

REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175



volume 75

número 240

janeiro / dezembro 2014

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 75 número 240 jan/dez 2014

ISSN 0034-7175

R. Bras. Estat., Rio de Janeiro, v. 75, n. 240, p. 1-117, jan./dez. 2014

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

© IBGE. 2015

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística - ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos através de divulgação de artigos inéditos tratando de aplicações da Estatística nas mais diversas áreas do conhecimento. Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para a produção e uso de estatísticas públicas.

Os originais para publicação deverão ser submetidos para o site <http://rbes.submitcentral.com.br/login.php>. Os artigos submetidos à RBES não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editor Responsável

Lúcia Pereira Barroso (IME-USP)

Editores Executivos

Pedro Luis do Nascimento Silva (ENCE/IBGE)

Mário de Castro Andrade Filho (ICMC-USP)

Editor de Metodologias

Fernando Antonio da Silva Moura (UFRJ)

Editor de Estatísticas Oficiais

José André de Moura Brito (ENCE/IBGE)

Editores Associados

Ana Maria Nogaes Vasconcelos (UNB)

Beatriz Vaz de Melo Mendes (UFRJ)

Cristiano Ferraz (UFPE)

Dalton Francisco de Andrade (UFSC)

Flávio Augusto Ziegelmann (UFRGS)

Francisco Louzada Neto (ICMC-USP)

Gleici Castro Perdoná (FMRP-USP)

Gustavo da Silva Ferreira (ENCE/IBGE)

Ismênia Blavatski de Magalhães (IBGE)

Thelma Sáfadi (UFLA)

Josmar Mazucheli (UEM)

Juvêncio Santos Nobre (UFC)

Luis A Milan (UFSCar)

Marcel de Toledo Vieira (UFJF)

Maysa Sacramento de Magalhães (ENCE/IBGE)

Paulo Justiniano Ribeiro Junior (UFP)

Pledson Guedes de Medeiros (UFRN)

Ronaldo Dias (UNICAMP)

Rosângela Helena Loschi (UFMG)

Solange Trindade Corrêa (Univ. Southampton)

Thelma Sáfadi (UFLA)

Viviana Giampaoli (IME-USP)

Editoração

Marilene Pereira Piau Câmara - ENCE/IBGE

Daniel Vitor de Araujo Azevedo - ENCE/IBGE

Dyana Cristina da Silva Braga - ENCE/IBGE

Impressão

Gráfica Digital / Centro de Documentação e Disseminação de Informações - CDDI/IBGE

Capa

Renato J. Aguiar - Coordenação de

Marketing/CDDI/IBGE

Ilustração da Capa

Marcos Balster - Coordenação de

Marketing/CDDI/IBGE

Revista brasileira de estatística / IBGE, - v.1, n.1
(jan./mar.1940), - Rio de Janeiro : IBGE, 1940 .v.

Trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística. Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989). Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística - Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

Gerência de Biblioteca e Acervos Especiais CDU 31(05)
RJ-IBGE/88-05 (rev.2009) PERIÓDICO

Impresso no Brasil/Printed in Brazil

Sumário

Nota da Editora	5
-----------------------	---

Artigos

A utilização da curva ROC para identificar fatores de risco, associados à mortalidade infantil	7
--	---

*Laís Helen Loose
Caroline Pafiadache da Silva
Luciane Flores Jacobi
Anaelena Bragança de Moraes*

Modelagem Estatística para previsão de resultados de Jogos de Futebol: Quem será o Campeão Brasileiro de 2010?	27
---	----

*Adriano Kamimura Suzuki
Francisco Louzada
José Fernando Francini
Luis Ernesto Bueno Salasar*

Comparação de modelos CAViaR e CARE para mensuração de risco	47
--	----

*Francyelle de Lima e Silva
Clélia Maria de Castro Toloí*

Aspectos Socioeconômicos e Ecológicos por escalonamento multidimensional e modelos de equações estruturais: Um estudo de caso no ZEE-MG	89
--	----

*Marcelo Angelo Cirillo
Lúcia Pereira Barroso*

Nota da Editora

O presente volume da RBEs reúne uma lista variada de contribuições. O artigo de Laís Helen Loose, Caroline Pafiadache da Silva, Luciane Flores Jacobi e Anaelena Bragança de Moraes apresenta uma aplicação da curva ROC (*Receiver Operating Characteristic*) com abordagem diferente da usual. No artigo de Adriano Kamimura Suzuki, Francisco Louzada, José Fernando Francini e Luis Ernesto Bueno Salazar são abordadas duas metodologias estatísticas para previsão de resultados de jogos de futebol. No artigo de Francielle de Lima e Silva e Clélia Maria de Castro Toloí é feita uma comparação entre os modelos de Valor de Risco Autorregressivo Condicional (CAViaR) e os modelos de *Expectil Autorregressivo Condicional* (CARE) aplicados na série do IBOVESPA. Finalmente o artigo de Marcelo Angelo Cirillo e Lúcia Pereira Barroso propõe um Modelo de Equações Estruturais após a aplicação do Escalonamento Multidimensional para regiões do Estado de Minas Gerais

Aproveito a oportunidade para agradecer a colaboração dos Editores Executivos Pedro Luis do Nascimento Silva (ENCE/IBGE) e Mário de Castro Andrade Filho (ICMC-USP), o Editor de Estatísticas Oficiais José André de Moura Brito (ENCE/IBGE) e o Editor de Metodologias Fernando Antonio da Silva Moura (UFRJ). Agradeço também aos Editores Associados, aos autores que submeteram seus artigos, IBGE, ABE, aos revisores, que anonimamente contribuíram para mais este número da Revista Brasileira de Estatística e a Marilene Pereira Piau Câmara pela editoração.

Desejo a todos que tenham uma excelente leitura.

