  - Rebuy one BBAS3 at R$ 21.16
  - Sell 0.6723 RENT3 at R$ 31.25.
  - Spread rebought at R$ 0.15.
- **Result: 21% of gross profit**

Second trade:
- Entrance 28-jan-2016 →
  - Short in one BBAS3 at R$ 22.75
  - Long in 0.6723 ($\hat{\beta}$) RENT3 at 27.03
  - Spread sold at R$ 4.57
- Exit 18-fev-2016 →
  - Rebuy BBAS3 at R$ 20.27
  - Sell RENT3 at R$ 29.93.
  - Spread at R$ 0.17
- **Result: 31% of gross profit**

Accumulated result: 58.79% against -9.88% of Ibovespa!

ISBN 978-85-98026-63-3

Extracting datasets from websites and making them handy.

Renato Leripio (PPGE/UFF) / e-mail: LERIPIORENATO@GMAIL.COM
Daniel Karp (PPGE/UFF) / e-mail: DANIELKARP@ID.UFF.BR
Anna Carolina Barros (IBRE/FGV) / e-mail: ANNA.BARRO@FGV.BR

Purpose of the paper

- Extract data from websets (either datasets or page content);
- Treat data and;
- Get them handy!

How?

- Webscrapping via *Rselenium* package;
- Aggregate and change data frequency with *zoo* package.

ISBN 978-85-98026-63-3

Motivating example

Data: Inflation expectations
Source: Central Bank of Brazil

Troubles

1. Data can only be downloaded within a two years span: need to repeat the *fill in* proccess many times;

2. Data are delivered on a daily basis (not well suited frequency to work).

Challenges

1. Download all data automatically (hands free);

2. Convert data from daily to monthly basis.

Benefits

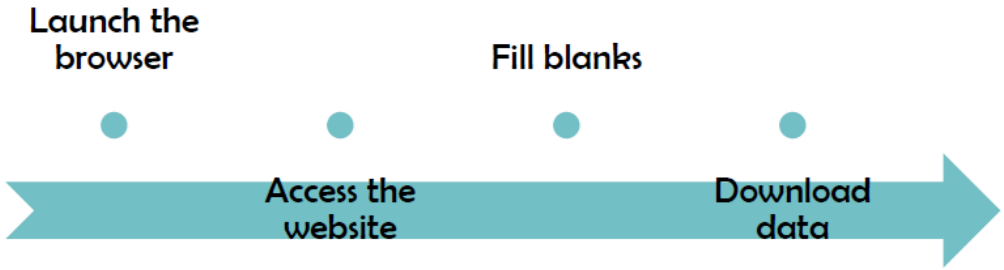
1. Save a huge amount of time;

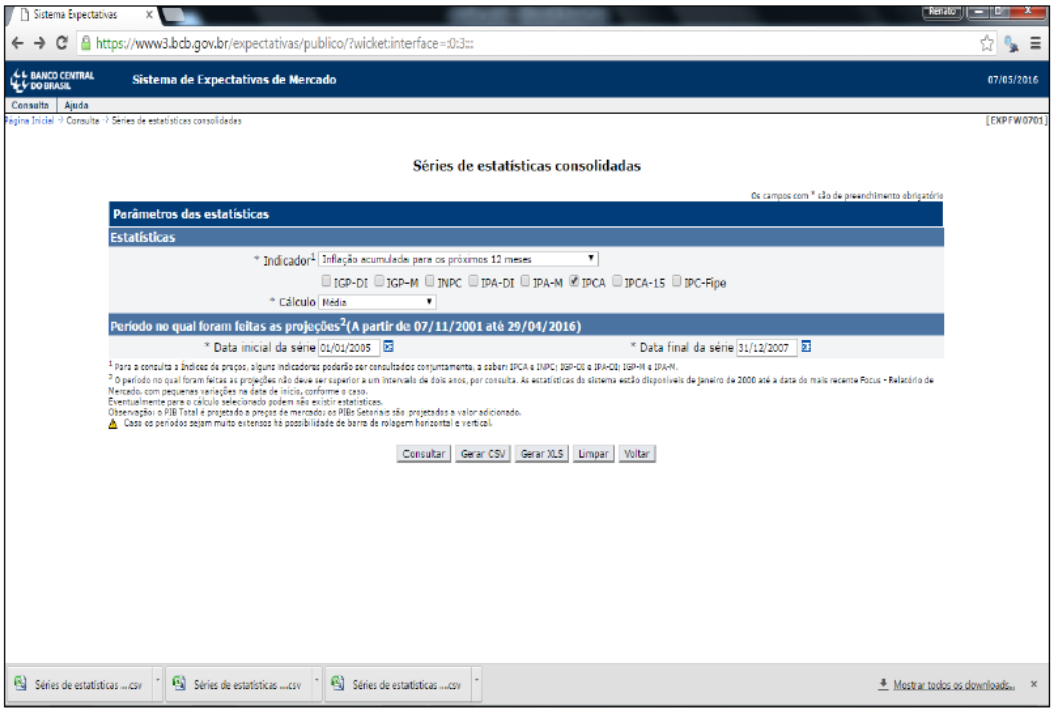
2. Completely eliminate operational errors.

Why on R?

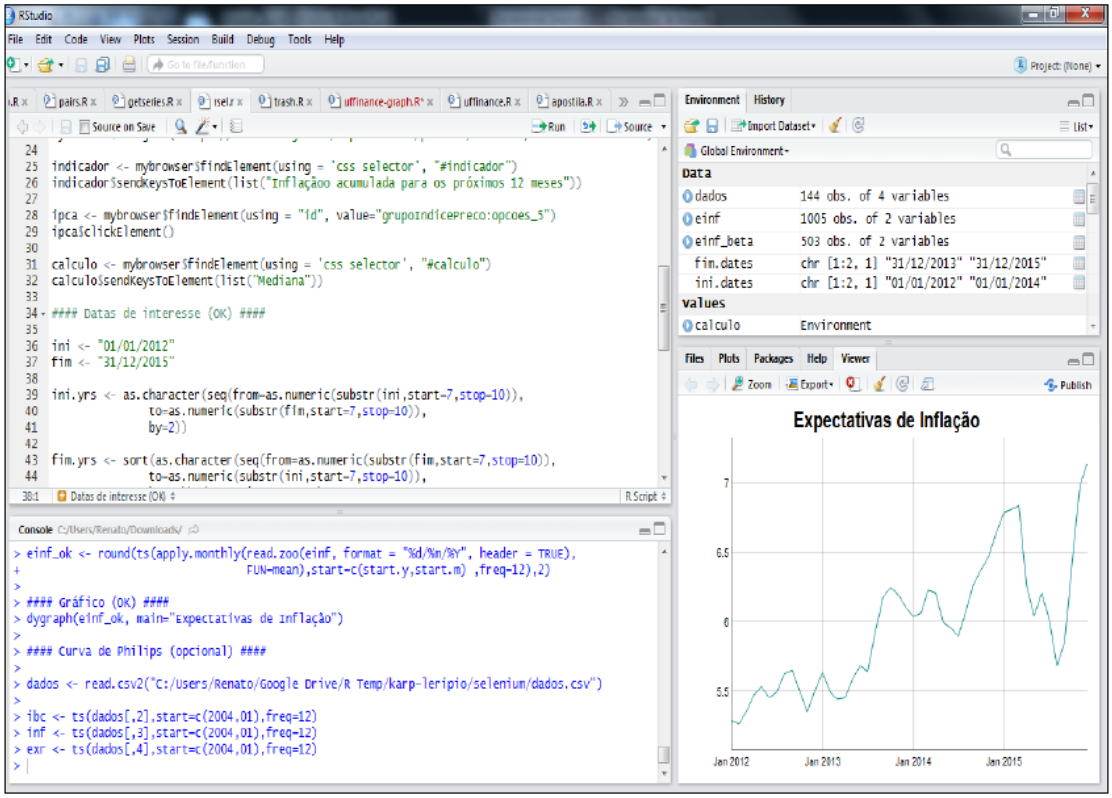
1. Data can be extracted, transformed and analyzed within the same enviroment.

ISBN 978-85-98026-63-3





ISBN 978-85-98026-63-3



ISBN 978-85-98026-63-3

Brasillegis: um pacote R para a câmara dos deputados brasileira.

Alexia Aslan (USP) / alexia.aslan@gmail.com
Leonardo Sangali Barone (USP) / leobarone@gmail.com

APRESENTAÇÃO

O bRasilLegis é um pacote de R que reúne funções para a coleta de dados disponibilizados pela Câmara dos Deputados.

A Câmara dos Deputados reúne diferentes informações sobre o processo legislativo. Por meio de um projeto chamado Dados Abertos, os dados brutos são disponibilizados em XML em um web service. O webservice oferece métodos de coleta de deputados, órgãos legislativos, proposições e sessões plenárias. A partir do web service, desenvolvemos o bRasilLegis.

Observando a difícil automatização que ocorre com a coleta de dados legislativos no campo da Ciência Política, buscamos criar uma ferramenta que tanto pesquisadores da área quanto outros profissionais pudessem acessar rapidamente as informações da Câmara dos Deputados com poucas linhas de código.

DADOS GOVERNAMENTAIS ABERTOS COM R

Seguindo o exemplo da Câmara dos Deputados, o Senado, o Governo Federal e diversas outras casas legislativas e governos locais passaram a disponibilizar seus dados via Web Services. O próximo passo de nosso projeto é incorporar esses novos Web Services dentro do bRasilLegis e de outro pacote atualmente em desenvolvimento, o bRasilGov.

Além disso, estamos atualizando o bRasilLegis para que os dados de votação no legislativo e de discursos parlamentares seja organizados respectivamente como objetos das classes rollcall (matrizes de votação) e VCorpus (corpus), de forma a permitir integração com outros pacotes de R para análise de dados, tais como pscl, wnominate (análise espacial de modelos de escolha) e tm (processamento de linguagem natural).

ISBN 978-85-98026-63-3

Planos para 2016:

- Inclusão dos dados aberto do Senado Federal no bRasilLegis;
- Criação do bRasilGov a partir da API de Compras Governamentais;
- Inclusão da API de Convênios (SICONV) no bRasilGov;
- Integração com pacotes para análise espacial de votos e processamento de linguagem natural;
- Publicação de tutoriais do bRasilLegis e bRasilGov em português e inglês.

Algumas Funções do Pacote bRasilLegis

Função	Descrição
obterAndamento	Returns a data frame containing detailed information of the progress of the requested proposition in the Brazilian Chamber of Deputies. sigla, numero and ano are required parameters.
obterDeputados	Returns a data frame containing detailed information of the active legislators at Camara dos Deputados and their respective information.
obterDetalhesDeputado	Returns a data frame containing detailed information of the legislator (for example, political parties, leaderships, committee positions, etc) at Camara dos Deputados.
obterEmendasSubstitutivoRedacaoFinal	Returns a data frame containing detailed information on the ammendments, substitutive draft and final draft of the requested proposition at Camara dos Deputados.
obterIntegraComissoesRelator	Returns a data frame containing detailed information on the committees reports of the requested proposition at Camara dos Deputados.
obterInteiroTeorDiscursosPlenario	Returns a data frame containing detailed information on every speech given in a legislative session. All the parameter of the function are required.
obterLideresBancadas	Returns a data frame containing leaders and vice-leaders of parties/coalitions at Camara dos Deputados. There are no required parameters.
obterMembrosOrgao	Returns a data frame containing all the legislators that are part of a Camara dos Deputados Organization.
obterOrgaos	Returns a data frame that lists internal Camara dos Deputados organizations (comittees for example) and respectives identification codes at the web service.
obterPartidosBlocoCD	Returns a data frame containing details of the coalitions made by the parties with representation at Camara dos Deputados.
obterPartidosCD	Returns a data frame containing details of the parties with representation at Camara dos deputados.
obterPauta	Returns a data frame with information on the legislative agenda of a Camara dos Deputados stance between an initial and a final date.
obterProposicao	Returns a data frame containing detailed information the requested proposition at Camara dos Deputados and respectives attached propositions ("proposicoes apensadas").
obterProposicaoPorID	Returns a data frame containing detailed information the requested proposition at Camara dos Deputados and respectives attached propositions ("proposicoes apensadas").
obterRegimeTramitacaoDespacho	Returns a data frame containing detailed information on the processing status of the requested proposition at Camara dos Deputados.
obterVotacaoProposicao	Returns a data frame containing all the roll call votes for the requested proposition and individual votes at Camara dos Deputados.

ISBN 978-85-98026-63-3

INSTALAÇÃO E EXEMPLO

O pacote está disponível em:
https://github.com/leobarone/bRasilLegis

Para instalar, utilizamos o pacote devtools e os comandos:

```
> library(devtools)
> install_github("leobarone/bRasilLegis")
```

Apresentamos abaixo exemplos do uso do bRasilLegis para votações em plenário na Câmara dos deputados. Por exemplo, se quisermos saber todas as proposições votadas em 2014 pelo plenário basta fazer:

```
> plenario <- listarProposicoesVotadasEmPlenario(2014)
> head(plenario)

  codProposicao nomeProposicao dataVotacao
1         19319   PL 3232/1992  10/06/2014
2         43617   PLP 275/2001  22/04/2014
3         44196   PLP 276/2002  23/04/2014
4        302638   PL 6025/2005  04/02/2014
5        302638   PL 6025/2005  11/02/2014
6        304008   PEC 471/2005  06/05/2014
```

Com as informações sobre o nome de uma proposição específica votada em plenário – por exemplo, o PL 6025 de 2005 – podemos observar o voto individual dos parlamentares utilizando a função:

```
> votos <- obterVotacaoProposicao("PL", "6025",
"2005")
> head(votos[,c(9:13)])

          Nome   ideCadastro Partido   UF  Voto
1  Isaias Silvestre     74152 PSB    MG Não
2 Francisco Tenório    141467 PMN    AL Não
3  Nilmário Miranda     74751 PT     MG Não
4  Manuel Rosa Neca    171622 PR     RJ Obstrução
5      Urzeni Rocha    141551 PSD    RR Não
6        Fabio Reis    171623 PMDB   SE Não
```

listarProposicoesVotadasEmPlenario() e obterVotacaoProposicao() podem ser facilmente combinadas para retornar todas as votações em plenário para uma legislatura específica ou para um período mais longo de anos.

ISBN 978-85-98026-63-3

Portfolio optimization and risk analysis in R.

Daniel Karp (UFF) / e-mail: danielkarp@id.uff.br
Victor Mamede (UFF) / e-mail: victorhfmamede@gmail.com

Purpose of the paper: show how R can efficiently

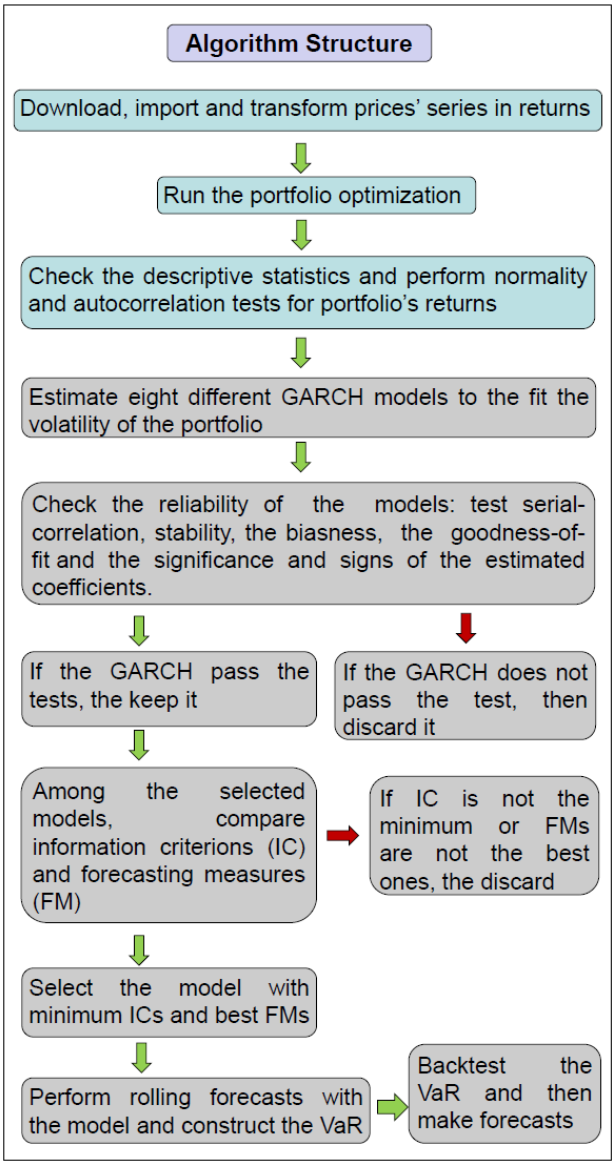
- Run portfolio optimizations, with different constraints;
- Estimate risk models, with different specifications.

Why is it efficient?

- Just one software and one code for both routines

⇒ Reduce the probability of manual errors;

⇒ Enhance productivity.