Saudações cordiais

Lúcia Pereira Barroso
Editora Responsável

A utilização da curva ROC para identificar fatores de risco associados à mortalidade infantil

*Laís Helen Loose*¹
*Caroline Pafiadache da Silva*²
*Luciane Flores Jacobi*³
*Anaelena Bragança de Moraes*⁴

Resumo

A curva ROC (Receiver Operating Characteristic) é um método gráfico simples e de grande aplicabilidade prática. É particularmente utilizada para comparação de diferentes testes nas mais diversas áreas de estudo como psicologia, medicina e técnicas de imagem, economia e em previsões do tempo. O presente trabalho objetiva apresentar um material teórico referente à metodologia curva ROC, assim como uma aplicação sob abordagem diferente da usual. Foi desenvolvido um estudo transversal considerando 6963 crianças, em que, por meio da curva ROC, foram identificadas variáveis associadas aos fatores de risco relacionados à mortalidade infantil. Os dados utilizados são referentes aos nascidos vivos de mães residentes nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006. Através de um padrão sugerido por Câmara (2009) verificou-se que as variáveis: peso ao nascer e escores de Apgar são fatores de risco para o óbito durante o primeiro ano de vida.

Palavras-chave: Curva ROC, sensibilidade, especificidade, mortalidade infantil

¹ Universidade Federal do Rio Grande do Norte, Programa de Pós-Graduação em Matemática Aplicada e Estatística. E-mail: laisloose@gmail.com

² Universidade Federal de Santa Maria, Engenharia de Produção. E-mail: carolpafiadache@gmail.com

³ Universidade Federal de Santa Maria, Departamento de Estatística. E-mail: lucianefj8@gmail.com.

⁴ Universidade Federal de Santa Maria, Departamento de Estatística. E-mail: anaelena@smail.ufsm.br

1. Introdução

A análise ROC, Receiver-Operating Characteristic, é uma técnica bastante útil para visualização, avaliação, organização e seleção de sistemas ou predição. Por meio de um método gráfico simples, permite analisar a variação da sensibilidade e especificidade para os diferentes pontos de corte, ou seja, para diferentes valores possíveis da variável de decisão. Assim, o gráfico ROC nos permite verificar o limiar entre taxas de acertos e taxas de erro (alarmes falsos) de classificadores (Fawcett, 2005).

Conforme Zou et al. (2007), a análise ROC pode ser vista de forma mais geral, como um método para avaliar a precisão de um modelo estatístico (por exemplo, a regressão logística, análise discriminante linear) que classifica os objetos em uma de duas categorias. No contexto de mortalidade infantil, foco da aplicação do presente trabalho, classifica quais as variáveis que discriminam um recém-nascido com ou sem risco de vir a óbito.

A área sob a curva ROC foi introduzida na literatura médica por Hanley e McNeil (1982), sendo um critério frequentemente usado para examinar a eficácia discriminativa de um teste diagnóstico. Esta área pode ser reduzida a um escalar, o que permite quantificar a capacidade discriminativa. Nesse sentido quantificar as variáveis que discriminam um recém-nascido com ou sem risco de vir a óbito torna-se possível.

No Brasil, a mortalidade infantil, apesar de apresentar uma tendência decrescente, ainda constitui um importante problema para a saúde pública. Dentre os principais fatores de risco associados à mortalidade infantil, pode-se ressaltar: características do recém-nascido, condições socioeconômicas, características maternas e do sistema de saúde (Barros et al., 1987; Menezes et al., 1998; Mezzomo, 2009).

Em virtude da importância que a mortalidade infantil tem sobre os indicadores de saúde pública e pela necessidade de se analisar esses fatores, este trabalho apresenta um histórico sobre a curva ROC e propõe a sua utilização para identificar os fatores de risco associados à mortalidade neonatal e pós-natal infantil de nascidos vivos, bem como as variáveis candidatas a preditoras dos óbitos. Foram consideradas as mães residentes nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006, em função de características da mãe e do recém-nascido.

O presente trabalho está organizado conforme segue. A seção 2 apresenta o método da curva ROC, um breve histórico, a construção da curva ROC, o cálculo da área abaixo da curva e a comparação entre áreas. Na seção 3, é apresentada uma aplicação da metodologia curva ROC. Na seção 4, são expostos os resultados da análise dos dados e a discussão. Por fim, na seção 5 são descritas as conclusões.

2. A curva ROC

2.1 Histórico

A análise da curva ROC é baseada nos princípios fundamentais da teoria da decisão estatística e na teoria da detecção de sinais. Esta análise foi originalmente desenvolvida por um grupo de psicólogos matemáticos da Universidade de Michigan, durante a Segunda Guerra Mundial. O objetivo era quantificar a habilidade do operador de radar, ao diferenciar sinal de ruído (Zweig e Campbell, 1993; Decarlo, 1998; Harvey Jr., 2004; Zou et al., 2007; Metz, 2008; Martinez et al., 2003). Na teoria de detecção de sinal, a análise era realizada via teste de hipóteses; testava-se o ruído como hipótese nula, H_0 , e o ruído mais sinal como hipótese alternativa H_1 . Erros do tipo I eram chamados falhas, e do tipo II, falsos alarmes. A tarefa do observador era decidir qual dos estímulos é resultado de um ruído mais sinal, ou apenas ruído (Braga, 2000). Em seguida teve aplicabilidade na área da psicologia, através da avaliação do desempenho de tarefas em psicologia sensorial experimental e psicofísica (Swets, 1973).

Atualmente essa análise vem sendo considerada em diferentes áreas. Obuchowski (2003), Grunkemeier e Jin (2001) e Lusted (1971) consideraram essa metodologia na área médica e em técnicas de imagem médica radiológica. Prati et al. (2008) apresentaram uma introdução à análise ROC no contexto de aprendizado de máquina e mineração de dados. Na área da economia, Brito e Neto (2008) a utilizaram para validação de modelos de classificação de risco de crédito de empresas. Ainda no contexto de previsões, Gorr (2009) utilizou a curva ROC para avaliar a qualidade de predições.

Verifica-se ainda a grande aplicabilidade da curva ROC na área da saúde. Em medicina, por exemplo, o custo de classificar incorretamente um paciente doente como sadio é muito maior do que classificar um paciente sadio como doente, dado que no

primeiro caso, a falha do diagnóstico pode levar à morte do paciente (Martinez et al., 2003; Prati et al., 2008).

2.2 Tabela de Contingência

No contexto de análise de diagnóstico, temos quatro resultados possíveis. Se o evento é positivo e classificado como positivo, o resultado é contabilizado como verdadeiro positivo, e se classificado como negativo, é contabilizado como falso negativo. Se o evento é negativo e classificado como negativo, o resultado é contabilizado como verdadeiro negativo, e se classificado como positivo, é contabilizado como falso positivo (Fawcett, 2005). Resultados verdadeiros positivos e verdadeiros negativos são considerados como acertos, e falsos negativos e falsos positivos considerados erros (Heeger, 1997). Esses possíveis resultados podem ser representados em uma tabela 2 X 2, chamada matriz de confusão ou tabela de contingência (Tabela 1).