ISBN 978-85-98026-63-3

Portfolio Optimization

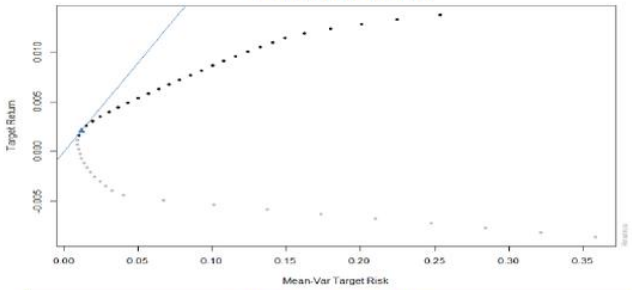
1. Find the combination of a given set of assets that yields a maximum mean/variance ratio.

Risk Analysis (VaR with GARCH models)

1. Calculate the volatility of the portfolio;
2. Perform backtesting to test the fit of your model;
3. Make forecasting analysis to test the exposure of the financial position for future periods.

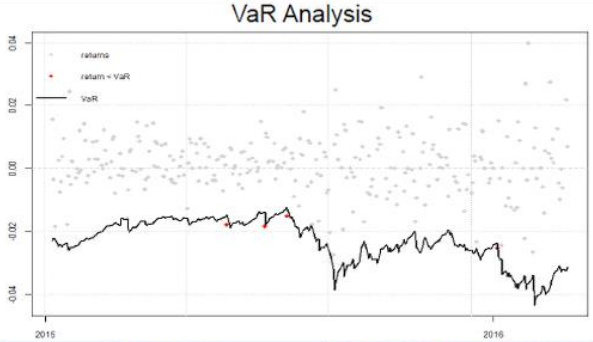
Example

Efficient Frontier



Efficient Portfolio			
Assets	**Weights**	**Assets**	**Weights**
ABEV3	8.92%	HYPE3	3.56%
BRFS3	8.82%	KROT3	0.01%
BRKM5	0.80%	QUAL3	1.99%
BRPR3	3.43%	RADL3	8.79%
CTIP3	14.78%	RUMO3	0.04%
EMBR3	10.80%	SMLE3	5.10%
EQTL3	13.13%	SUZB5	5.89%
FIBR3	10.12%	VIVT4	1.88%

VaR Analysis



Kupiec test		VaR Backtest Report	
Null-Hypothesis	Correct Exceedances	**alpha**	1.0%
LR.uc Statistic	0.305	**Expeceted exceed**	3
LR.uc Critical	3.841	**Actual VaR exceed**	4
Reject Null	NO	**Actual %**	1.3%

ISBN 978-85-98026-63-3

Análise espacial de dados socioeconômicos como suporte para caracterização e identificação de áreas degradadas na bacia hidrográfica do rio imboaçu, São Gonçalo.

Antonio da Cunha Nunes(UFF e Universo) / e-mail: nunes.antoniocunha@gmail.com
Fernando Benedicto Mainier (UFF) / e-mail: mainier@vm.uff.br
Viviane da Silva de Alcântara (Universo) / e-mail: vialcantara@gmail.com

Introdução

As práticas humanas de ocupação do espaço acarretam o comprometimento do ar, do solo, da água, da fauna e flora, produzindo infinidades de demandas ambientais, econômicas, sociais e demográficas. A bacia hidrográfica do Rio Imboaçu (São Gonçalo) (Figura 1) sofre com tais fenômenos, oriundos, principalmente, da carência de infraestrutura urbana, isto é, dos serviços básicos fornecidos pelo poder público e que visam garantir a qualidade de vida da população e, concomitantemente, a preservação do meio ambiente. O objetivo é identificar as áreas que carecem de infraestrutura urbana na Bacia Hidrográfica do Rio Imboaçu (BHRI).

Figura 1: Bacia hidrográfica do Rio Imboaçu, São Gonçalo.

ISBN 978-85-98026-63-3

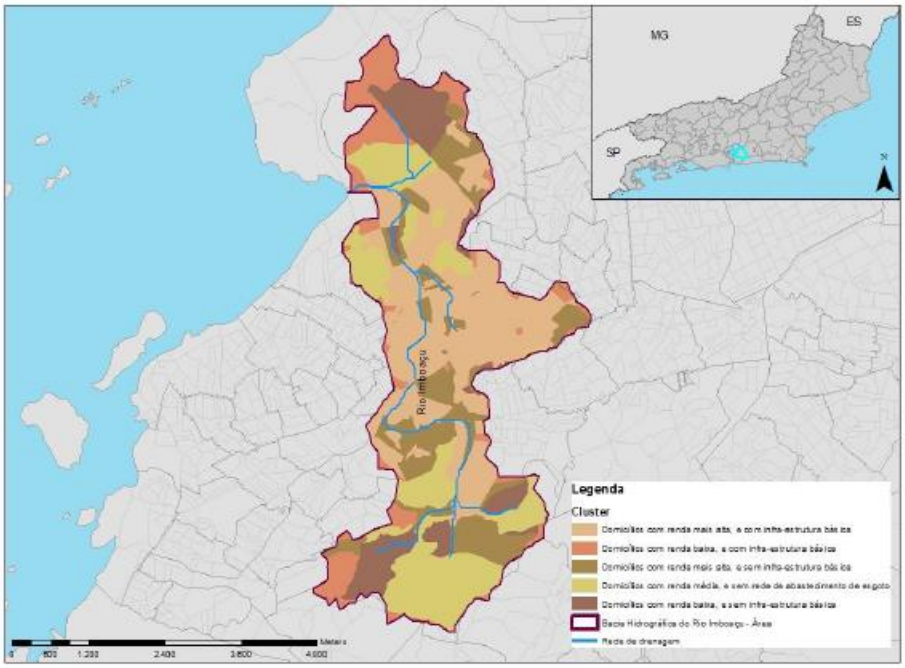


Figura 3: Análise de cluster e sua espacialização na BHRI. Evidenciam áreas que carecem de infraestrutura urbana (Figura 3).

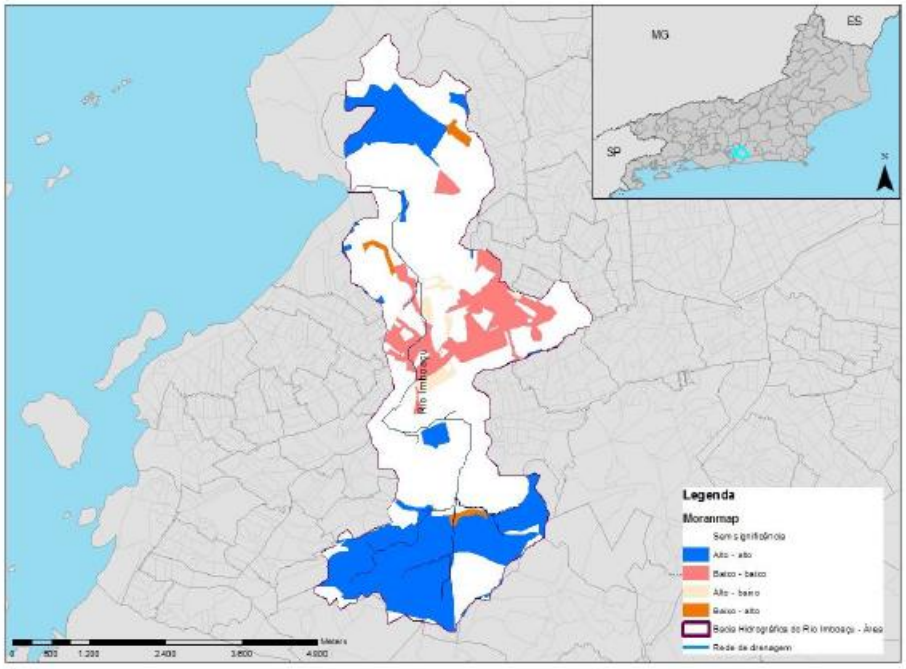


Figura 4 : Índice Global de Moran dos clusters.

ISBN 978-85-98026-63-3

Metodologia

Consistiu de três etapas: A primeira etapa consistiu em identificar as variáveis socioeconômicas e demográficas que manifestam as ações antrópicas no ambiente, essas foram: média de moradores por domicílio, renda média por domicílio, destino dos resíduos sólidos, abastecimento irregular de água, esgotamento inadequado (Figura 2). A segunda etapa consistiu na análise exploratória dos dados. E, a última etapa, consistiu na identificação dos clusters que evidenciam áreas que carecem de infraestrutura urbana (Figura 3).

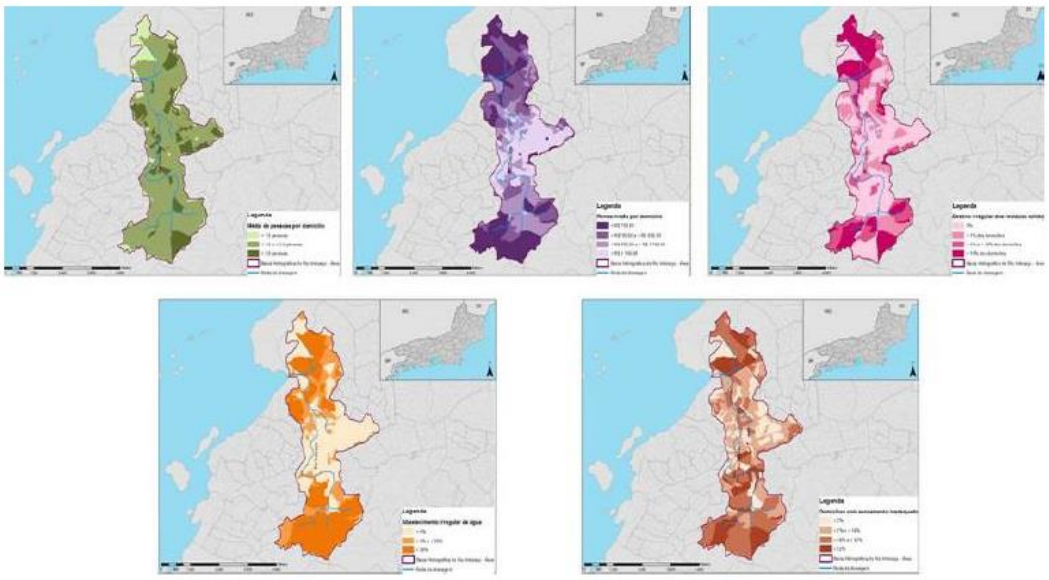


Figura 2: Análise exploratória dos dados e a sua respectiva espaciliazação, na sequência: média de moradores, renda média, destino dos resíduos sólidos, abastecimento irregular de água, esgotamento inadequado por domicilio da BHRI.

ISBN 978-85-98026-63-3

Análise espacial do clusters criados

Como resultados dessa investigação espacial dos clusters, verificou-se que o índice global de Moran (Figura 4) igual a 0,345, com p-valor igual a 0,001, logo há autocorrelação espacial positiva estatisticamente entre os indicadores selecionados com cada setor censitário e com os seus vizinhos. Numa análise mais detalhada, realizada através dos índices locais de Moran (Figura 5), observa-se que a existência de associações espaciais muito significativas entre as áreas que carecem de infraestrutura urbana e áreas degradadas, com confiabilidade de 99,9% em todos os setores da bacia.

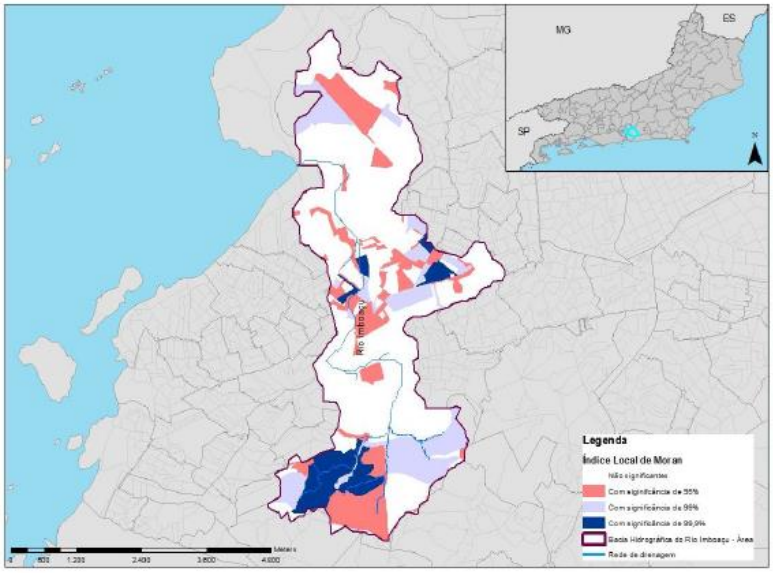


Figura 5: Índice Local de Moran.

ISBN 978-85-98026-63-3

Regras de associação em R: análise do pacote "arules".

Eduardo C. Gonçalves (ENCE/IBGE) / e-mail: eduardo.correa@ibge.gov.br
André Bruno de Oliveira (IBGE) / e-mail: andre.oliveira@ibge.gov.br
Elon Martins de Sá (IBGE) / e-mail: elon.sa@ibge.gov.br

Introdução

A extração de **regras de associação** [1,2,3] é um dos mais conhecidos problemas de análise de *Big Data*. Neste problema, **algoritmos** são utilizados para identificar, de forma automática, **relacionamentos entre itens** de uma base de dados. Este trabalho apresenta **arules**, o pacote de regras de associação do ambiente **R**.

Regras de Associação

Uma regra de associação é uma expressão da forma $A \Rightarrow B$, onde *A* e *B* podem ser conjuntos compostos por um ou mais itens. Este tipo de regra representa um **relacionamento** extraído de uma base de dados. O componente *A* é chamado de **antecedente** da regra e *B* de **consequente**.

A análise de cestas de compras (*market basket analysis*) é uma das aplicações típicas das regras de associação. Neste problema, o objetivo é examinar uma base de dados de transações de compras, como a apresentada na **Figura 1**, para determinar produtos que costumam ser adquiridos em conjunto.

TID	Lista de Itens
1	biscoito, cerveja, chá, salaminho
2	cerveja, pão, queijo
3	café, pão
4	chá, pão
5	café, cerveja, pão, salaminho

Figura 1 – Base de dados de transações

Um exemplo de regra de associação que poderia ser extraída da base de dados acima é:

- *67% dos clientes que compram o produto 'cerveja' também compram o produto 'salaminho' ; 40% de todas as compras possuem ambos os produtos"*
- **Representação:** $\{cerveja\} \Rightarrow \{salaminho\}$
  - Confiança = 67% Suporte = 40%

ISBN 978-85-98026-63-3

Como Trabalham os Algoritmos de Regras de Associação?

Normalmente, os algoritmos dividem o trabalho em duas fases:

• **Fase 1 (DIFÍCIL)**: determinar **todos** os conjuntos de itens que possuam suporte igual ou superior a *SupMin* (suporte mínimo especificado pelo usuário).

Para *n* itens, a quantidade de combinações possíveis (espaço de busca) é igual a $2^n - 1$.

{café}	{café, leite}	{café, leite, manteiga}
{leite}	{café, manteiga}	{café, leite, pão}
{manteiga}	{café, pão}	{café, manteiga, pão}
{pão}	{leite, manteiga}	{leite, manteiga, pão}
	{leite, pão}	{café, leite, manteiga, pão}
	{manteiga, pão}	

• **Fase 2 (FÁCIL)**: Para cada conjunto frequente encontrado na Fase 1, gerar as regras que possuem confiança igual ou superior a *ConfMin* (confiança mínima).

O algoritmo mais conhecido (e um dos mais eficientes) para contagem de suporte é o **APRIORI** [1]. Suas características principais:

• Faz a contagem de suporte **por etapas**, cada uma tratando de conjuntos de diferentes comprimentos (iniciando pelo comprimento 1).

• **Não** tenta contabilizar o suporte de todas as combinações de itens ($2^n - 1$), mas apenas das "boas combinações".

  • **Ex.**: se {*café*, *pão*} não é frequente, não é preciso avaliar {*café*, *leite*, *pão*}.

  • "*todo superconjunto de um conjunto infrequente é necessariamente infrequente*"

O Pacote "arules"

O pacote "arules" [3] oferece um **ambiente completo** para a exploração da técnica de regras de associação em **R**. Seus principais pontos positivos são: (i) utilização do APRIORI para determinar os conjuntos frequentes; (ii) disponibilização de mais de 30 medidas para avaliar o grau de interesse de uma regra de associação. Neste trabalho, além do suporte e da confiança, foram utilizadas duas medidas adicionais:

• ***phi*** : coeficiente de correlação de Pearson.

• ***lift*** : quanto mais frequente torna-se *B*, quando A ocorre: Conf($A \Rightarrow B$) ÷ Sup(B).

ISBN 978-85-98026-63-3

Experimento

Para avaliar o pacote "arules", foi realizado um experimento em que a técnica de regras de associação foi utilizada sobre uma base de dados que armazena as **sessões de usuários** de um portal da Internet. O intuito foi o de identificar os padrões de associação entre as diferentes páginas. Foi utilizada uma base de dados pública, que registra os acessos ao portal da *Microsoft* (**Figura 2**). Ela contém 32.711 transações (sequência de páginas visitadas por um usuário em uma sessão), 284 itens (páginas do portal) e está no formato *single* (cada linha tem o TID + página visitada).

```
10001 End_User_Produced_View
10001 Support_Desktop
10001 regwiz
10002 Knowledge_Base
10002 Support_Desktop
10003 Knowledge_Base
10003 Microsoft.com_Search
10003 Support_Desktop
10004 Norway
10005 misc
10006 Knowledge_Base
10006 Microsoft.com_Search
10007 International_IE_content
```

Figura 2 – Base de dados da Microsoft

O pacote "arules" foi utilizado para resolver o seguinte problema: encontrar todas as regras de associação de comprimento menor ou igual a 3, que possuam suporte, confiança e *lift* maiores ou iguais, respectivamente, a 0,1%, 30% e 1,0. A **Figura 3** ilustra o *script* R utilizado.

```r
#importa o package
library("arules")

#importa a base para um objeto do tipo "transactions"
ms_td <- read.transactions("c:\\tmp\\ms.txt", format = "single", cols = c(1,2))

#faz a contagem de suporte e gera as regras utilizando o algoritmo Apriori
#parâmetros: SupMin=0,1%, ConfMin =30%, comprimento máximo = 3
rules <- apriori(ms_td,parameter = list(support = 0.001, confidence = 0.3, maxlen=3))

#corte no lift > 1.0
rulesLift <- subset(rules, lift > 1.0)

#adiciona a medida de interesse "phi"
#(o "suporte", "confiança" e "lift" são computados por default)
quality(rulesLift) <- cbind(quality(rulesLift),
interestMeasure(rulesLift,c("phi"),transactions=ms_td,reuse=TRUE))

#salva os resultados (regras geradas) em um arquivo CSV separado por ";"
write(rulesLift, file = "c:\\tmp\\regras.csv", sep = ";", col.names = NA)
```

Figura 3 – *Script* R

ISBN 978-85-98026-63-3

Resultados

Foram extraídas 2.814 regras. A **Tabela 1** apresenta alguns exemplos:

Tabela 1 – Exemplos de regras de associação extraídas pelo "arules"

REGRA	Sup	Conf	Lift	Phi
{Knowledge Base} ⇒ {Support Desktop}	5,52%	60,85%	4,47	**0,43**
{Internet_Explorer} ⇒ {Free_Downloads}	**16,08%**	56,05%	1,69	0,31
{MS_Word} ⇒ {MS_Word_News}	0,94%	79,03%	**30,70**	**0,53**
{Internet Development ∧ Web Site Builders Gallery} ⇒ {Developer Workshop}	0,17%	**91,80%**	20,02	0,18
{Windows Family of OSs} ⇒ {Free_Downloads}	**7,79%**	55,08%	1,66	0,19
{Job Listings for Pre-Grads} ⇒ {Job Openings}	0,13%	**89,13%**	**73,62**	0,30

Os resultados evidenciam que as regras obtidas pelo pacote "arules" representam uma ferramenta valiosa para a caracterização dos diferentes perfis de usuários do portal. Observe, por exemplo, a regra ***{Knowledge Base} ⇒ {Support Desktop}***:

- 5,52% dos visitantes consultaram ambas as páginas (suporte).
- A probabilidade de um usuário ter visitado a página **Support Desktop**, dado que ele também visitou **Knowledge Base** é de 60,85% (confiança).
- A visita à página **Support Desktop** é 4,47 vezes maior entre os usuários que também visitam **Knowledge Base** (*lift*). E o coeficiente de correlação entre as páginas é de 0,43 (*phi*).

Referências

[1] HAN, J.; KAMBER, M.; PEI, J. Data mining: Concepts and techniques. Morgan Kaufmann, 3rd ed., 2011.

[2] GONÇALVES, E. C. Regras de associação e suas medidas de interesse objetivas e subjetivas. INFOCOMP, v.4, n.1, 2005.

[3] HAHSLER, M.; BUCHTA, C.; GRUEN, B.; HORNIK, K. arules: Mining association rules and frequent itemsets. 2009.

ISBN 978-85-98026-63-3

Propriedades eletromagnéticas mostram potencial para mapear atributos do solo correlacionados em R.

Hugo M Rodrigues (UFF) / hugomr@id.uff.br
Gustavo M Vasques (Embrapa Solos) / gustavo.vasques@embrapa.br

1. Introdução:

Enquanto o Brasil apresenta grande potencial para suprir a demanda mundial de alimentos, fibras e energia, a base de informações sobre os solos, incluindo índices de fertilidade e aptidão para uso agrícola, encontra-se defasada e inapropriada para dar suporte à tomada de decisões estratégicas para o futuro da agropecuária no país. O manejo ideal ou ótimo dos solos carece de informações sobre a distribuição espacial contínua dos atributos do solo, atualizadas e em escalas detalhadas (>1:25.000) ou ultra detalhadas (>1:5.000). No entanto, o levantamento de solos usando os métodos clássicos já consagrados é oneroso. Abordagens mais recentes para a caracterização e mapeamento de solos, utilizando métodos estatísticos modernos, como o mapeamento digital de solos, permitem incorporar, dados advindos de bases cartográficas, sensores remotos e observações pontuais de solos (incluindo dados de sensores proximais). Para então elaborar mapas contínuos dos atributos do solo. Nesse sentido, a utilização de sensores proximais como solução para redução dos custos de amostragem para a caracterização da variação dos solos em maior escala espacial (>1:25.000) torna-se promissora.

2. Objetivos:

1) Mapear propriedades eletromagnéticas (EM) e atributos do solo usando métodos geoestatísticos; e 2) Avaliar a qualidade dos mapas quanto à incerteza.

3. Material e método

3.1 Área de estudo e malha de pontos

A área de estudo possui aproximadamente 3,4 ha e está no município de Seropédica, RJ. Para leitura das propriedades EM do solo usando sensores proximais de campo, foram alocados 377 pontos amostrais distribuídos em malha de 10 x 10 m, constituída por 13 transectos longitudinais à topossequência com 29 pontos cada um. A partir dessa malha, foi definida uma malha reduzida de 20 x 20 m (105 pontos), acrescida de 25 pontos alocados usando o método do hipercubo latino condicionado, para coleta de amostras em 0-10 e 10-20 cm de profundidade, destinadas à análise dos atributos do solo em laboratório.

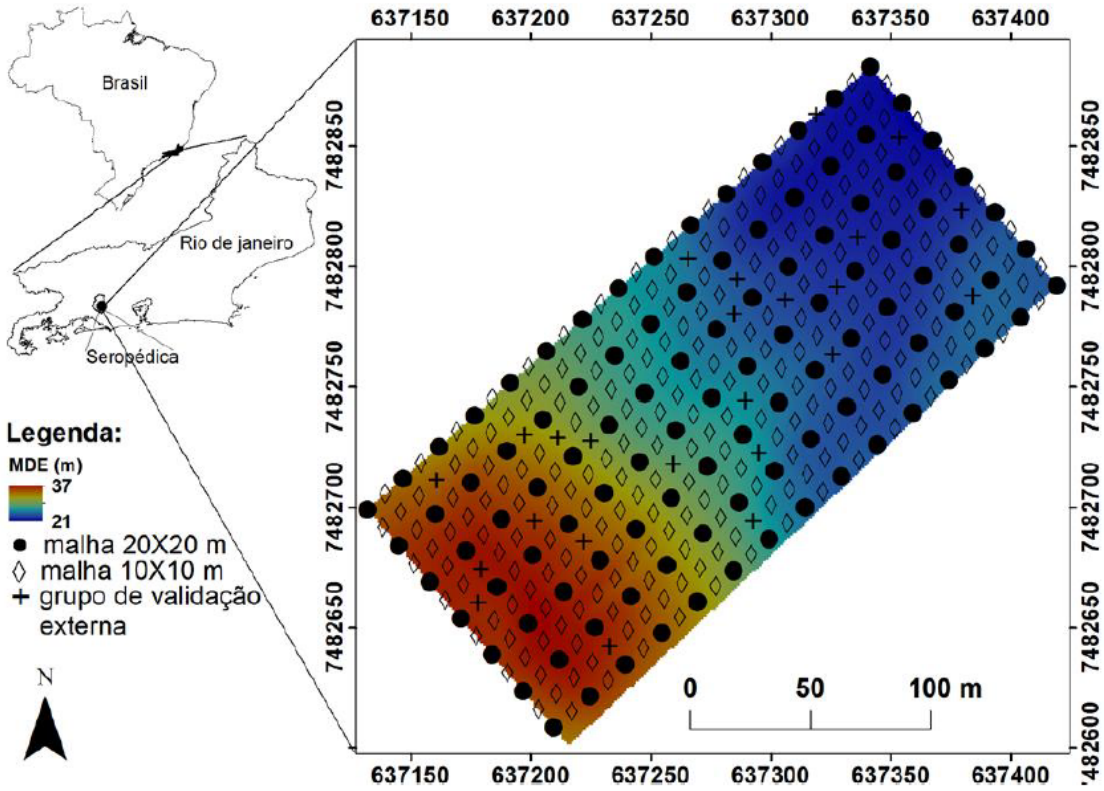


Figura 1. Mapa de localização da área de estudo, modelo digital de elevação (MDE) e delineamento amostral. A malha de 20 x 20 m e os 25 pontos de validação externa estão dispostos sobre a malha de 10 x 10 m. Projeção WGS84 UTM 23S.

ISBN 978-85-98026-63-3

3.2. Aquisição de amostras por Sensores e laboratório

Nos 377 pontos da malha de 10 x 10 m, foram obtidas leituras dos seguintes sensores proximais: 1) KT-10 S/C, que mede a condutividade elétrica aparente (CE) e a susceptibilidade magnética (SM); e 2) RS-230 BGO, que mede o teor dos elementos gamarradioativos potássio (K), urânio (U) e tório (Th). Os dados de laboratório utilizados foram: teor de argila (Argila), teor de carbono orgânico (COrg) e teor de ferro total extraído por ácido sulfúrico (Fe total).

Figura 2. (A) Pesquisador realizando leitura do topo do solo com sensor KT-10 S/C enquanto o sensor RS-230 BGO está funcionando; (B) Mensurando em sub-superfície com o sensor KT-10 S/C (B); (C) Preparando trincheira de até 20 cm de profundidade para coleta das amostras para laboratório, enquanto está sendo preparado o sensor RS-230 BGO; (D, F) Exemplo dos painéis dos sensores com leituras realizadas pelo KT-10 S/C e RS-230 BGO (respectivamente); (E) Abertura de trincheira e coleta de amostra de laboratório, também pode ser visto o equipamento RS-230 BGO em funcionamento.

3.3. Interpolação e análise estatística

A partir das amostras de 0-10 cm da malha de 20 x 20 m, utilizou-se krigagem ordinária para interpolação dos dados e produção dos mapas das propriedades EM e atributos do solo. Usou-se transformação para logaritmo nas variáveis Fe total, CE, SM e U, enquanto o Argila, COrg e Th apresentaram distribuição de frequência aproximadamente normal. A qualidade dos mapas foi avaliada usando-se índices de incerteza obtidos por validação externa usando as 25 amostras alocadas via hipercubo latino. As análises estatísticas foram feitas no programa R, utilizando o pacote gstat.

ISBN 978-85-98026-63-3

4. Resultados

4.1. Estatística descritiva dos dados

Estatística	COrg (g kg^{-1})	Argila (g kg^{-1})	CE (S m^{-1})	Fe2O3 (g kg^{-1})	SM (10^{-3} SI)	Th (ppm)	U (ppm)
	Valores originais						
Observações	129	129	129	129	129	129	129
Mínimo	3.70	25.00	0.00	10.00	0.02	0.00	0.50
Máximo	28.00	380.00	4.90	84.00	3.80	13.87	3.40
Média	11.27	175.08	1.20	28.57	0.35	6.29	1.27
Mediana	11.40	160.00	0.80	26.00	0.18	5.80	1.20
Desvio padrão	3.86	93.34	1.09	14.39	0.45	2.77	0.43
Assimetria	0.65	0.26	1.17	1.35	4.22	0.45	1.31
	Valores em logaritmo						
Observações			129	129	129		129
Mínimo			0.00	2.40	-3.91		0.41
Máximo			1.77	4.44	1.34		1.48
Média			0.68	3.28	-1.54		0.80
Mediana			0.59	3.30	-1.71		0.79
Desvio padrão			0.46	0.46	0.98		0.18
Assimetria			0.31	0.15	0.12		0.50

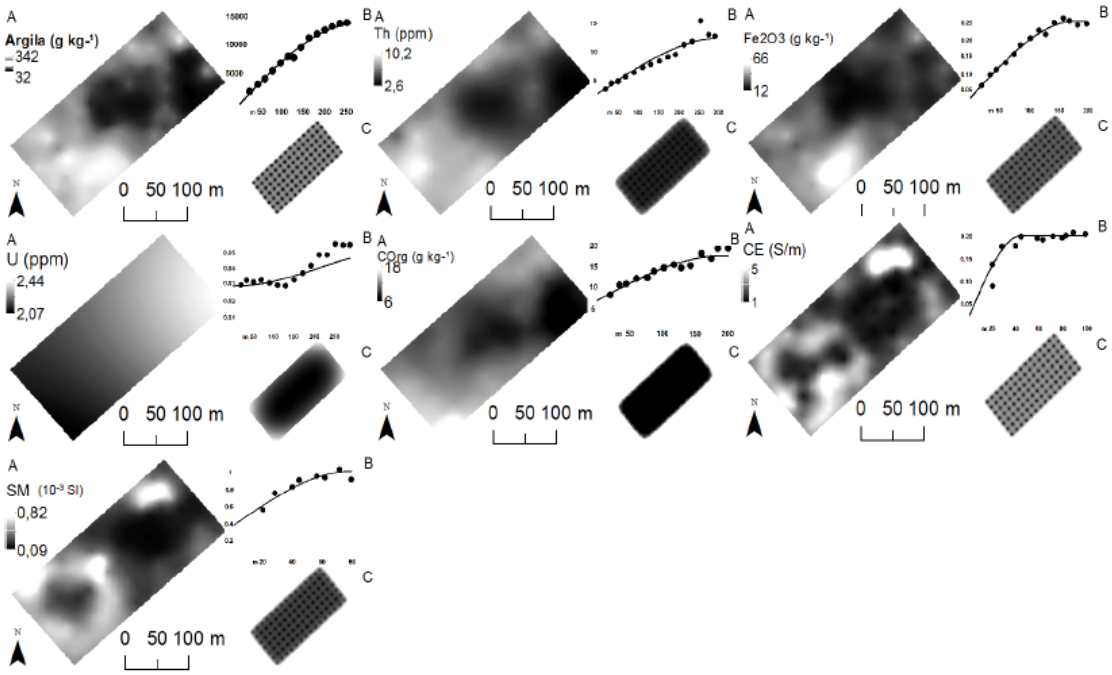
Tabela 1. Estatísticas descritivas das variáveis do solo.

	SM	U	Fe2O3	CE	Argila	COrg	Th
SM	1	-0.23*	0.48*	0.72*	0.54*	0.34*	0.38*
U		1	-0.11ns	-0.25*	-0.1ns	-0.02ns	0.09ns
Fe2O3			1	0.48*	0.87*	0.54*	0.66*
CE				1	0.49*	0.38*	0.36*
Argila					1	0.65*	0.78*
COrg						1	0.54*
Th							1

Tabela 2. Índice de correlação de Pearson. *, significativo ao nível de 0.05; ns, não significativo

ISBN 978-85-98026-63-3

Os índices de correlação linear foram significativos ($p > 0.05$) entre a maioria das variáveis do solo, principalmente com o teor de argila. Pode-se inferir que as variáveis correlacionadas com a argila (COrg, Fe total e Th) encontram-se ou derivam de constituintes e processos que se desenvolvem predominantemente nessa fração (SM, CE). Por exemplo, o Fe total seria um constituinte com participação importante na SM e CE do solo. A exceção foi o teor de U, que inclusive mostrou semivariogramas e padrões espaciais contrastantes. Os semivariogramas do teor de argila, Th, Fe total e COrg apresentaram valores de alcance maiores (177 a 304 m) do que os de CE e SM (44 a 80 m). Observou-se uma forte relação entre o padrão de distribuição espacial das variáveis do solo e o relevo da área. Por exemplo, no sentido topo-baixada, para Argila, Th, COrg e Fe total, valores maiores foram observados nas cotas mais altas da área (topo plano), seguidos de valores menores na encosta suave, que culminam em uma planície de inundação (baixada), com valores menores de atributos ao longo do canal de drenagem, e os valores mais baixos no aluvião de entorno, incluindo o extremo leste-nordeste da área. Para a CE e SM, essas relações na planície de inundação são aproximadamente inversas, com menores valores ao longo do canal de drenagem e maiores valores nas áreas aluviais ao redor. Ambos os sensores utilizados (KT-10 S/C e RS-230 BGO) apresentam potencial para estimar o padrão de distribuição espacial da Argila, COrg e Fe total do solo, sendo que as causas dos padrões observados, especialmente daqueles divergentes na área do canal de drenagem, devem ser investigadas mais a fundo.



Figuras 3 a 9. Mapas das propriedades Argila, Th, Fe total, U, COrg, CE e SM para 0-10 cm de profundidade; A: valores preditos; B: semivariogramas empírico (pontos) e ajustado (linha); C: variância da krigagem.