Tabela 1: Tabela de Contingência 2 x 2 para modelos de classificação

		Evento		
Diagnóstico		Positivo	Negativo	
	Positivo	<i>VP</i> (verdadeiro positivo)	<i>FP</i> (falso positivo)	<i>VP + FP</i>
	Negativo	<i>FN</i> (falso negativo)	<i>VN</i> (verdadeiro negativo)	<i>FN + VN</i>
		<i>VP + FN</i>	<i>FP + VN</i>	<i>VP + FP + FN + VN = N</i>

Swets, 1988; Heeger, 1997; Fawcett, 2005; Prati et al., 2008.

2.3 Sensibilidade e Especificidade

A curva ROC é baseada em dois conceitos, sensibilidade e especificidade. A sensibilidade é definida como a probabilidade do resultado de um evento sob observação ter diagnóstico positivo, dado que o evento realmente é positivo, ou ainda fração de verdadeiros positivos, dada na equação (1). A especificidade é definida como a probabilidade do resultado de um evento sob observação ter diagnóstico negativo, dado que o evento é realmente negativo, ou fração de verdadeiro negativo, dada na equação (2).

Conforme Castanho et al. (2004), as medidas de sensibilidade e especificidade fornecem, respectivamente, a probabilidade de que um teste classifique corretamente um indivíduo doente e um indivíduo saudável. Tais medidas são dadas pelas equações (1) e (2), respectivamente:

$$S_e = \frac{VP}{VP + FN} \quad (01)$$

$$E_s = \frac{VN}{VN + FP} \quad (02)$$

em que VP representa os verdadeiros positivos, FP os falsos positivos, VN os verdadeiros negativos e FN os falsos negativos (Fawcett, 2005; Swets, 1979) .

No plano unitário, a curva ROC é representada graficamente pelos pares sensibilidade e 1-especificidade, ordenada e abscissa, respectivamente, resultantes da variação do ponto de corte ao longo de um eixo de decisão (Braga, 2000).

2.4 Característica para avaliação da curva ROC

Quando a sensibilidade e 1-especificidade são projetadas no plano (x, y), os valores das coordenadas variam de zero a um. Dado que são probabilidades, a curva é limitada por um quadrado unitário. Há ainda uma linha diagonal positiva neste quadrado, chamada linha do acaso (Braga, 2000; Fawcett, 2005; Prati et al., 2008).

A curva deve começar em (0,0) e terminar em (1,1), sendo não decrescente em todo o seu percurso. O ponto (0,0) representa a estratégia de nunca classificar um exemplo como positivo. Modelos que correspondem a esse ponto não apresentam nenhum falso positivo, mas também nenhum verdadeiro positivo.

A estratégia inversa, de sempre classificar um novo exemplo como positivo, é representada pelo ponto (100%,100%). O ponto (0,100%) representa o modelo perfeito, isto é, todos os exemplos positivos e negativos são corretamente classificados. O ponto (100%, 0) representa o modelo que sempre faz previsões erradas (Fawcett, 2005; Prati et al., 2008).

Em relação à linha do acaso, os pontos pertencentes ao triângulo superior esquerdo dessa diagonal representam modelos que tem desempenho melhor que o aleatório. Ainda pontos pertencentes ao triângulo inferior direito representam modelos piores que o aleatório (Egan, 1975 apud Braga, 2000).

Espera-se que o teste seja, ao mesmo tempo, altamente sensível e altamente específico. Em relação ao desempenho de diferentes sistemas para uma curva ROC, quanto mais próxima do canto superior esquerdo, maior o poder discriminante.

Um ponto no espaço ROC é melhor que outro se e somente se ele estiver acima e à esquerda do outro ponto. Ou seja, tem maior taxa de verdadeiros positivos e menor taxa de falsos positivos (Prati et al., 2008).

Quanto mais distante a curva estiver da diagonal principal, melhor será o desempenho do sistema de aprendizado para aquele domínio. Ao se comparar duas (ou mais) curvas, caso não haja nenhuma intersecção, a curva que mais se aproxima do ponto (0, 100%) é a de melhor desempenho. Caso haja intersecções, cada um dos sistemas tem uma faixa operacional na qual é melhor que o outro. Idealmente, a curva deve ser convexa e sempre crescente (Swets, 1988; Braga, 2000; Prati et al., 2008).

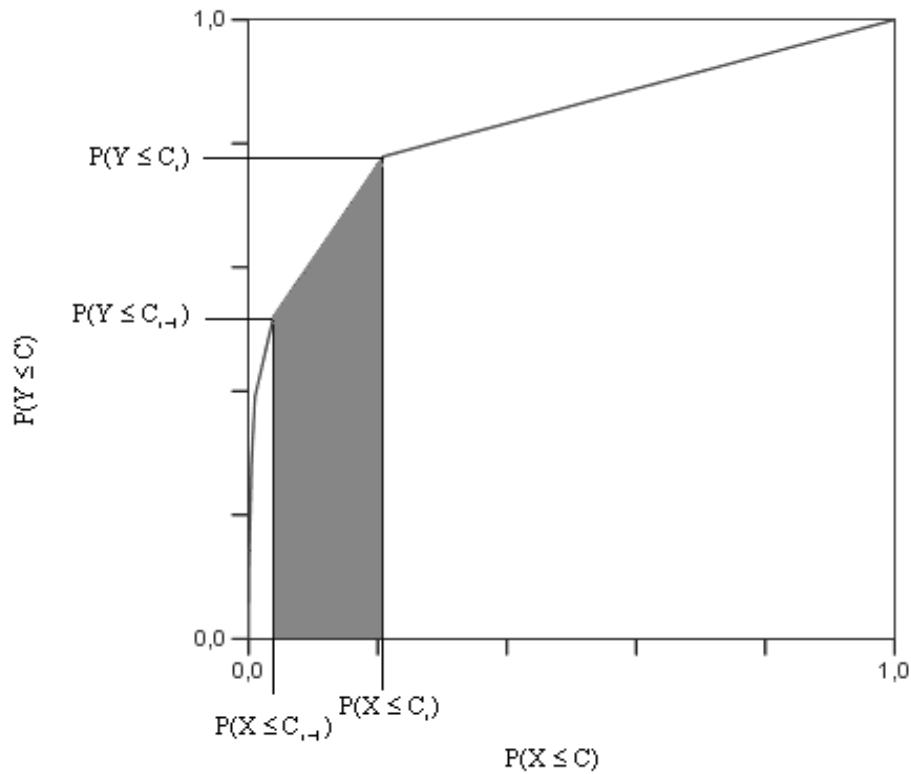
2.5 Cálculo da área abaixo da curva ROC

A curva ROC é uma demonstração bidimensional da performance de um classificador. Então, faz-se necessário reduzir a curva a um valor escalar, este valor é a área abaixo da curva (Fawcett, 2005), sendo o principal índice para avaliar a “qualidade” da curva (Hanley e McNeil, 1982).