5. Conclusões:

As propriedades EM do solo, de mais fácil medição em campo usando sensores proximais, mostram excelente potencial para avaliar a distribuição espacial de atributos correlacionados do solo, complementando ou substituindo as análises de laboratório. A utilização do R permitiu a integração dos dados espaciais e geração de resultados em pouco tempo, com a vantagem de ser um programa gratuito e colaborativo. Outras questões ainda passíveis de investigação incluem: 1) Que fatores e/ou processos de formação do solo podem explicar os padrões espaciais observados? 2) Qual a influência do número de observações nos parâmetros de dependência espacial e mapas derivados das propriedades EM do solo? e 3) Que outros atributos do solo podem estar correlacionados com as propriedades EM estudadas?

6. Referência:

EMBRAPA SOLOS. **Manual de métodos de análise de solo**. 2.ed. rev. Rio de Janeiro: Embrapa Solos, 2011. (Documentos, 132)
MINASNY, B., MCBRATNEY, A.B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, v. 32, p. 1378–1388, 2006.
PEBESMA, E.J. Multivariable geostatistics in S: the gstat package. **Computers & Geosciences**, v. 30, p. 683–691, 2004.
R CORE TEAM. **R**: a Language and Environment for Statistical Computing. Viena, Áustria: R Foundation for Statistical Computing, 2015.

ISBN 978-85-98026-63-3

Concentração de renda e sua associação com algumas características das maiores empresas da Europa: uma análise usando modelo de regressão ordinal.

Selma Alves Dios (UFF) / e-mail: selmadios@vm.uff.br
Gabriel de Aguiar Mendonça (UFF) / e-mail: gabrieldeaguiarmendonca@gmail.com
José Rodrigo de Moraes (UFF) / e-mail: jrodrigo78@gmail.com

1. Introdução

Uma relevante forma de avaliar a responsabilidade social das empresas é a que considera sua contribuição a uma distribuição de renda mais equitativa entre os agentes que contribuem para a geração de riqueza pela empresa. Se a riqueza gerada pela atividade da empresa se concentra, não há desenvolvimento do entorno. Isso se pode avaliar pela conformação do quadro de distribuição e concentração de renda baseado na maior ou menor proporção do valor adicionado concentrado. Considerando que uma maior geração de valor adicionado depende de uma maior produtividade do trabalho, assim como gera maior rentabilidade do capital, importa saber de que forma a produtividade e a rentabilidade estão relacionadas com a concentração de renda realizada pelas empresas, assim como variáveis intervenientes neste quadro, como o setor de atividade e o faturamento bruto das empresas.

2. Objetivo

Este trabalho tem como objetivo estimar a chance de concentração de renda pelas maiores empresas da Europa no ano de 2012, a partir de suas características, tais como produtividade, rentabilidade, setor de atividade e faturamento bruto.

3. Método

Foi utilizado o modelo de regressão logística ordinal, também denominado modelo de chances proporcionais, onde a variável resposta do modelo é um índice de concentração de renda (ICR), criado com base nas rubricas componentes do valor adicionado das empresas, especificamente Depreciação e Resultado do Exercício. Este indicador foi categorizado em três níveis: baixa ($ICR \leq 0{,}25$), moderada ($0{,}25 < ICR \leq 0{,}436$) e alta ($ICR > 0{,}436$) intensidade de concentração de renda. Utilizou-se o programa R, versão 3.2.4, para ajustar o referido modelo, considerando uma amostra de 286 empresas, para as quais se tinha informação sobre o ICR. Os dados foram obtidos da base de dados Amadeus, que reúne dados contábeis de aproximadamente 11 milhões de empresas em toda a Europa.

ISBN 978-85-98026-63-3

O modelo de regressão logística multinomial é representado pela seguinte equação:

$$ln\left(\frac{p_{ij}}{p_{ij}}\right) = \boldsymbol{X}_i'\boldsymbol{\beta}_j \qquad \begin{matrix} i = 1, 2, \dots, n \\ j = 1, 2, \dots, J-1 \end{matrix}$$

onde:

$\boldsymbol{X}_i = \left(1, X_{i1}, X_{i2}, \dots, X_{ik-1}\right)$` é o vetor coluna, de dimensão k x 1, composto por k-1 valores das variáveis explicativas referentes ao i-ésimo elemento;

$\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jk-1})'$ é o vetor coluna, de dimensão k x 1, composto por k parâmetros desconhecidos referentes a j-ésima categoria da variável resposta do modelo;

$p_{ij} = P(Y_i = j)$ é a probabilidade do i-ésimo elemento pertencer a j-ésima categoria da variável resposta do modelo;

$p_{ij} = P(Y_i = J)$ é a probabilidade do i-ésimo elemento pertencer a J-ésima categoria (categoria de referência) da variável resposta do modelo;

4. Resultados

Usando o teste de Wald, verificou-se que as variáveis rentabilidade e produtividade têm efeitos estatisticamente significantes na intensidade da concentração de renda, a um nível de significância de 5%. As empresas com rentabilidade muito alta (rentabilidade superior a 34,51) apresentaram chance de concentrar renda 2,45 (OR=1/0,408; p-valor=0,046) vezes maior que as empresas com rentabilidade muito baixa (até 13,22). Empresas com produtividade muito baixa (até 56,49) e produtividade baixa (entre 56,50 e 96,47) apresentaram chances similares de concentrar renda, mas menor que as empresas com produtividade alta (entre 96,48 e 212,68), quando comparadas com as empresas com produtividade muito alta (212,69 ou mais). Com relação a capacidade preditiva do modelo selecionado, verificou-se, em termos globais, que o modelo classificou corretamente 71,3% das empresas analisadas.

Tabela 1: Teste de significância geral do modelo logístico multinomial explicativo da concentração de renda, com todas as variáveis explicativas.

Variável	p-valor
Setor	0,119
Receita	0,858
Rentabilidade	0,001
Produtividade	<0,001

ISBN 978-85-98026-63-3

Tabela 2: Resultados do ajuste do modelo logístico multinomial explicativo da concentração de renda, com todas as variáveis explicativas.

Variável		Estimativa	P-valor*	Razão de chance
	Baixa	-4,225	<0,001	,015
	Moderada	-1,574	,008	,207
Setor	Sem informação	1,884	,191	6,579
	Primário e educação	-	-	-
	Petróleo e gás	1,223	,041	3,398
	Químico	,541	,125	1,717
	Consumo	-,531	,204	,588
	Construção	,187	,715	1,206
	Transporte	,173	,732	1,189
	Telecomunicação	-,047	,940	,954
	Financeiro	0[a]	.	1
Receita	Sem informação	-	-	-
	Receita até 9906239	,089	,802	1,093
	Receita entre 9906239 e 13897796	-,166	,671	,847
	Receita entre 13897796 e 22769100	,170	,657	1,185
	Receita superior à 22769100	0[a]	.	1
Rentabilidade	Sem informação	-2,118	,007	,120
	Rentabilidade até 13,22	-,887	,056	,412
	Rentabilidade entre 13,23 e 22,31	-,288	,566	,750
	Rentabilidade entre 22,32 e 34,50	,414	,469	1,513
	Rentabilidade superior à 34,51	0[a]	.	1
Produtividade	Sem informação	-3,926	<0,001	,020
	Produtividade até 56,49	-4,338	<0,001	,013
	Produtividade entre 56,50 e 96,47	-4,366	<0,001	,013
	Produtividade entre 96,48 e 212,68	-2,184	<0,001	,113
	Produtividade superior à 212,69	0	-	1

* Test de Wald de significância individual dos parâmetros do modelo.

Tabela 3: Teste de significância geral do modelo logístico multinomial explicativo da concentração de renda, com as variáveis explicativas selecionadas.

Variável	p-valor
Rentabilidade	0,002
Productividade	<0,001

ISBN 978-85-98026-63-3

Tabela 4: Resultados do ajuste do modelo logístico multinomial explicativo da concentração de renda, com as variáveis explicativas selecionadas.

Variável		Estimativa	P-valor*	Razão de chance
	Baixa	-4,509	<0,001	,011
	Moderada	-1,930	<0,001	,145
Rentabilidade	Sem informação	-2,293	**,003**	,101
	Rentabilidade até 13,22	-,895	**,046**	,408
	Rentabilidade entre 13,23 e 22,31	-,363	,448	,696
	Rentabilidade entre 22,32 e 34,50	,198	,719	1,218
	Rentabilidade superior à 34,51	0[a]	.	1
Produtividade	Sem informação	-3,942	**<0,001**	,019
	Produtividade até 56,49	-4,291	**<0,001**	,014
	Produtividade entre 56,50 e 96,47	-4,425	**<0,001**	,012
	Produtividade entre 96,48 e 212,68	-2,357	**<0,001**	,095
	Produtividade superior à 212,69	0	-	1

* Test de Wald de significância individual dos parâmetros do modelo.

5. Conclusões

Na análise da concentração de renda, verificou-se que as influências de produtividade, de modo que as empresas com maior produtividade têm uma chance maior de concentração de renda. É dizer que as empresas com produtividade muito elevada têm distribuição de renda muito pequena.

Também foi encontrada correlação com a concentração de renda. No entanto, apenas as empresas com elevada rentabilidade têm uma maior probabilidade de concentrar renda que as empresas com rendimentos mais baixos. Isso é um indicativo de que a participação dos lucros e depreciação do valor adicionado são independentes da rentabilidade do capital.

6. Referências bibliográficas

ANMINISTIA INTERNACIONAL, ECONOMISTAS SIN FRONTERAS, INTERMON OXFAM, Setem (2002):"Empresas más responsables para una Europa más justa". Boletin Económico de ICE, nº. 2728, 13-19, maio.

DOBSON, A. J. (1945): "An introduction to generalized linear models" / Annette J. Dobson. - 2nd ed. p. cm. - (Chapman & Hall/CRC texts in statistical science series).

ISBN 978-85-98026-63-3

Pesquisa sobre a utilização do programa R no curso de estatística da Universidade Federal do Paraná.

Bruna Davies Wundervald (UFPR) / e-mail: brunadaviesw@gmail.com

O QUE É ESTE QUESTIONÁRIO?

É um questionário sobre a utilização do programa R. Contendo onze questões simples, o questionário cobre desde o conhecimento do aluno até o seu relacionamento com o software.

PORQUE APLICAR ESTE QUESTIONÁRIO?

A fim de levantar dados sobre a experiência dos alunos da UFPR com o R. Estes dados, por sua vez, servirão como suporte para uma análise inicial dentro do Departamento de Estatística da UFPR. A análise irá compreender como está sendo o ensino do R no Curso de Estatística da Universidade. A ideia inicial é que esta pesquisa seja realizada anualmente.

COMO FOI APLICADO O QUESTIONÁRIO E COMO OS DADOS SÃO APRESENTADOS?

O questionário foi aplicado pessoalmente, em sala de aula, visto que essa seria a forma de se obter uma maior quantidade de respostas. Os dados são apresentados em um aplicativo Shiny.

ISBN 978-85-98026-63-3

QUESTIONÁRIO

1. Você sabe o que é o R?
   - Sim
   - Não
2. Você já cursou a matéria (e foi aprovado) Estatística Computacional I?
   - Sim
   - Não
3. Você sabe o que são o RStudio e o Emacs?
   - Sim, sei o que é o Emacs
   - Sim, sei o que é o RStudio
   - Sim, sei o que são ambos
   - Não
4. Você utiliza o R na sua graduação?
   - Sim
   - Não utilizo o R em nenhum momento da minha graduação
5. Se você utiliza o R na sua graduação, sua preferência é pelo RStudio ou Emacs?
   - Nenhum
   - RStudio
   - Emacs
6. Você tem facilidade de usar o R?
   - Sim
   - Não
7. Você gosta de utilizar o R?
   - Sim
   - Não
8. Você acha que o R facilita o seu aprendizado e realização de trabalhos/tarefas?
   - Sim
   - Não
9. Quais disciplinas você já cursou obrigatoriamente fizeram uso do R?

   _____

   _____
10. Você tem interesse/sente necessidade em aprimorar seus conhecimentos em R?
    - Sim
    - Não
11. O que faria você compreender melhor a lógica do R e poder fazer melhor uso do programa?

    _____

ISBN 978-85-98026-63-3

GRÁFICOS

Gráfico 1

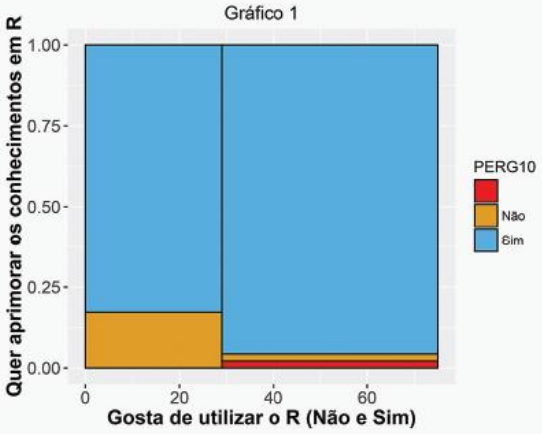


Gráfico 2

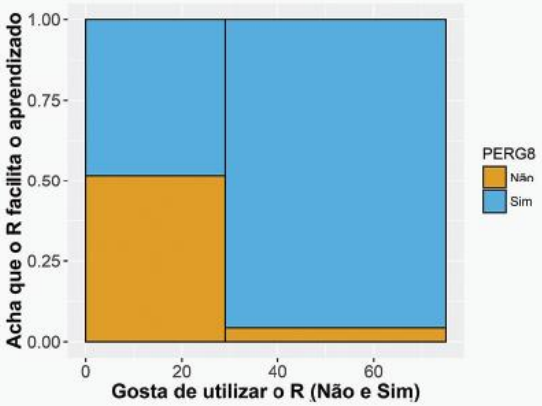


Gráfico 3

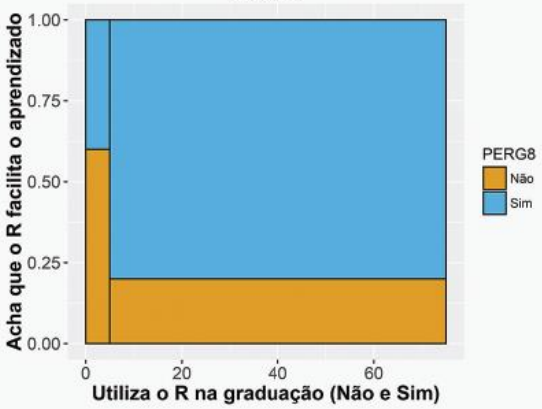
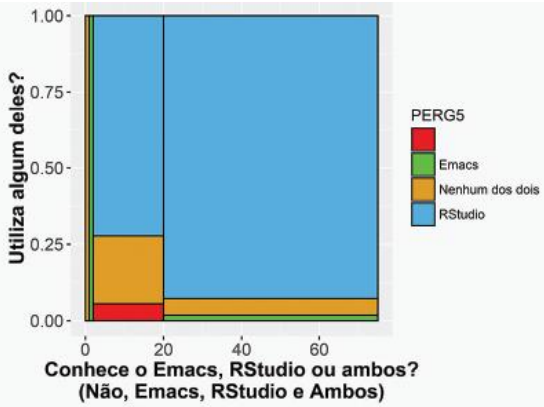


Gráfico 4

ISBN 978-85-98026-63-3



CONCLUSÃO

A partir desta pesquisa nota-se, por exemplo, que apesar de alguns alunos não gostarem de utilizar o R, quase 50% destes reconhecem que ele facilita o aprendizado, e a maior parte quer aprimorar seus conhecimentos sobre o programa. Além disso, há uma porcentagem de quase 40% de alunos que disseram não utilizar o R na graduação que reconhece o fato de ele ajudar no aprendizado. Já para a pergunta 11, a maior parte das respostas compreenderam ações como "usar mais o R em sala" e "praticar mais", o que já está sendo posto em prática e estimulado, tanto por alunos quanto por professores.

Curso de Estatística da
Universidade Federal do Paraná

ISBN 978-85-98026-63-3

Avaliação do escoamento superficial e da perda de solo sob diferentes coberturas e declividades utilizando análise de variância e modelos lineares em R.

Hugo M Rodrigues (UFF) / hugomr@id.uff.br
Gustavo M Vasques (Embrapa-Solos) / gustavo.vasques@embrapa.br
Marcelo W A Lemes (UFF) / marcelowlemes@hotmail.com

1. Introdução:

A erosão hídrica é um problema de escala global e, no Brasil, é responsável pela perda de áreas agricultáveis. Quantificar e estimar o escoamento superficial (ES) da água e a consequente perda de solo (PS) em função dos fatores condicionantes é uma necessidade para o manejo consciente e sustentável da produção agrícola, minimizando impactos negativos como a diminuição da produção e o assoreamento de corpos d'água. Para isso quantificou-se as taxas de ES e PS em diferentes condições ambientais, e foi avaliado a influência da declividade e da cobertura do solo sobre ES e PS a fim de entender quais fatores agravam os processos erosivos.

2. Material e Métodos

2.1. Coleta de amostras e filtragem em laboratório

Foram selecionadas duas encostas, com 28% e 51% de declividade na Unidade de Gestão Santo Antônio do Maratuã-RJ (Figura 1). Em cada encosta, foram instalados, um pluviômetro e duas parcelas de erosão, conforme Wischmeier e Smith (1978), uma coberta por gramínea (GR1 e GR2) e outra sem cobertura (SC1 e SC2) (Figura 2). No final de cada parcela foram instalados tambores para coletar o ES. A quantidade de chuva e o ES foram medidos em campo após cada evento de chuva e a concentração de sedimentos foi medida posteriormente em laboratório (Figura 3). A PS foi calculada multiplicando-se o ES pela concentração de sedimentos.