Alguns métodos para a obtenção da área abaixo da curva são apresentados na literatura, dentre eles temos a estimação via máxima verossimilhança (Hanley e McNeil, 1983; Metz, 2008), aproximação à estatística de Wilcoxon – Mann – Whitney (Hanley e McNeil, 1982; 1983) e a regra dos trapézios, apresentada a seguir.

Pelo método dos trapézios, sendo X e Y discretas finitas, considera-se $c_1, \dots, c_k$ o conjunto ordenado de valores que X e Y podem tomar com probabilidade não nula. Seja, ainda, c_0 um valor arbitrário menor que c_1 . A área $[A(X, Y)]$ é calculada dividindo o gráfico em trapézios. Calcula-se a área de cada trapézio (A_i), conforme a Figura 1.

Figura 1: Área da curva ROC dividida em trapézios para o cálculo da área.



Analiticamente pode-se calcular a área dos trapézios por meio de:

$$\begin{aligned} A_i &= P(X = c_i) \left[\frac{1}{2} P(Y \leq c_i) + \frac{1}{2} P(Y \leq c_{i-1}) \right] \\ &= P(X = c_i) \left[P(Y \leq c_{i-1}) + \frac{1}{2} P(Y = c_i) \right] \end{aligned}$$

Como $A(X,Y)$ é igual à soma dos diversos A_i , temos:

$$\begin{aligned} A(X,Y) &= \sum_{i=1}^k P(X = c_i) \left[P(Y \leq c_{i-1}) + \frac{1}{2} P(Y = c_i) \right] \\ &= \sum_{i=1}^k P(X = c_i) P(Y \leq c_{i-1}) + \frac{1}{2} \sum_{i=1}^k P(X = c_i) P(Y = c_i) \\ &= P(Y < X) + \frac{1}{2} P(X = Y) \end{aligned}$$

Para X e Y contínuas, $P(X=Y)$ é nula, assim, a equação é válida para ambas as situações, para variáveis aleatórias discretas finitas e variáveis aleatórias contínuas. A (X, Y) pode tomar qualquer valor com mínimo igual a zero e máximo igual a um.

Quanto ao valor calculado da área da curva, Câmara (2009) sugere um padrão para a classificação da área: se esta estiver entre 0,5 e 0,6, a probabilidade é insignificante; entre 0,6 e 0,7, é ruim; entre 0,7 e 0,8, é regular; entre 0,8 e 0,9, é bom e, acima de 0,9, é excelente, sendo limitada por 1.

2.6 Teste para a diferença entre duas áreas abaixo da curva ROC

Para testar se a diferença entre duas áreas abaixo das curvas ROC derivadas de diagnósticos alternativos com base no mesmo conjunto de dados é aleatória ou real, tem-se como hipótese nula as áreas abaixo da curva ROC não diferem, contra a hipótese alternativa que as áreas diferem. A estatística de teste é dada pela razão crítica calculada z , a qual tem distribuição normal assintótica (Martinez et al., 2003; Braga, 2000), definida pela equação (3):

$$z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \quad (03)$$

em que A_1 e SE_1 e A_2 e SE_2 representam as áreas e erros estimados da curva ROC para as modalidades 1 e 2, respectivamente, r a correlação entre A_1 e A_2 , sendo esta obtida com base no método sugerido por Hanley e McNeil (1983), por meio de uma tabela que determina a correlação entre as áreas.

A quantidade Z refere-se às tabelas de distribuição normal padrão, os valores de Z em (3) acima de um determinado valor evidenciam, estatisticamente, que as verdadeiras áreas entre as curvas ROC diferem. O termo $2rSE_1SE_2$ no denominador é de grande importância devido ao fato de os dados estarem correlacionados, pois são provenientes de uma mesma amostra. A ausência desse termo implicaria um denominador de maior valor e, conseqüentemente, o valor de Z menor, o que reduziria a possibilidade de se verificar diferença significativa entre as duas curvas (Braga, 2000; Hanley e McNeil, 1983).

Pelo método sugerido por Hanley e McNeil (1983), utiliza-se uma tabela para a determinação do coeficiente de correlação (r) entre as áreas. Calcula-se o coeficiente de correlação entre os valores das modalidades 1 e 2, para as observações que possuem a característica observada c_p e para os que não a possuem c_n , podendo ser obtida pelo método da correlação de Pearson ou pela correlação Tau de Kendall. Calcula-se a média das correlações $c = \frac{c_p + c_n}{2}$ e a média das áreas $A = \frac{A_1 + A_2}{2}$, que vão constituir as entradas na tabela construída por Hanley e McNeil (1983), a partir do qual se verifica o valor de r .

3. Aplicação da Metodologia curva ROC

Nesta seção, apresentamos uma aplicação da metodologia curva ROC, por meio de um estudo relacionado à mortalidade infantil envolvendo os nascidos vivos de mães residentes nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006. Com a utilização da curva ROC, verificamos quais as variáveis seriam as que melhor explicariam os óbitos dos recém-nascidos. Para tanto, realizaram-se etapas iniciais, com a análise descritiva dos dados. Em seguida, foi feita a seleção das variáveis consideradas como boas preditoras dos óbitos e, após, apresentamos os resultados.