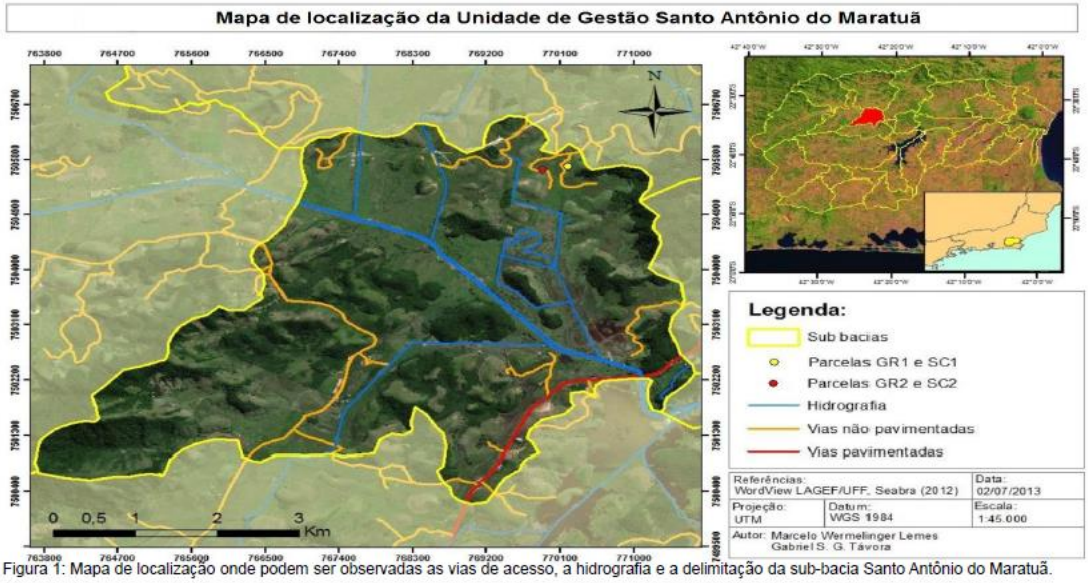


Figura 1: Mapa de localização onde podem ser observadas as vias de acesso, a hidrografia e a delimitação da sub-bacia Santo Antônio do Maratuã.

ISBN 978-85-98026-63-3

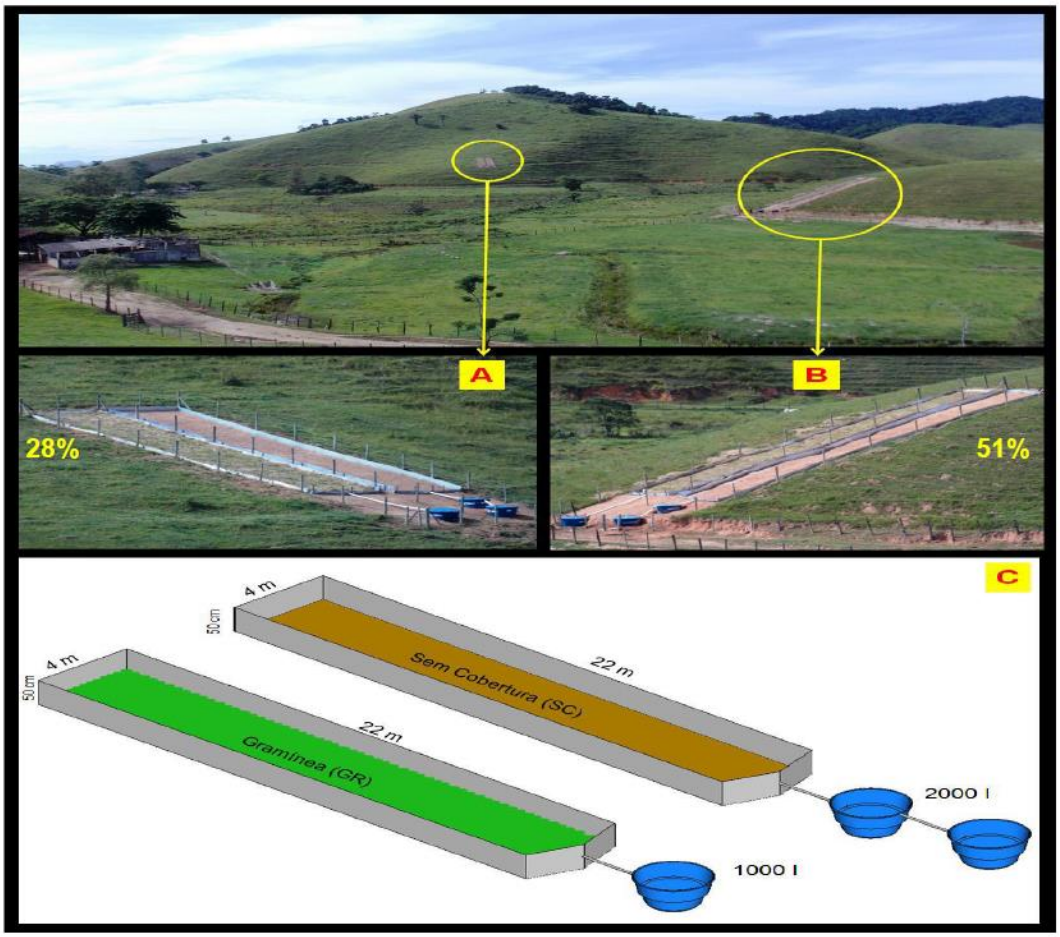


Figura 2: (A) Parcelas com cobertura de gramínea e sem cobertura da encosta 1, GR1 e SC1, com 28% de declividade; (B) Parcelas com cobertura de gramínea e sem cobertura na encosta 2, GR2 e SC2 , com 51% de declividade e (C) Esquema da estrutura da parcelas montadas em cada encosta.

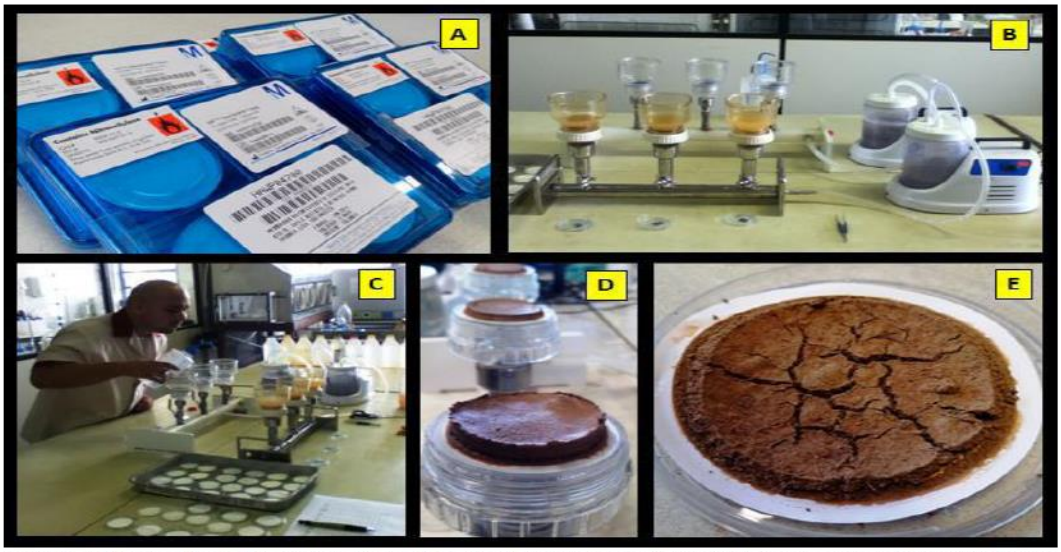


Figura 3: (A) Filtros de celulose, (B) Manifold para filtragem dos sedimentos, (C) Processo de filtragem, (D) Sedimentos retidos no filtro e (E) Sedimentos pós-estufa pronto para ser pesado.

ISBN 978-85-98026-63-3

2.2. Análise estatística

Os dados coletados foram agrupados mensalmente e transformados para logaritmo natural. A influência da declividade e cobertura sobre o ES e a PS, respectivamente, foi analisada utilizando análise de variância (ANOVA). Em seguida, foram ajustados modelos lineares para predição do ES e da PS, respectivamente, em função da quantidade de chuva e dos fatores significativos sobre essas variáveis encontrados na ANOVA. Por último, considerando a medição do ES mais simples e direta do que a da PS, se ajustou um modelo de predição da PS (de difícil medição) em função do ES (de fácil medição). A qualidade dos modelos de predição foi avaliada pelo coeficiente de determinação (R2), erro médio (EM) e raiz do erro quadrado médio (REQM). Todos os procedimentos estatísticos foram realizados no programa R.

3. Resultados

3.1. Estatística descritiva dos dados

Estatística	Chuva	ES (L)	PS (Mg ha^{-1})	Chuva	ES (L)	PS (Mg ha^{-1})
	Original			Logaritmo natural		
	Declividade de 28% / Gramínea					
Média	177,64	638,58	0,13	4,87	3,83	-5,90
Desvio padrão	125,36	997,08	0,25	0,95	5,15	3,96
Mediana	143,30	73,00	0,00	4,96	4,29	-6,54
Mínimo	15,28	0,00	0,00	2,73	-11,51	-11,51
Máximo	449,07	2737,00	0,76	6,11	7,91	-0,27
Assimetria	0,63	1,31	1,63	-0,83	-2,08	0,07
	Declividade de 28% / Sem cobertura					
Média	177,64	3381,92	1,61	4,87	5,98	-1,64
Desvio padrão	125,36	3178,15	2,15	0,95	5,72	4,08
Mediana	143,30	2122,50	0,69	4,96	7,65	-0,38
Mínimo	15,28	0,00	0,00	2,73	-11,51	-11,51
Máximo	449,07	10057,00	7,51	6,11	9,22	2,02
Assimetria	0,63	0,62	1,61	-0,83	-2,33	-1,38
	Declividade de 51% / Gramínea					
Média	178,30	483,83	0,07	4,87	3,71	-5,95
Desvio padrão	125,08	653,95	0,10	0,96	5,16	4,04
Mediana	142,86	219,50	0,01	4,96	5,00	-5,54
Mínimo	14,64	0,00	0,00	2,68	-11,51	-12,67
Máximo	448,82	2071,00	0,31	6,11	7,64	-1,18
Assimetria	0,63	1,23	1,11	-0,87	-2,05	-0,25
	Declividade de 51% / Sem cobertura					
Média	178,30	3301,92	1,61	4,87	5,92	-1,77
Desvio padrão	125,08	3097,17	2,22	0,96	5,74	4,26
Mediana	142,86	2472,50	0,64	4,96	7,81	-0,45
Mínimo	14,64	0,00	0,00	2,68	-11,51	-11,51
Máximo	448,82	9980,00	7,65	6,11	9,21	2,03
Assimetria	0,63	0,69	1,60	-0,87	-2,29	-1,33

Tabela 1: Estatística descritiva da quantidade de chuva, escoamento superficial (ES) e perda de solo (PS), por parcela.

ISBN 978-85-98026-63-3

3.2. Análises de variância

Fonte	GL	SQ	QM	F
Escoamento superficial (ES)				
Declividade	1	0,09	0,09	0,48
Cobertura	1	57,18	57,18	36,96**
Declividade x Cobertura	1	0,01	0,01	0,10
Mês	11	1286,54	116,96	
Declividade x Mês	11	2,06	0,19	
Cobertura x Mês	11	17,02	1,55	
Declividade x Cobertura x Mês	11	1,39	0,13	
Perda de solo (PS)				
Declividade	1	0,09	0,09	0,16
Cobertura	1	213,68	213,68	28,01**
Declividade x Cobertura	1	0,02	0,02	0,07
Mês	11	640,70	58,25	
Declividade x Mês	11	6454,00	0,59	
Cobertura x Mês	11	83,91	7,63	
Declividade x Cobertura x Mês	11	3547,00	0,32	

Tabela 2: GL, graus de liberdade; SQ, soma de quadrados; QM, quadrado médio ** Significativo ($p < 0.05$)

3.3. Modelos de predição de escoamento superficial (ES) e perda de solo (PS)

Equação	R^2	EM	REQM
ln(ES) = -6,81 + 2.36 x ln(chuva) + 2.38 x sem cobertura	0,89	0,50	4,67
ln(PS) = -24.92 + 3.90 x ln(chuva) + 4.22 x sem cobertura	0,87	2,23	9,90
ln(PS) = -14.90 + 1.86 x ln(ES)	0,97	0,40	4,20

Tabela 3: R^2, coeficiente de determinação; EM, erro médio; REQM, raiz do erro quadrado médio; ln, logaritmo natural

4. Discussão

Enquanto a quantidade de chuva e o tipo de cobertura influenciaram significativamente o ES e a PS, a declividade do terreno não demonstrou influência sobre essas variáveis. Ainda, as parcelas em encostas mais declivosas tiveram, em média, menor ES e menor PS. Esses resultados foram inesperados e contrariam resultados anteriores (Joshi e Tambe, 2010; Li et al., 2014). Contudo, eles podem ser explicados pela maior porosidade total do solo e maior exposição ao sol da encosta mais declivosa (Lemes, 2014), o que, por um lado, aumenta a infiltração da água da chuva e, por outro, retarda o processo de saturação do solo após a chuva, consequentemente reduzindo o ES.

A PS foi fortemente explicada pelo ES (R^2: 0,97), sendo que a cobertura não foi significativa nesse modelo (terceira equação). Por outro lado, na ausência do ES como variável independente, a cobertura tornou-se significativa (segunda equação). Ainda, a cobertura influenciou o ES, sendo que a área com capim apresentou menor ES. Isso implica na cobertura vegetal controlando a PS por meio de resistência ao ES, mas não afeta a PS em função da chuva. Mingguo et al. (2007) observaram comportamento similar na escala de bacia hidrográfica, porém, na escala de parcela, a cobertura afetou a relação entre ES e PS . A falta de interação entre chuva e cobertura nos modelos deve-se provavelmente à pouca diferença de chuva entre as encostas, enquanto a interação entre chuva e declividade não faz sentido, já que cada encosta possuía somente um pluviômetro

5. Conclusões

O escoamento superficial e a consequente perda de solo foram influenciados pela quantidade de chuva e cobertura vegetal. Por sua vez, somente o escoamento superficial explicou 97% da variância da perda de solo, havendo, portanto, possibilidade de suprimir a etapa de medição da perda de solo em laboratório. Contudo, essa possibilidade deve ser avaliada em outras áreas de estudo sob diferentes condições.

A declividade do terreno não teve influência significativa no escoamento superficial ou na perda de solo, provavelmente devido a outros fatores conflitantes, como porosidade do solo e grau de exposição ao sol da encosta. Portanto, a influência da declividade e de outros possíveis fatores sobre o ES e PS precisa ser melhor entendida na área de estudo

6. Referências

Joshi, V.U.; Tambe, D.T. Estimation of infiltration rate, run-off and sediment yield under simulated rainfall experiments in upper Pravara Basin, India: Effect of slope angle and grass-cover. Journal of Earth System Science, v. 119, pp. 763-773, 2010.

Lemes, W.M. Análise dos solos, dos processos erosivos e do comportamento hidrológico em colinas dissecadas e morros rebaixados sob diferentes usos na sub-bacia Santo Antônio do Maratuã, Silva Jardim – RJ. Dissertação (Mestrado em Geografia) – Universidade Federal Fluminense, Niterói. 2011.

Li, X.; Niu, J.; Xie, B. The effect of leaf litter cover on surface runoff and soil erosion in Northern China, PLoS ONE, v. 9, e107789, 2014.

Mingguo, Z.; Qiangguo, C.; Hao, C. Effect of vegetation on runoff-sediment yield relationship at different spatial scales in hilly areas of the Loess Plateau, North China. Acta Ecologica Sinica, v. 27, p. 3572-3581, 2007.

Wischmeier, W.H.; Smith, D.D. Predicting rainfall erosion losses: a guide to conservation planning. Washington, DC: United States Department of Agriculture, 1978. (Agriculture Handbook, v. 537).

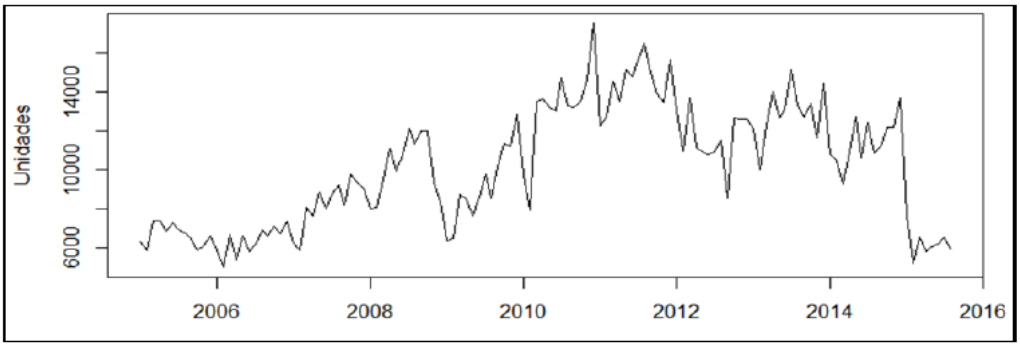
ISBN 978-85-98026-63-3

Utilização do R para previsão de vendas de caminhões no Brasil através do método bagging arima.

Luiz Campos de Sá Lucas (MC 15 Consultoria) / e-mail: luizsa.lucas@mc15.com.br
Felipe Lobo Umbelino de Souza (PUC-Rio) / e-mail: felipelobodesouza@yahoo.com.br

Introdução: O processo da indústria de caminhões envolve setores com grande expressão no país: transporte, logística e comércio. Destaca-se a predominância do modal rodoviário que em Janeiro/2016 representava cerca de 61,10% do transporte de carga no país (CNT, 2016). Assim, o setor de transportes, em especial o segmento de caminhões no País, possui uma dinâmica de acordo com a evolução da conjuntura econômica. A análise recente de vendas de caminhões demostra uma tendência de redução no número de licenciamentos com o agravamento da recessão econômica no país, conforme indica a Figura 1. Dessa forma, a dificuldade de realizar previsões de vendas nesse cenário inibe os investimentos das empresas, que, assim, optam por postergá-los.

Figura 1: Vendas de Caminhões no Brasil (Janeiro/2005 – Agosto/2015)



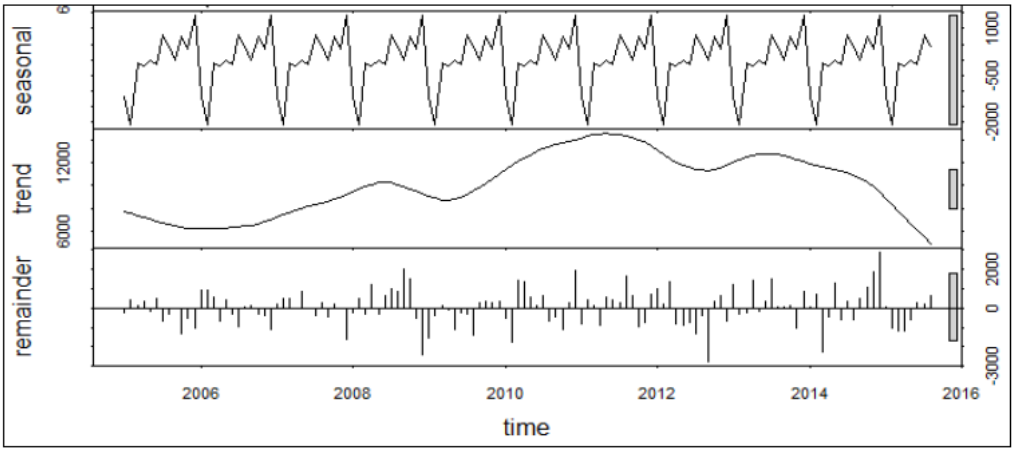
Fonte: ANFAVEA, 2016.

Objetivo: Este trabalho busca utilizar a série de vendas de Caminhões no Brasil, divulgada pela ANFAVEA (2016), com o objetivo de aplicar uma metodologia de previsão de séries temporais combinando técnicas de simulação *Moving Block Bootstrap*, decomposição STL e o método ARIMA, visando obter previsões mais eficientes em relação aos métodos tradicionais da literatura. Para análise da série, foram usados os dados de janeiro/2005 até agosto/2015. Para o processo de validação e comparação dos resultados através do MAPE (*Mean Absolute Percentage Error - out-of-sample*) foram utilizados as últimas sete observações disponíveis à época da realização do estudo (setembro/2015 a março/2016).

Método: A série temporal de vendas de Caminhões é decomposta utilizando o método STL, resultando em três partes: Componente Sazonal, Tendência e "*Remainder*" (Figura 2).

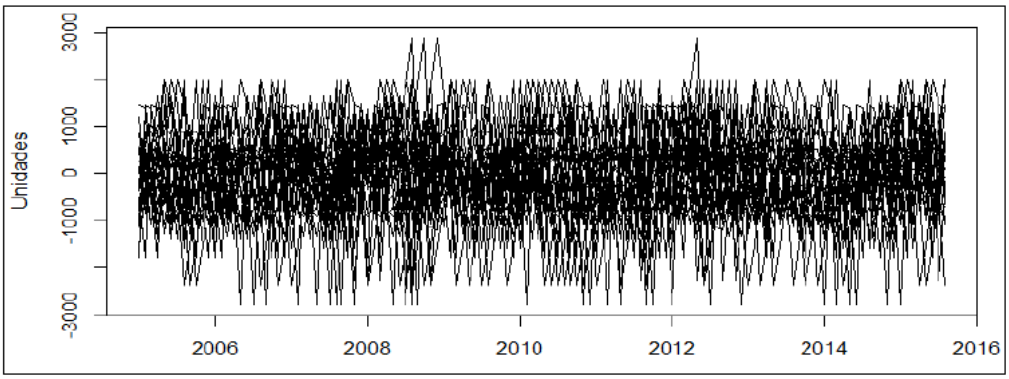
ISBN 978-85-98026-63-3

Figura 2: Decomposição STL da série de vendas de Caminhões



Breiman (1996) destaca que o número de replicações consideradas necessárias varia de 25 a 50. Assim, à série "*Remainder*" é aplicado o método *Moving Block Bootstrap*, com 30 replicações, sendo geradas 30 séries pelo método, conforme a Figura 3.

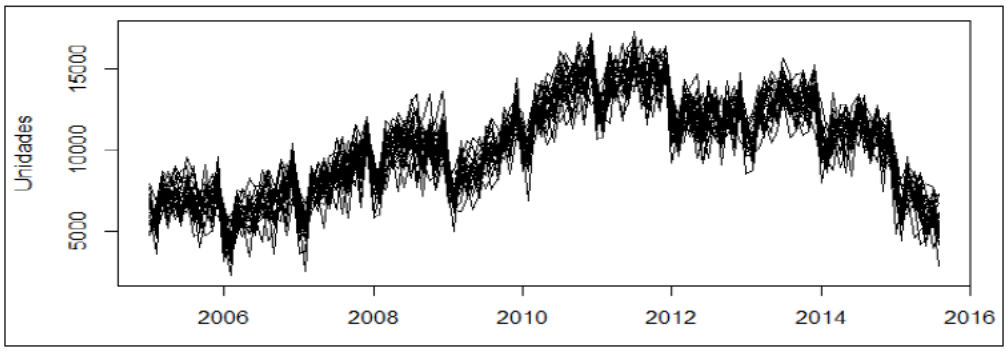
Figura 3:Geração de séries a partir da simulação *Moving Block Bootstrap*



A cada uma das replicações geradas são adicionados os componentes sazonais e de tendência da série original obtidos pela decomposição STL, resultando assim em 30 novas séries sintéticas, como demonstrado na Figura 4.

ISBN 978-85-98026-63-3

Figura 4: 30 séries sintéticas geradas pelo método



Resultados: Para este estudo, todo o método descrito foi replicado 100 vezes no R. Utilizando a média e mediana dos resultados de previsão gerados pela simulação, pode-se realizar uma comparação entre as previsões geradas através dos métodos de *Bagging* ARIMA e SARIMA tradicional para os sete últimos meses observados. Analisando pela métrica MAPE (*out-of-sample*), a abordagem que considera a agregação de resultados por amostras geradas via *Bagging* ARIMA apresenta melhores resultados que a aplicação direta do método SARIMA simples, e até do modelo *Holt Winters*, conforme apresentado nas Figuras 5 e 6, e na Tabela 1.

Figura 5: Comparação entre as previsões geradas pelo método *Bagging* Arima e as vendas reais no período set/15-mar/16

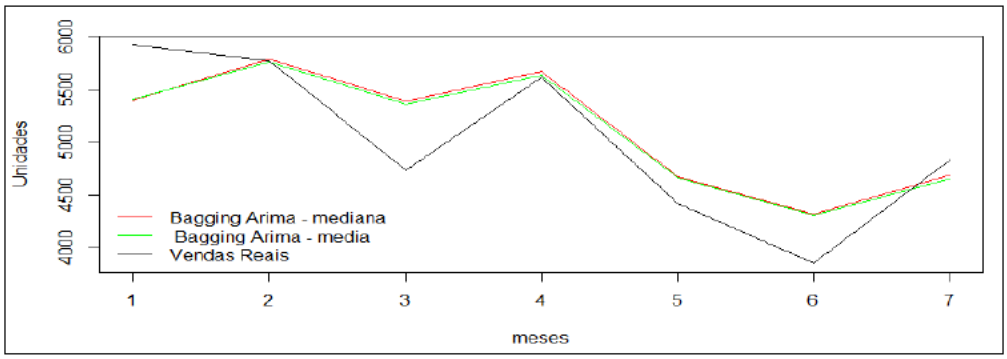
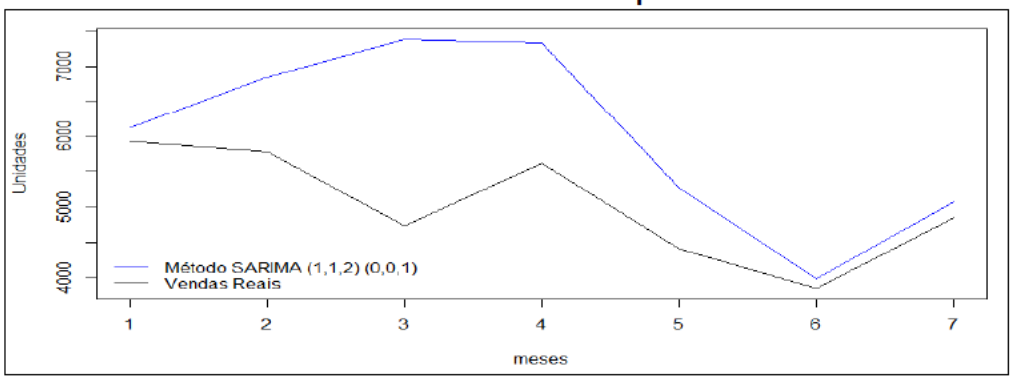


Figura 6: Comparação entre a previsão gerada pelo método SARIMA tradicional e as vendas reais no período set/15-mar/16



ISBN 978-85-98026-63-3

Tabela 1 – Resultados dos métodos

Métodos		MAPE (*out of sample*)
Holt Winters		13,16%
SARIMA (1,1,2) (0,0,1)		19,43%
Bagging Arima	Média	6,55%
	Mediana	6,21%

Conclusão: Verifica-se, pela Tabela 1, que métodos tradicionais podem muitas vezes serem aperfeiçoados com o uso de técnicas de simulação. Trabalhos desenvolvidos como Dantas e Oliveira (2014; 2015) já haviam demonstrado que este método pode produzir previsões mais precisas. Como contribuição, este trabalho buscou utilizar uma metodologia para a previsão de séries temporais no contexto das vendas de caminhões no Brasil, visando oferecer informações mais acuradas ao mercado automobilístico, nesse período de instabilidade econômica.

Referências Bibliográficas

ANFAVEA(2016).Séries Temporais .(http://www.anfavea.com.br/tabelasnovo.html); **Breiman, L (1996)**. *Bagging predictors*. Machine Learning; **CNT**. Boletim estatístico. Janeiro/2016;**Dantas, Oliveira (2014).**. Previsão de Velocidade de Vento: Uma abordagem utilizando *Bagging Holt Winters* com Decomposição STL. ; **Dantas; Oliveira (2015)** *Bagging* Arima Para Previsão De Demanda De Transporte Aéreo.

ISBN 978-85-98026-63-3

Pricing a Self-funded Health Plan Applying Generalized Linear Models Using R.

Helano Silva Eugênio de Souza (MSc, IBMEC) helanosouza@uol.com.br
Luiz Carlos da Silva Leão (BSc, UFF) luizcarlosleao@id.uff.br

In the Brazilian market of Health Insurance Plans, self-funded operators are those in which the company itself or other organization establishes and manages with non-profit purpose, the health care program for its employees and families (RN 279 ANS). We can mention CASSI plan, from Banco do Brasil Company, AMS Plan, from Petrobras Company, among others.

These operators have a tradition of applying simple methods, from the statistical and actuarial point of view, in order to pricing their health plans. Two of them are: a fixed percentage of the participant salary or a co-participation single contribution table. Sometimes both of them.

So, given that the medical utilization of health plans traditionally follows a probability distribution of the exponential family (Jones, 2010), which was confirmed with the medical utilization experience used for this poster, it was possible to apply the Generalized Linear Models to obtain the values of tariff that should be charged to participants of a self-funded health plan, in order to obtaining a new way of pricing it.

This methodology considers the profile of each participant (individual risk), similar to what is done today in the open health insurance market, considering variables as age, gender, salary and others.

A 300.069 lives medical plan database of a self-funded health plan, through the period from 2007 to 2010, both including, with outpatient (lower) and inpatient (higher) costs was considered in this paper. A project in the statistical software R was designed consolidating it in a one-group mass.

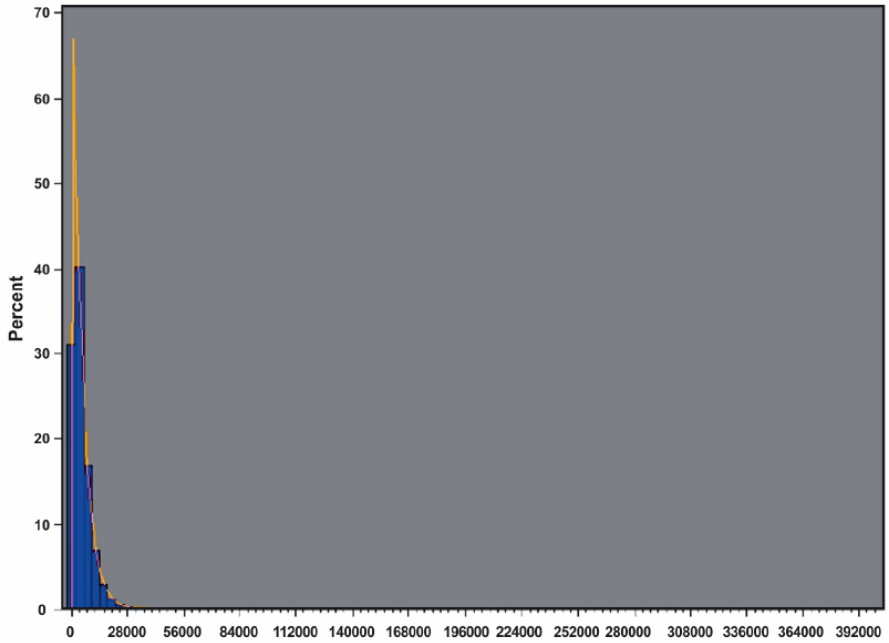
ISBN 978-85-98026-63-3

Additionally, it was also modeled, by GLM, the number of days hospitalized of this same health plan participants. Finally, by manipulating the GLM´s results statistically obtained, the model proposed was able to generate tables containing tariffs that can be charged to the participants, covering all kind of average risk (expected medical utilization).

In order to assure the financial time position consistence of the medical cost considered, the 2007, 2008 and 2009 expenses were monetarily updated to 2010 (the last year considered) by the Hospital Costs' Index of Variation - HCIV[1] observed in the same plan experience.

The paper inferred that the Outpatient and Inpatient expenses both follow a Gamma distribution, what was critical to applying GLM (Generalized Linear Model).

CHART 1 - HISTOGRAM OF OUTPATIENT EXPENSES



The following is one of the 8 Tables of Tariffs generated by GLM for a self-funded health plan (the header of the table refers to salary ranges: 1 – Lower than R$ 882,14; 2 - Up to R$ 882,14 and Lower than R$ 1.628,57; 3 - Up to R$ 1.628,57 and Lower than R$ 3.257,14; 4 - Up to R$ 3.257,14 and Lower than R$ 6.075,07; 5 - Up to R$ 6.075,07 and Lower than R$ 12.150,14; 6 - Greater than R$ 12.150,14):

ISBN 978-85-98026-63-3

TABLE 1 – TARIFF'S TABLE GLM MODELED / GENDER: FEMALE / USER TYPE: HOLDER

	1	2	3	4	5	6
0 - 18	R$ 349,13	R$ 282,67	R$ 253,27	R$ 234,32	R$ 222,79	R$ 225,27
19 - 23	R$ 372,27	R$ 298,48	R$ 265,82	R$ 244,28	R$ 230,90	R$ 233,04
24 - 28	R$ 294,26	R$ 240,20	R$ 216,29	R$ 201,22	R$ 192,23	R$ 194,65
29 - 33	R$ 379,70	R$ 306,71	R$ 274,42	R$ 253,49	R$ 240,69	R$ 243,26
34 - 38	R$ 412,00	R$ 336,27	R$ 302,79	R$ 281,66	R$ 269,07	R$ 272,45
39 - 43	R$ 492,85	R$ 400,95	R$ 360,30	R$ 334,43	R$ 318,88	R$ 322,69
44 - 48	R$ 586,74	R$ 476,49	R$ 427,73	R$ 396,55	R$ 377,72	R$ 382,13
49 - 53	R$ 692,42	R$ 560,46	R$ 502,09	R$ 464,45	R$ 441,53	R$ 446,41
54 - 58	R$ 838,77	R$ 673,54	R$ 600,42	R$ 552,37	R$ 522,59	R$ 527,60
> = 59	R$ 1.398,66	R$ 1.102,06	R$ 970,69	R$ 880,97	R$ 823,42	R$ 828,18

The variable "number of days hospitalized" is an important parameter for health plans managers, especially in regard to actuarial provisions, because more the number of days hospitalized, more tends to be the medical costs.