3.1 Descrição do Estudo

As informações utilizadas sobre os nascidos vivos (NV) foram as provenientes das Declarações de Nascido Vivo (DN), cujas informações constam no Sistema de Informações sobre Nascidos Vivos (SINASC), e sobre os óbitos neonatais e pós-neonatais, as provenientes das Declarações de Óbitos (DO), no ano de 2006, que constam no Sistema de Mortalidade (SIM). A população considerada neste estudo foi de 6963 NV, pertencentes à 4ª Coordenadoria Regional de Saúde do Rio Grande do Sul. A identificação dos óbitos foi realizada mediante compartilhamento de banco de dados, vinculando-se o número da DN comum às bases de nascimentos (Nascimento et al., 2008).

Foram selecionadas algumas variáveis do banco de dados original. As consideradas no estudo são apresentadas no Quadro 1, referente às características da mãe e dos recém-nascidos.

Quadro 1: Variáveis consideradas no estudo relativas à mãe e às características do recém – nascido nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006.

Variáveis consideradas no estudo:			
Relacionadas à mãe:		Relacionadas ao bebê:	
Idade (Idade mãe)	Anos	Peso ao nascer (Peso)	Gramas
Número de Partos (Paridade)	1 ou 2 filhos	Índice de Apgar 1 minuto (Apgar1)	De 0 a 10
	3 filhos ou mais		
	Nuríparas		
Escolaridade (Escola mãe)	0 a 3 anos	Índice de Apgar 5 minuto (Apgar5)	De 0 a 10
	4 a 11 anos		
	12 anos concluídos ou mais de estudo		
Número de consultas pré-natais (Consultas)	Nenhuma	Raça/Cor (Racacor)	Branca
	De 1 a 6		Preta
	7 ou mais		Amarela
Tipo de gravidez (Gravidez)	Única		Parda
	Múltipla		Indígena
Tipo de parto (Parto)	Vaginal	Sexo (Sexo)	Masculino
	Cesáreo		Feminino
Local de nascimento (Locnas)	Hospital		
	Outro		

Inicialmente foi realizada a análise descritiva das variáveis quantitativas do Quadro 1, em seguida, com o objetivo de verificar a sua eficiência para prever os óbitos, calculou-se a sensibilidade e a especificidade.

Através do software SPSS 17.0, obteve-se o gráfico da curva ROC para as variáveis relativas à mãe e as relativas ao recém-nascido e, ainda, obteve-se a área abaixo da curva e os erros-padrão da curva. Conforme o padrão adotado, sugerido por Câmara (2009) do banco de dados inicial, foram selecionadas apenas algumas variáveis, sendo estas as que possuíam maior área. Em seguida, para estas variáveis, calculou-se a correlação pelo Tau de Kendall e, pelo método sugerido por Hanley e McNeil (1983), testou-se a diferença significativa entre as áreas que, segundo o padrão adotado, seriam as melhores preditoras dos óbitos. Foi considerado um nível de significância de 5%.

4. Resultados e Discussões

O número total de nascidos vivos de mães residentes nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, no ano de 2006, foi de 6963 nascidos vivos, 6883 sobreviveram e 80 vieram a óbito, sendo os óbitos representados por 1,1%. Em relação ao sexo dos bebês, 3564 foram registrados nascidos do sexo masculino (51,2%) e 3395 nascidos do sexo feminino (48,8 %).

Entre os bebês que permaneceram vivos (n=6883), observa-se, na Tabela 2, que a variável peso apresentou média de 3181,91 g com mínimo de 560,00 g e máximo de 5380,00 g. Para os recém-nascidos que morreram (n=80), verificou-se média de peso 2016,38 g, mínimo de 500,00 g e máximo de 4650,00 g.

Em relação à idade da mãe (dos bebês que permaneceram vivos), observa-se média de 26,28 anos, mínimo de 12 anos e máximo de 52 anos, com desvio-padrão de 6,86 anos. Ainda, para os óbitos, observa-se média de 24,84 anos, mínimo de 14 anos e máximo de 44 anos.

As variáveis Apgar 1º e 5º minuto apresentaram mediana e moda iguais sendo 9 e 10 os seus valores, respectivamente, para os vivos. Para os óbitos, verifica-se mediana 6 e múltiplas modas para a variável Apgar1, sendo que, para o Apgar5, a moda foi igual a 9 e a mediana igual a 8.

Então, temos para os recém-nascidos que vieram a óbito, as mães em geral são mais jovens, os bebês têm menor peso e valores inferiores de Apgar1 e Apgar5, quando comparados aos que sobreviveram.

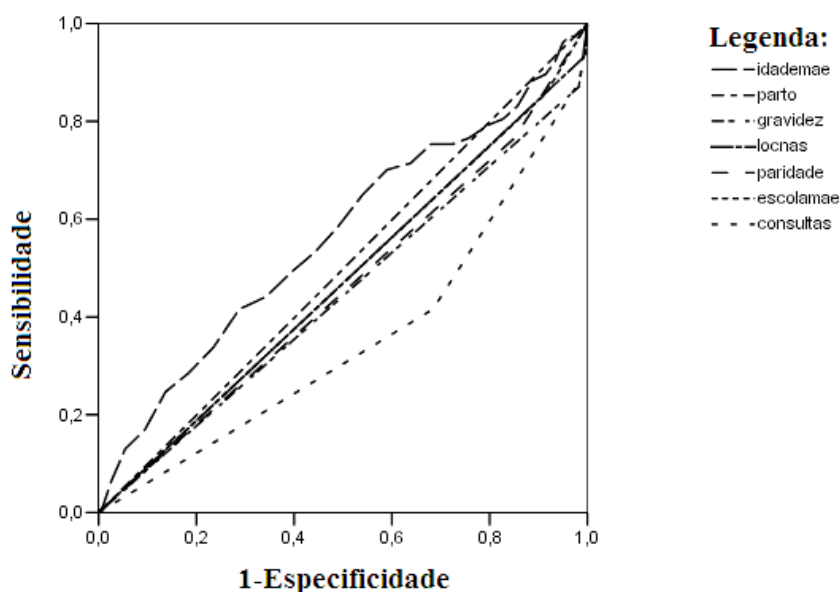
Tabela 2: Medidas descritivas para as variáveis: idade da mãe (em anos), peso (em gramas), Apgar1 e Apgar5 (escore de 0 a 10) de recém-nascidos vivos nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006, separados segundo o desfecho (óbitos e vivos).