[1] *The Hospital Costs' Index of Variation – HCIV is obtained by calculating the variation of amount of per capita expenses among two consecutive years. Internal expertise of the data provider suggests that considering a defined basket of medical items (as done in a common price index measurement), three key elements compose the HCIV: Current basket prices variations, current basket demand variation and basket itens changings. This way, we can enunciate medical inflation, new technologies, plan rules changings, aging factor, governmental regulation, among others, as components of those key elements.*

ISBN 978-85-98026-63-3

The following is an example of how many days is expected that an individual, "male gender", age group "Greater than 58 years", "Holder", with an income above R$ 12,150.14 and an "Operative Employee", remain hospitalized by the GLM results:

TABLE 2 - RESULTS OF TESTING 1 (ONE) INDIVIDUAL MODEL OF TIME REMAIN IN HOSPITAL

Expected Number of Days Hospitalized	12,97 days

As shown above, it is expected that an individual, with the characteristics described before, remains approximately 13 days hospitalized. This model can be used for a large number of experiments with a lot of implications. Generalized Linear Modeling fitted very well on the medical costs database analyzed.

Using the statistical software R, it was possible to estimate a pricing model that considered different characteristics between the plan participants, according to current practices in the insurance market.

Unlike the common practices of self-funded health insurance market, which is applying a percentage of participant´s salary or collecting values of a fixed table, by this model, each participant, with own his or her individual characteristics (age, salary, gender and type) will has his or her specific tariff, considering the degree of risk of the plan (levels of medical utilization) similar to what is done in the opened health insurance market.

From the actuarial provisions point of view, the model proposed in this paper differs from current practices in self-funded health plans, because it tends to collect a different amount of premium, which is statistically adherent to the medical utilization experience of the plan.

ISBN 978-85-98026-63-3

```
###### R SCRIPT ######

## 1º Seminário Internacional de Estatística com R ##

### Pricing a Self-funded Health
Plan Applying Generalized Linear Models Using R##

## SOUZA & LEÃO ##

#####################################################

amb<-read.csv("ambulatorial.csv",header=TRUE,sep=",",stringsAsFactors=T)

int<-read.table("INT.txt",header=TRUE,sep=",")

hist(amb$SUM_of_PRMF)

hist(int$SUM_of_GRANDE.RISCO)

hist(int$SUM_of_NUM_DIAS_INT)

plot(density(amb$SUM_of_PRMF),main="Densidade Pequeno Risco")

plot(density(int$SUM_of_GRANDE.RISCO),main="Densidade Grande Risco")

plot(density(int$SUM_of_NUM_DIAS_INT),main="Densidade Numero de Dias Interna-
cao")

amb2<-subset(amb, amb$SUM_of_PRMF > 0) #amostra Despesa Ambulatorial onde Pe-
queno Risco maior que Zero pois no GLM Gamma foi informado ter valores negati-
vos

int2<-subset(int, int$SUM_of_GRANDE.RISCO > 0) #amostra Despesa Internacao onde
Grande Risco maior que Zero pois no GLM Gamma foi informado ter valores negati-
vos

int3<-subset(int2,int2$SUM_of_NUM_DIAS_INT > 0)

amb3<-data.frame(amb2$NUM_ID_USUARIO,amb2$IDENTIFICACAO,amb2$TIPO_USUARIO,amb2
$FAIXA_SAL,amb2$SEXO,amb2$Faixa.Etaria,amb2$SUM_of_PRMF,amb2$COUNT_of_PRMF)

int4<-data.frame(int3$NUM_ID_USUARIO,int3$IDENTIFICACAO,int3$TIPO_USUARIO,int3
$FAIXA_SAL,int3$SEXO,int3$Faixa.Etaria,int3$SUM_of_GRANDE.RISCO,int3
$SUM_of_NUM_DIAS_INT)

amb3$amb2.NUM_ID_USUARIO<-as.character(amb3$amb2.NUM_ID_USUARIO)

is.character(amb3$amb2.NUM_ID_USUARIO)

amb3$amb2.IDENTIFICACAO<-as.factor(amb3$amb2.IDENTIFICACAO)

is.factor(amb3$amb2.IDENTIFICACAO)

amb3$amb2.TIPO_USUARIO<-as.factor(amb3$amb2.TIPO_USUARIO)

is.factor(amb3$amb2.TIPO_USUARIO)

amb3$amb2.FAIXA_SAL<-as.factor(amb3$amb2.FAIXA_SAL)

is.factor(amb3$amb2.FAIXA_SAL)

amb3$amb2.SEXO<-as.factor(amb3$amb2.SEXO)

is.factor(amb3$amb2.SEXO)

amb3$amb2.Faixa.Etaria<-as.factor(amb3$amb2.Faixa.Etaria)

is.factor(amb3$amb2.Faixa.Etaria)

int4$int3.NUM_ID_USUARIO<-as.character(int4$int3.NUM_ID_USUARIO)

is.character(int4$int3.NUM_ID_USUARIO)

int4$int3.IDENTIFICACAO<-as.factor(int4$int3.IDENTIFICACAO)

is.factor(int4$int3.IDENTIFICACAO)

int4$int3.TIPO_USUARIO<-as.factor(int4$int3.TIPO_USUARIO)

is.factor(int4$int3.TIPO_USUARIO)

int4$int3.FAIXA_SAL<-as.factor(int4$int3.FAIXA_SAL)

is.factor(int4$int3.FAIXA_SAL)

int4$int3.SEXO<-as.factor(int4$int3.SEXO)

is.factor(int4$int3.SEXO)

int4$int3.Faixa.Etaria<-as.factor(int4$int3.Faixa.Etaria)

is.factor(int4$int3.Faixa.Etaria)
```

ISBN 978-85-98026-63-3

```
library(MASS)

glmsemident<-glm(amb3$amb2.SUM_of_PRMF ~ amb3$amb2.TIPO_USUARIO + amb3
$amb2.FAIXA_SAL + amb3$amb2.Faixa.Etaria + amb3$amb2.SEXO, family=Gamma(link =
"log"))

summary(glmsemident)

glmcomident<-glm(amb3$amb2.SUM_of_PRMF ~ amb3$amb2.TIPO_USUARIO + amb3
$amb2.IDENTIFICACAO + amb3$amb2.FAIXA_SAL + amb3$amb2.Faixa.Etaria + amb3
$amb2.SEXO, family=Gamma(link = "log"))

gamma.shape(glmcomident,verbose=T)

summary(glmcomident, dispersion=gamma.dispersion(glmcomident))

glmcomidentlinkinverse<-glm(amb3$amb2.SUM_of_PRMF ~ amb3$amb2.TIPO_USUARIO +
amb3$amb2.IDENTIFICACAO + amb3$amb2.FAIXA_SAL + amb3$amb2.Faixa.Etaria + amb3
$amb2.SEXO, family=Gamma(link = "inverse"))

gamma.shape(glmcomidentlinkinverse,verbose=T)

summary(glmcomidentlinkinverse, dispersion=gamma.dispersion
(glmcomidentlinkinverse))

glm_int_link_log<-glm(int4$int3.SUM_of_GRANDE.RISCO ~ int4$int3.TIPO_USUARIO +
int4$int3.IDENTIFICACAO + int4$int3.FAIXA_SAL + int4$int3.Faixa.Etaria + int4
$int3.SEXO, family=Gamma(link = "log"))

gamma.shape(glm_int_link_log,verbose=T)

summary(glm_int_link_log, dispersion=gamma.dispersion(glm_int_link_log))

glm_dias_link_log<-glm(int4$int3.SUM_of_NUM_DIAS_INT ~ int4$int3.TIPO_USUARIO +
int4$int3.IDENTIFICACAO + int4$int3.FAIXA_SAL + int4$int3.Faixa.Etaria + int4
$int3.SEXO, family=Gamma(link = "log"))

gamma.shape(glm_dias_link_log,verbose=T)

summary(glm_dias_link_log, dispersion=gamma.dispersion(glm_dias_link_log))

med<-merge
(amb3,int4,by.x="amb2.NUM_ID_USUARIO",by.y="int3.NUM_ID_USUARIO",all=TRUE)

med[is.na(med)] <- 0

summary(med$int3.SUM_of_GRANDE.RISCO)

summary(int4$int3.SUM_of_GRANDE.RISCO)

summary(med$int3.SUM_of_NUM_DIAS_INT)

summary(int4$int3.SUM_of_NUM_DIAS_INT)
```

ISBN 978-85-98026-63-3

```
pr<-sum(med$amb2.SUM_of_PRMF)
gr<-sum(med$int3.SUM_of_GRANDE.RISCO)
pr2<-sum(med$amb2.SUM_of_PRMF/3.5)
gr2<-sum(med$int3.SUM_of_GRANDE.RISCO/3.5)
med$total<-med$int3.SUM_of_GRANDE.RISCO+med$amb2.SUM_of_PRMF
med$totalN<-med$int3.SUM_of_NUM_DIAS_INT+med$amb2.COUNT_of_PRMF
Ambulatorial_X_Fitado<-rgamma(300000,1.277270381,0.003038649)
hist(Ambulatorial_X_Fitado)
Internacao_X_Fitado<-rgamma(100000,0.612003403,0.002582486)
hist(Internacao_X_Fitado)
Ambulatorial_N_Fitado<-rgamma(300000,1.277270381,0.003038649)
hist(Ambulatorial_N_Fitado)
Internacao_N_Fitado<-rgamma(100000,0.612003403,0.002582486)
hist(Internacao_N_Fitado)
glm_total_link_log<-glm(med$total ~ med$amb2.TIPO_USUARIO +
med$amb2.IDENTIFICACAO + med$amb2.FAIXA_SAL + med$amb2.Faixa.Etaria +
med$amb2.SEXO, family=Gamma(link = "log"))
gamma.shape(glm_total_link_log,verbose=T)
summary(glm_total_link_log, dispersion=gamma.dispersion(glm_total_link_log))
glm_total_N_link_log<-glm(med$totalN ~ med$amb2.TIPO_USUARIO +
med$amb2.IDENTIFICACAO + med$amb2.FAIXA_SAL + med$amb2.Faixa.Etaria +
med$amb2.SEXO, family=Gamma(link = "log"))
gamma.shape(glm_total_N_link_log,verbose=T)
summary(glm_total_N_link_log, dispersion=gamma.dispersion(glm_total_link_log))
total_Fitado<-rgamma(100000,0.643222996,0.001438805)
hist(total_Fitado)
totalN_Fitado<-rgamma(100000,1.565102552,0.003782347)
hist(totalN_Fitado)
```

End of the R Script SOUZA & LEÃO

ISBN 978-85-98026-63-3

Critérios de seleção baseados no poder predito: um estudo de simulação interagindo o R com o openbugs.

Bruno Leonardo dos Santos Nobrega UFF/ brunoleonardo.nave@gmail.com
Jony Arrais Pinto Junior UFF/ jarrais@est.uff.br

Introdução

- Muitos autores têm discutido a difícil tarefa de selecionar modelos tanto do ponto de vista frequentista quanto Bayesiano. Dentre os métodos mais utilizados pode-se citar o fator de Bayes e os critérios AIC,BIC,DIC,entre outros.
- Os modelos utilizados, foram os de Cox log-Gaussiano devido a sua importância na modelagem de padrões de pontos(Møller et al. 1998), os modelos são (Benes et al. 2002):

$$\begin{aligned} X &\sim PP(\Lambda(\cdot)), \\ \Lambda(\mathbf{s}) &= r(\mathbf{s})\lambda(\mathbf{s}), \forall \mathbf{s} \in S, \\ \log\lambda(\mathbf{s}) &= w(\mathbf{s}) + \boldsymbol{\beta}'\mathbf{z}(\mathbf{s}), \\ w(\cdot) &\sim PG(\mu, \tau, \rho_\phi), \end{aligned}$$

em que $r(\mathbf{s})$ é uma constante conhecida, que pode representar a densidade populacional, por exemplo, $\mathbf{z}(\mathbf{s})$ são as covárias associadas ao espaço, β é o vetor de incrementos das covariáveis na intensidade do processo e $w(\cdot)$ é um processo Gaussiano com média μ, precisão τ e função de correlação espacial ρ_ϕ.

- O enfoque deste trabalho é totalmente Bayesiano e os critérios de comparação trabalhados são aqueles que avaliam o poder preditivo de cada modelo para uma região não observada.

Objetivos

Este trabalho tem por objetivo entender,comparar e propor critérios paraa seleção de modelos para padrões de pontos com base em métodos preditivos, isto é, o modelo M preferível será aquele que consegue prever melhor observações futuras geradas de um mesmo processo que os dados originais.

E, além de estudar critérios básicos propostos inicialmente, estudar como se comportam critérios ja conhecidos na literatura com WAIC, LOO e K-fold (Vehtari and Gelman 2014), no contexto de padrões de pontos.

ISBN 978-85-98026-63-3

Metodologia e Implementação

- Primeiramente, era simulado um conjunto de dados a partir do medelo descrito acima, então com esse conjunto de dados ajustava-se modelos para eles, um que de fato os dados foram gerados e outros diferentes.
- Para obter as estimativas bayesianas foi utilizado no R o pacote "R2OpenBUGS", que permite fazer a integração direta entre os dois softwares. O pacote possui a função "bugs"que executa a compilação e gera as cadeias a posterioris para os parâmetros solicitados. O uso esta sendo mostrado abaixo.

```
require(R2OpenBUGS)
modelo = "umacov.txt"
dado = list(N = 99, phi.w = 4.248685,
            x = centros[,1], y = centros[,2],
            dados = X, r = r[-1], z = z[-1])
inits = list(
  list(media.w = 0, tau.w = 1,
       beta = 2, w = w),
  list(media.w = 0.2, tau.w = 1.2,
       beta = 1.8, w = w) )

parametros = c("media.w", "tau.w", "beta","w")

resul <- bugs(model.file = modelo,data = dado,
              inits = inits, n.burnin = 1000,
              parameters.to.save = parametros,
              n.chains = 2, n.iter = 10000)
```

- A partir das estimativas obtidas com cada ajuste, utilizava-se os critérios em cada um dos ajustes e verficava-se qual tinha sido o melhor de acordo com cada critério.

Critérios de Comparação Básicos

Os primeiros critérios avaliados foram o desvio quadrático médio das previsões com as reais contagens (C_1), a quantidade de contagens contidas nos seus respectivos intervalos de credibilidade (C_2), a amplitude dos intervalos de credibilidade (C_3) e os desvios quadráticos ponderados pela amplitude dos intervalos (C_4). Os quatro critérios citados acima são definidos por:

ISBN 978-85-98026-63-3

- $C_1 = \frac{\sum_{i=1}^{N}(\hat{n}_{S_i} - n_{S_i})^2}{N}$,
- $C_2 = \#(n_{S_i} \subset IC_{n_{S_i}})$,
- $C_3 = \frac{\sum_{i=1}^{N} \text{amplitude do } IC_{n_{S_i}}}{N}$,
- $C_4 = \frac{\sum_{i=1}^{N} \text{amplitude do } IC_{n_{S_i}}(\hat{n}_{S_i} - n_{S_i})^2}{N}$,

em que $\hat{n}_{S_i}$ é a estimativa para as contagens na região i, n_{S_i} é a contagem observada na região i, $IC_{n_{S_i}}$ é o intervalo de credibilidade para n_{S_i} e N é o número total de regiões.