	Variáveis	Média	Moda	Mediana	Mínimo	Máximo	Desvio Padrão
Vivo (n=6883)	Idade mãe	26,28	22,00	26,00	12,00	52,00	6,86
	Peso	3181,91	3000,00	3200,00	560,00	5380,00	538,11
	Apgar1		9,00	9,00	0,00	10,00	
	Apgar5		10,00	10,00	0,00	10,00	
Óbitos (n=80)	Idade mãe	24,84	18,00	23,50	14,00	44,00	7,54
	Peso	2016,38	**	1947,50	500,00	4650,00	1076,65
	Apgar1		**	6,00	0,00	10,00	
	Apgar5		9,00	8,00	0,00	10,00	

**múltiplas modas

Na Figura 2 e na Tabela 3, são apresentados o gráfico da curva ROC para as variáveis relativas à mãe e as áreas abaixo da curva ROC para essas variáveis, respectivamente.

Figura 2: Gráfico das variáveis relativas à mãe de recém-nascidos nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006.



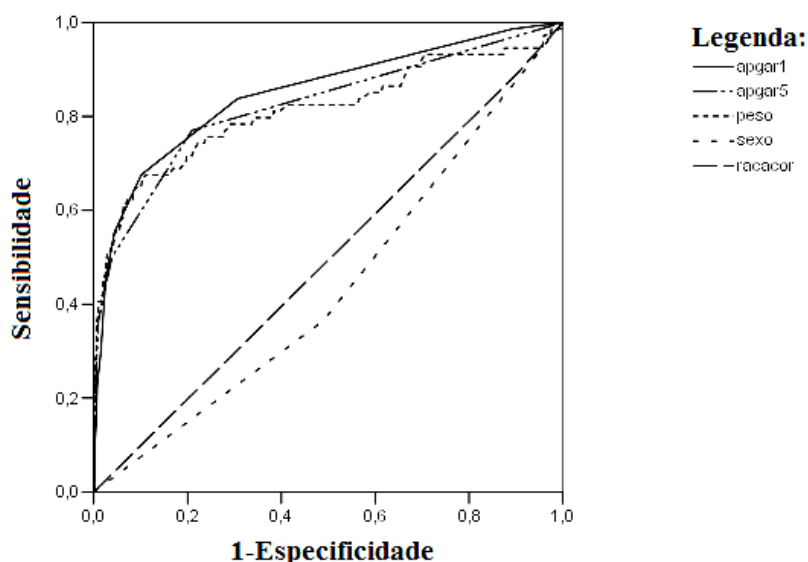
Analisando as variáveis relativas à mãe, observa-se, na Figura 2, que nenhuma curva se aproxima do canto superior esquerdo e, na Tabela 3, observa-se, com base no padrão adotado, que as mesmas são insignificantes para prever os óbitos, pois nenhuma tem valor acima de 0,6.

Tabela 3: Valores das áreas situadas abaixo da curva ROC e erros-padrão para as variáveis relacionadas à mãe dos recém-nascidos vivos nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006.

Variáveis relacionadas à mãe	Áreas	Erros-padrão
Paridade	0,458	0,035
Escola mãe	0,471	0,034
Consultas	0,352	0,034
Idade mãe	0,565	0,036
Parto	0,499	0,033
Gravidez	0,444	0,036
Locnas	0,469	0,035

A Figura 3 e a Tabela 4 apresentam respectivamente as áreas referentes à curva e os valores das áreas abaixo da curva para as variáveis relacionadas às características do recém – nascido.

Figura 3: Gráfico da curva ROC das variáveis relativas a recém-nascidos de mães residentes nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006.



Analisando as variáveis relativas ao recém-nascido, observa-se, na Figura 3, que as que mais se aproximam do canto superior esquerdo são peso, Apgar1 e Apgar5. Ainda, analisando a Tabela 4, pode-se concluir, conforme o padrão adotado, que o valor da área encontrada para essas variáveis é considerado bom, estando entre 0,8 e 0,9.

Comparando os resultados obtidos neste trabalho com outros estudos que utilizaram diferentes abordagens temáticas, observa-se que o peso ao nascer é considerado um forte determinante de risco relacionado a óbitos antes do primeiro ano de vida, o que confirma o descrito na literatura. Sendo o mesmo associado aos nascimentos prematuros e retardo do crescimento intra-uterino, situações decorrentes de problemas de saúde materna. Esse determinante é mencionado como um indicador do risco social relacionado às precárias condições socioeconômicas e nutricionais de uma população, mostrando, assim, a complexidade dos fatores associados e a necessidade de um aprofundamento nos estudos das causas desse determinante (Morais Neto e Barros, 2000; Sarinho et al., 2001; Martins e Velásquez-Meléndez, 2004).

Tabela 4: Valores das áreas situadas abaixo da curva ROC e erros-padrão para as variáveis relacionadas aos recém-nascidos vivos de mães residentes nos municípios pertencentes à 4ª Coordenadoria Regional de Saúde, RS, no ano de 2006.

Variáveis relacionadas ao recém-nascido	Áreas	Erros-padrão
Apgar1	0,848	0,027
Apgar5	0,823	0,031
Peso	0,812	0,034
Sexo	0,438	0,033
Racacor	0,495	0,034

O índice de Apgar é amplamente utilizado para mensurar a vitalidade do recém-nascido, varia de 0 a 10 e avalia cinco sintomas objetivos: frequência cardíaca, respiração, irritabilidade reflexa, tônus muscular e cor (Kilsztajn et al., 2007).

Em relação aos resultados encontrados quanto ao Apgar 1º minuto e Apgar 5º minuto, confirma-se o que Sarinho et al. (2001), Silva et al. (2006) e Paulucci e Nascimento (2007) encontraram em seus estudos: que os mesmos são considerados como importantes variáveis relacionadas aos óbitos. Esses autores verificaram que, quanto menor o escore do Apgar, maior a probabilidade do bebê vir a óbito. Neste estudo, observou-se que as variáveis Apgar 1º minuto e Apgar 5º minuto são também consideradas importantes para explicar os óbitos, porém, não foram associadas aos seus níveis de escores.

A variável sexo, neste trabalho, não apresentou relação com os óbitos, concordando com alguns estudos (Morais Neto e Barros, 2000; Martins e Velásquez-Meléndez, 2004). Também a variável raça/cor, não se mostrou associada aos óbitos.