Outros Critérios

▸ WAIC

Suponha $y_1, ..., y_n$ variáveis independentes que são modeladas por um vetor de parâmetros θ e tomando como notação para a distribuição a posteriori $p_{\text{post}}(\theta)$ e para distribuição preditiva $p_{\text{post}}(\tilde{y})$. O critério WAIC será dado por (Vehtari and Gelman 2014):

$$\text{WAIC} = -2\ \widehat{\text{elpd}}_{\text{waic}},$$

em que o elpd é uma medida para avaliar a precisão preditiva, sendo assim, para calcular o $\widehat{\text{elpd}}_{\text{waic}}$ pode-se utilizar a seguinte expressão:

$$\widehat{\text{elpd}}_{\text{waic}} = \widehat{\text{lpd}} - \widehat{p}_{\text{waic}},$$

$$\text{lpd} = \sum_{i=1}^{n} \log p_{\text{post}}(y_i),$$

$$p_{\text{waic}} = \sum_{i=1}^{n} \text{var}_{\text{post}}(\log p(y_i|\theta)).$$

▸ LOO

Diferentemente do primeiro critério, neste a avaliação será feita de modo que, ao estudarmos o resultado da previsão para uma variável vamos supor que ela não participou do processo de estimação. Porém , o modelo será estimado apenas uma vez e para usar essa suposição cada região terá um peso. Para calcular esse critério temos (Vehtari and Gelman 2014):

$$\widehat{\text{elpd}}_{\text{is-loo}} = \sum_{i=1}^{n} \widehat{\text{elpd}}_{\text{is-loo}_i} = \sum_{i=1}^{n} \log\left(\frac{\sum_{t=1}^{T} p(y_i|\theta^t)\tilde{w}_t}{\sum_{t=1}^{T} \tilde{w}_t}\right),$$

em que o peso $\tilde{w}$ para a variável t vai ser calculado como:

$$w_t = \frac{1}{p(y_i|\theta^t)},$$

ISBN 978-85-98026-63-3

Contudo com os pesos sendo definidos dessa forma, eles causam instabilidades pois podem gerar valores muito grandes ou até infinitos, por isso os pesos serão dados por:

$$\tilde{w}_t = \min(w_t, \sqrt{T}\overline{w}).$$

Outra medida importante é o número efetivo de parâmetros que pela relação utilizada acima será escrito como:

$$\widehat{p}_{\text{is-loo}} = \widehat{\text{lpd}}_{\text{is-loo}} - \widehat{\text{elpd}}_{\text{is-loo}},$$

com $\widehat{\text{lpd}}_{\text{is-loo}}$ podendo ser calculado como:

$$\widehat{\text{lpd}}_{\text{is-loo}} = \sum_{i=1}^{n} \log p_{\text{post}(-i)}(y_i).$$

- K-fold

Para este caso, os dados serão divididos em partições e semelhante ao LOO, quando estudarmos determinada variável vamos supor que ela não participou do processo de estimação,aqui, porém, toda a partição a qual ela pertence não participará do processo de estimação. Com a recomendação de que se use poucas partições, sendo 5 ou 10 os valores mais comuns. Diferentemente dos outros dois critérios, neste caso o processo de estimação ocorrerá tantas vezes quanto o número de partições. Dado todos esses fatores o critério é calculado por (Vehtari and Gelman 2014):

$$\widehat{\text{elpd}}_{\text{xval}} = \sum_{i=1}^{n} \widehat{\text{ldp}}_i,$$

$$\widehat{\text{ldp}}_i = \log\left(\frac{1}{T}\sum_{t=1}^{T} p(y_i|\theta^{k,t})\right),$$

Em que, $\theta^{k,t}$ representa o t-ésimo valor da amostra a posteriori para o parâmentro, estimando sem utilizar a partição k a qual i pertence. Nesta situação, para calcular o número efetivo de parâmetros haverá a mesma relação do critério anterior:

$$\widehat{p}_{\text{xval}} = \widehat{\text{lpd}} - \widehat{\text{elpd}}_{\text{xval}}.$$

ISBN 978-85-98026-63-3

Resultados

Considerando o Processo de Cox log-Gaussiano definido no inicio e os seguintes modelos:

- Modelo Correto (MC): Modelo com 1 covariável ($z(s)$), em que $z(s)$ é a covariável utilizada na geração dos dados.
- Modelo Errado 1 (ME1): Modelo com 1 covariável ($y(s)$), uma covariável qualquer associada ao espaço.
- Modelo Errado 2 (ME2): Modelo com 2 covariáveis ($z(s)$) e ($y(s)$).

os resultados para cada critério básico foi:

	MC	ME1	ME2
C_1	**3.885,48**	35.730,07	3.982,94
C_2	91	**95**	92
C_3	**184,42**	658,80	188,99
C_4	**5.746.885**	93.274.832	6.333.291

Tabela 1: *Resultados dos critérios utilizando a mediana como estimativa.*

	MC	ME1	ME2
C_1	**4.763,54**	46.510,31	4.790,71
C_2	91	**95**	92
C_3	**184,42**	658,8007	188,99
C_4	**8.072.461**	169.923.040	8.686.339

Tabela 2: *Resultados dos critérios utilizando a média como estimativa.*

E quando os ajustes foram avaliados pelos outros critérios mais robustos os resultados foram:

	MC	ME1	ME2
WAIC	**-17,37**	33,58	369,34
p_{waic}	**33,70**	6.322,05	227,81
LOO	**-6,60**	-59,75	-805,40
$p_{is\text{-}loo}$	**48,99**	99,44	848,54
K-fold	**10,35**	8,36	-24,33
p_{xval}	**32,04**	36,15	67,48

Tabela 3: *Resultados dos critérios WAIC, LOO e K-fold para MC, ME1 e ME2*

ISBN 978-85-98026-63-3

Conclusões e Trabalhos Futuros

- Na maioria dos casos os critérios apontaram para o modelo que de fato tinha gerado os dados, o único critério que não conseguiu apontar corretamente foi o C_2, dos critérios básicos, o que faz sentido se olharmos também para o critério C_3 que aponta para uma grande amplitude dos IC's nos outros modelos, implica numa grande incerteza, mas conteve o valor verdadeiro do parâmetro em mais ocasiões.
- Quando compara-se MC com ME2 pelos critérios básicos a diferença não é tão alta, lembrando que para ME2 apenas foi adicionada uma covariável espacial que não era relevante para o modelo. Quando olhamos para os outros critérios a diferença ja é mais perceptível devido a penalização ao número de parâmetros.
- Utilizar a integração entre os dois programas demonstrou vantagem ao ponto que não era necessário ficar exportando e importando dados de um software para o outro. Houve uma grande economia de tempo.
- A partir de agora o objetivo será pensar em combinar os critérios estudados ou ainda desenvolver novos critérios com base no poder preditivo.
- Outro ponto ainda é estudar como diminuir os problemas observados durante a execução do trabalho com relação ao parâmetro de alcance do modelo.

Referências

[1] Benes, V and Bodlak, K and Møller, J and Waagepetersen, R (2002) Markov Chain Bayesian analysis of log Gaussian processes for disease mapping. *Research Report*

[2] Møller, Jesper and Syversveen, Anne Randi and Waagepetersen, Rasmus Plenge (1998) Log gaussian cox processes. *Scandinavian journal of statistics*

[3] Vehtari, Aki and Gelman, Andrew (2014) WAIC and cross-validation in Stan. *Submitted. http://www. stat. columbia. edu/˜ gelman/research/unpublished/waic_stan. pdf Accessed*

ISBN 978-85-98026-63-3

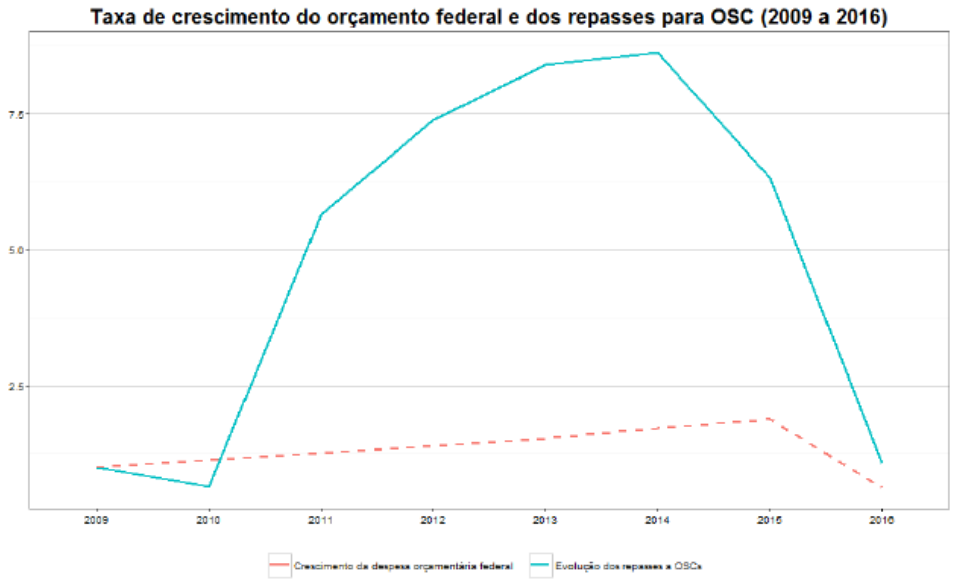
Análise dos repasses de recursos federais a organizações da sociedade civil (2009-2016).

André P. Vieira (UFRJ) / e-mail: andrehpv@gmail.com
Heraldo B. Filho (PUC/RJ) / e-mail: heraldoborges@gmail.com

INTRODUÇÃO

As organizações da sociedade civil (OSCs) são atores indispensáveis para o aprimoramento da administração pública em sociedades democráticas. A reconfiguração recente do papel e do lugar dessas organizações na provisão de políticas públicas tem sido acompanhada pela expectativa de que ocupem paulatinamente espaços maiores no orçamento público e no rol de programas e ações implementados pelos governos.

O Estado brasileiro está repassando mais recursos para as OSCs?



ISBN 978-85-98026-63-3

PROBLEMAS DE PESQUISA

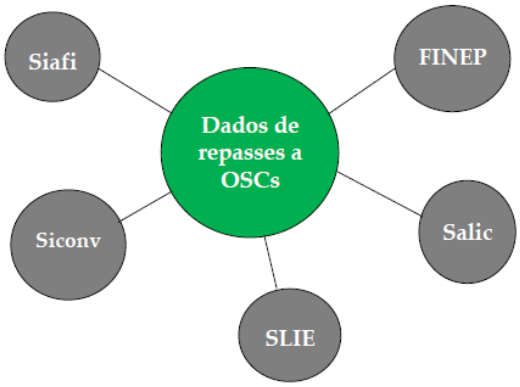
A pesquisa teve por objetivo verificar os seguintes pontos:

- Variações no volume de recursos transferidos ou repassados às organizações civis por instrumento de parceria;
- Proporção de transferências diretas da União vis-à-vis as transferências feitas indiretamente por meio de repasses da União a estados e municípios;
- Principais órgãos federais concedentes e áreas de políticas públicas recipientes de recursos;
- Distribuição de recursos transferidos ou repassados, por organização civil, natureza jurídica, tamanho, estado e região sede da organização.

As fundações privadas receberam, em média,
R$ 16 milhões desde 2009.

MÉTODOS

Os dados utilizados neste trabalho provêm de cinco fontes diferentes:



ISBN 978-85-98026-63-3

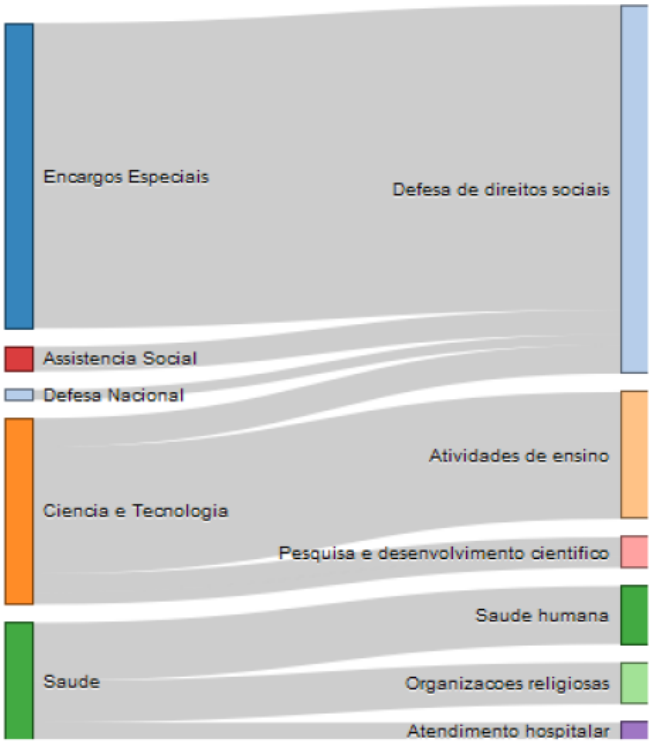
RESULTADOS

- 60% das OSCs que receberam recursos federais estavam em **SP, RJ ou MG**.
- Menos de 1% dos repasses no Siconv foi fruto de emendas parlamentares.
- 20% dos repasses para OSCs foram feitos indiretamente por meio de repasses a estados e municípios.

43%

dos repasses foram para OSCs de

defesa de direitos sociais.



ISBN 978-85-98026-63-3

Predição do Comportamento do Mercado Financeiro Utilizando Notícias

Heraldo Borges (PUC-Rio) / e-mail: heraldoborges@gmail.com

SUMÁRIO

Diversas teorias financeiras, entre elas a Hipótese do Mercado Eficiente, alegam a impossibilidade de prever o futuro do mercado de ações baseado na informação atualmente disponível. Entretanto, pesquisas recentes demonstram uma forte relação entre a hora da publicação de uma notícia e a consequente reação do mercado financeiro. Nosso objetivo é implementar um algoritmo de predição para mercado de ações que utiliza notícias jornalísticas sobre empresas de capital aberto. Utilizamos uma abordagem baseada em aprendizagem de máquina para a tarefa de predição do comportamento de um ativo nas posições de alta, baixa ou neutra, dentro de janelas de horários ao longo do dia.

ABORDAGEM

- Coleta de notícias jornalísticas sobre a Petrobras
- Coleta de dados históricos de informações *Intraday* do ativo *PETR4*

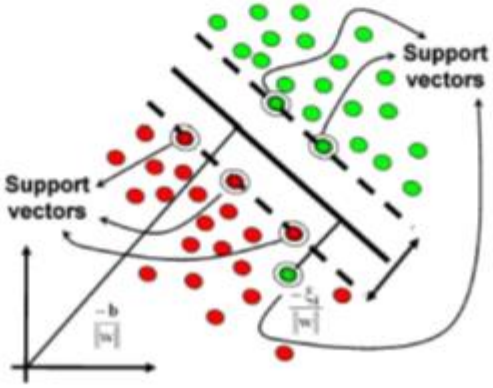
ISBN 978-85-98026-63-3

- Anotação automática das notícias
- Extração de atributos
- Construção de um classificador SVM para a classificação das notícias nas classes de alta, baixa ou neutra
- Avaliação do desempenho

CONHECIMENTO PRÉVIO

Classificação de Textos – processo de organização de documentos em classes previamente definidas

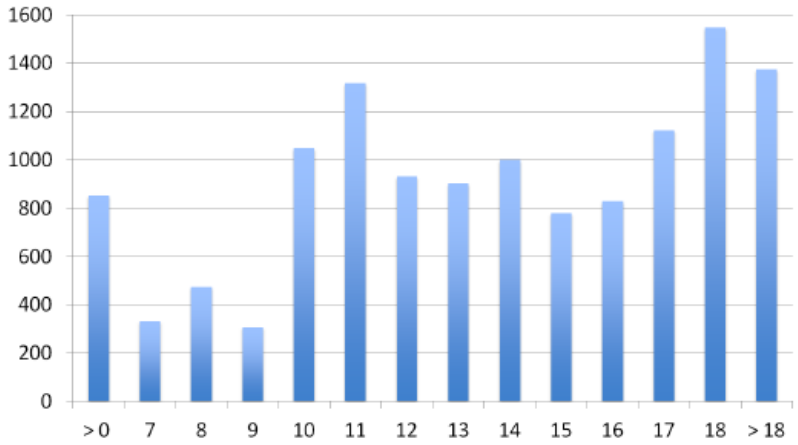
Support Vector Machine **(SVMs)** – modelo computacional associado com um algoritmo de aprendizado que recebendo um conjunto de dados de treino anotado gera um hiperplano que pode ser usado para classificar novos dados.



ISBN 978-85-98026-63-3

DADOS

- 9000 notícias jornalísticas sobre a Petrobras
- Período de 2009 à 2014
- Base de Dados de informações *Intraday* do ativo *PETR4* na BMF&Bovespa



Distribuição de notícias por hora de publicação

Horário	Abertura	Máx	Mín	Fecham.
2013-10-10 10:05:00	18.26	18.3	18.25	18.3
2013-10-10 10:06:00	18.3	18.31	18.27	18.28
2013-10-10 10:07:00	18.29	18.29	18.26	18.27
2013-10-10 10:08:00	18.28	18.3	18.26	18.29
2013-10-10 10:09:00	18.29	18.29	18.28	18.28
2013-10-10 10:10:00	18.28	18.28	18.27	18.28
2013-10-10 10:11:00	18.28	18.28	18.28	18.28
2013-10-10 10:12:00	18.28	18.29	18.28	18.29

Informação *Intraday* do ativo *PETR4*

ISBN 978-85-98026-63-3

EXPERIMENTO & RESULTADO

Seleção de Atributos:

- Representação por presença
- Atributos estruturais
- POS tag
- Bag of words
- N-gramms

Utilizou-se estratégia de múltiplos classificadores para determinadas frações de tempo durante o dia da negociação, sendo modelado apenas com notícias publicadas dentro desse intervalo de tempo. Nessa direção, foram criadas três faixas de horário:

- Notícias publicadas entre 7 e 11h
- Notícias publicadas entre 11 e 15h
- Notícias publicadas entre 15 e 18h

Modelo	Resultado Preliminar%
A. E. + 1,2,3-gramas + POS	62,60%

REFERÊNCIAS

Lavrenko, V.; Schmill, M.; Lawrie, D.; & Ogilvie, P. Language models for financial news recommendation. Proceedings of the Ninth International Conference of Information and Knowledge Management, p. 389-396, 2000.

Borges, H.; Milidiú, R.; Predição do Comportamento do Mercado Financeiro Utilizando Notícias emPortuguês. Tese (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, 2014.