Para a variável idade materna, estudos indicam que ela tem influência sobre os óbitos (Souza e Gotlieb, 1993; Araújo et al., 2000; Almeida et al., 2002). No entanto, neste estudo, não se verificou relação entre idade materna e mortalidade, confirmando o estudo de Duarte e Mendonça (2005).

A escolaridade materna também não se mostrou fator de risco, coincidindo com os resultados encontrados por Araújo et al. (2000), Moraes Neto e Barros (2000), Almeida e Barros (2004), e Paulucci e Nascimento (2007).

O número de consultas pré-natais, segundo Martins e Velásquez-Meléndez, (2004) e Silva et al. (2006), está relacionado à mortalidade infantil. Já, para Almeida e Barros (2004), o número de consultas pré-natais não se manteve associado ao óbito neonatal após este ser controlado pela duração da gestação. No presente estudo, essa variável não possui associação com os óbitos neonatais.

Nos achados de Menezes et al. (1998), o tipo de parto está diretamente ligado à mortalidade neonatal precoce. Para Almeida et al. (2002), o parto do tipo cesáreo mostrou-se protetor para a mortalidade neonatal. No entanto, neste estudo, o tipo de parto não teve relação com os óbitos, confirmando os estudos de Araújo et al. (2000) e Silva et al. (2006).

Nesta análise, a variável tipo de parto não apresentou associação com o óbito neonatal precoce o que condiz com os achados de Menezes et. al (1998) e Paulucci e Nascimento (2007).

A variável local de nascimento não apresentou relação com os óbitos, o que se confirma no estudo de Silva et al. (2006).

O tipo de gravidez (única ou múltipla) e também a paridade não tiveram relação com a mortalidade infantil, confirmando os resultados dos estudos de Sarinho et al. (2001), Martins e Velásquez-Meléndez, (2004). Já nos estudos de Machado e Hill (2003), houve associação entre a paridade acima de 3 e morte pós-neonatal.

Segundo o método proposto por Hanley e McNeil (1983), representado pela equação (3), deve-se testar a hipótese de que as áreas abaixo das curvas apresentam diferenças entre si. Assim, para as variáveis que apresentaram maior área: peso, Apgar 1º minuto e Apgar 5º minuto foram testadas essas diferenças, duas a duas, obtendo-se: Apgar1 versus peso $p = 0,332$, Apgar1 versus Apgar5 $p = 0,379$ e Apgar5 versus peso $p = 0,734$. Conclui-se que as curvas para as três variáveis testadas não diferem entre si considerando que, para cada comparação, foi adotado um nível de 5% de significância.

Com base neste teste, verifica-se que as variáveis peso, Apgar1 e Apgar5 podem ser consideradas preditoras dos óbitos, o que é confirmado e evidenciado em diversos estudos (Menezes et al., 1998; Araújo et al., 2000; Sarinho et al., 2001; Almeida et al., 2002; Machado e Hill, 2003; Martins e Velásquez-Meléndez, 2004; Almeida e Barros, 2004; Duarte e Mendonça, 2005; Paulucci e Nascimento, 2007).

5. Conclusão

O estudo aqui apresentado aponta, por meio da análise da curva ROC (Receiver Operating Characteristics), possíveis fatores preditores de óbitos neonatais e pós-neonatais em Santa Maria – RS.

As variáveis peso, Apgar 1º minuto e Apgar 5º minuto são as que melhor explicam o óbito de recém-nascidos. Dessa forma, é possível concluir que são necessárias ações de qualidade em saúde que resultem no peso adequado das crianças ao nascer. A assistência ao recém-nascido na sala de parto com abordagem sistemática e equipe preparada para prestar atendimento àqueles que apresentam baixos escores de Apgar também poderá diminuir a chance de óbito neonatal e pós-neonatal.

Ainda, em relação a essas três variáveis, observa-se que qualquer uma delas pode ser utilizada para “monitorar” os bebês, para uma melhora nos índices da mortalidade infantil.

Destaca-se ainda que a utilização da curva ROC possui vantagens frente a outras técnicas para determinação de fatores de risco. Em especial ao fato de basear-se em um gráfico de fácil obtenção, e simples interpretação. Espera-se que os pontos aqui levantados sejam úteis para difundir ainda mais a utilização da metodologia da curva ROC nas diferentes áreas de aplicabilidade.

Referências bibliográficas

- Almeida, M. F., Novaes, H. M. D., Alencar, G. P., Rodrigues, L. C. (2002). Mortalidade neonatal no Município de São Paulo: influência do peso ao nascer e de fatores sócio-demográficos e assistenciais. *Revista Brasileira de Epidemiologia* 5(1), 93 – 107.
- Almeida, S. D. M., Barros, M. B. A. (2004). Atenção à saúde e mortalidade neonatal: estudo caso-controle realizado em Campinas, SP. *Revista Brasileira de Epidemiologia* 7 (1), 22 – 35.
- Araújo B. F., Bozzetti M. C., Tanaka A. C. A. (2000). Mortalidade neonatal precoce no Município de Caxias do Sul: um estudo de coorte. *Jornal de Pediatria*, 76, 200-206.
- Barros, F. C., Victoria, C. G., Vaughan, J. P., Estanislau, H. J. (1987). Perinatal mortality in southern Brazil: a population-based study of 7392 births. *Bulletin of the World Health Organization*, 65 (1), 95 – 104.
- Braga, A. C. S. (2000). Curvas ROC: aspectos funcionais e aplicações. Braga: Universidade do Minho, Tese.
- Brito, G. A. S., Neto, A. A. (2008). Modelo de classificação de risco de crédito de empresas. *Revista Contabilidade & Finanças*, 46(19), 18 – 29.
- Câmara, F. P. (2009). *Psiquiatria e Estatística V: Validação de procedimentos diagnóstica pela Curva ROC. Coluna Psiquiatria Contemporânea*. 14 (4). Disponível em: <<http://www.polbr.med.br/ano09/cpc0409.php>> . Acesso em: 28 jun. 2010.
- Castanho, M. J. P., Yamakami, A., Barros, L. C. de, Vendite, L. L. (2004). Avaliação de um teste em medicina usando uma curva ROC fuzzy. *Biomatemática*, 14, 19 – 28.
- Decarlo, L. T. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods*, 3 (2), 186-205.
- Duarte, J. L. M. B., Mendonça, G. A. S. (2005). Fatores associados à morte neonatal em recém-nascidos de muito baixo peso em quatro maternidades no Município do Rio de Janeiro, Brasil. *Cadernos de Saúde Pública*, 21 (1), 181 – 191.
- Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, 861 – 874.
- Gorr, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristic curves. *International Journal of Forecasting*, 25(1), 48 – 61.
- Grunkemeier, G. L., Jin, R. (2001). Receiver Operating Characteristic Curve Analysis of Clinical Risk Models. *Annals of Thoracic Surgery*, 72, 323 – 326.
- Hanley, J. A., McNeil, B. I. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curves. *Radiology*, 143, 29 – 36.
- Hanley, J. A., McNeil, B. I. (1983). A Method of Comparing the Areas under Receiver Operating Characteristic curves derived from the same case. *Radiology*, 148, 839 – 843.
- Harvey Jr., L. O. (2004). Detection Theory: Sensitivity and Response Bias. *Psychology of Perception Psychology*.
- Heeger, D. (1997). Signal Detection Theory. Disponível em: <<http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html>> . Acesso em: 15 out. 2010.
- Kilsztajn, S., Lopes, E. S., Carmo, M. S. N., Reyes, A. M. A. (2007). Vitalidade do recém-nascido por tipo de parto no Estado de São Paulo, Brasil. *Caderno de Saúde Pública*, 23 (8), 1886 – 1892.
- Lusted, L. B. (1971). Signal Detectability and Medical Decision-Making. *Science*, 171, 1217 – 1219.

- Machado, C.J., Hill, K. (2003). Determinantes da mortalidade neonatal e pós-neonatal no Município de São Paulo. *Revista Brasileira de Epidemiologia*, 6(4), 345 – 358.
- Martinez, E. Z., Louzada-Neto, F. L., Pereira B. de B. (2003). A curva ROC para testes diagnósticos. *Cadernos de Saúde Coletiva*, 11 (1), 7 – 31.
- Martins, E. F.; Velásquez-Meléndez, G. (2004). Determinantes da mortalidade neonatal a partir de uma coorte de nascidos vivos, Montes Claros, Minas Gerais, 1997-1999. *Revista Brasileira de Saúde Materno Infantil*, 4 (4), 405 – 412.
- Menezes, A. M. B., Barros, F. C., Victora, C. G., Tomasi, E., Halpern, R., Oliveir, A. L. B. (1998). Fatores de risco para mortalidade perinatal em Pelotas, RS, 1993. *Revista de saúde pública*. 32 (3), 209 – 216.
- Metz, C. E. (2008). ROC analysis in medical imaging: a tutorial review of the literature. *Radiological Physics Technology*, 1, 2 – 12.
- Mezzomo, M. (2009). Estudo da Mortalidade Infantil – um Modelo de Regressão Logística Múltipla. Santa Maria: UFSM, Monografia
- Morais Neto, O. L.; Barros, M. B. A. (2000). Fatores de risco para a mortalidade neonatal e pós-neonatal na região centro-oeste do Brasil: linkage entre banco de dados de nascidos vivos e óbitos infantis. *Cadernos de Saúde Pública*, 16 (2), 477 – 485.
- Nascimento, E. M. R.; Costa, M da C. N.; Mota, E. L. A.; Paim, J. S. (2008). Estudo de fatores de risco para óbitos de menores de um ano mediante compartilhamento de bancos de dados. *Cadernos de Saúde Pública*, 24, 2593 – 2602.
- Obuchowski, N. A. (2003). Receiver Operating Characteristic Curves and Their Use in Radiology. *Radiology*, 229, 3 – 8.
- Paulucci, R. S., Nascimento, L. F. C. (2007). Mortalidade neonatal em Taubaté: um estudo de caso. *Revista Paulista de Pediatria*, 25 (4), 358 – 363.
- Prati, R. C., Batista, G. E. A. P. A., Monard, M. C. (2008). Curvas ROC para avaliação de classificadores. *IEEE Latin América Transactions*, 6 (2), 215 – 222.
- Sarinho, S. W.; Filho, D. A. M.; Silva, G. A. P.; Lima M. C. (2001). Fatores de risco para óbitos neonatais no Recife: um estudo caso-controle. *Jornal de Pediatria*, 77 (4), 294 – 298.
- Silva, C. F.; Leite, A. J. M.; Almeida, N. M. G. S. de; Gondim, R. C. (2006). Fatores de risco para mortalidade infantil em município do Nordeste do Brasil: linkage entre bancos de dados de nascidos vivos e óbitos infantis, 2000 a 2002. *Revista Brasileira de Epidemiologia*, 9 (1), 69 – 80.
- Souza, R. K. T., Gotlieb, S. L. D. (1993). Probabilidade de morrer no primeiro ano de vida em área urbana da região sul, Brasil. *Revista de Saúde Pública*, 27 (6), 445 – 454.
- Swets, J. A. (1973). The relative operating characteristic in psychology. *Science*, 182, 990 – 1000.
- Swets, J. A. (1979). ROC Analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14, 109 – 121.
- Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science*, 240, 1285 – 1293.
- Zou, K. H., O'malley, A. J., Mauri, L. (2007). Receiver-Operating characteristic analysis for evaluating diagnostic tests and predictive models, *Circulation*, 115, 654 – 657.
- Zweig, M. H., Campbell, G. (1993). Receiver- Operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39 (4), 561 – 577.

Abstract

ROC curve (Receiver Operating Characteristic) is a simple and highly practical applicability graphical method. It is particularly used for comparisons of different tests in different areas of study such as Psychology, Medicine Imaging techniques, Economy and Weather forecasts. This study presents a theoretical material regarding the ROC curve methodology, as well as an application under a different approach than the usual one. A sectional study considering 6963 children was carried out, in which, through the ROC curve, variables associated with risk factors related to infant mortality were identified. The used data referred to live births from mothers resident in the municipalities belonging to the 4th Regional Health RS state in 2006. Through a pattern suggested by Câmara (2009), it was found that the variables: birth weight and the Apgar score are risk factors for death during the first year of life.

Keywords: ROC curve, sensitivity, specificity, infant mortality

Modelagem Estatística para previsão de resultados de Jogos de Futebol: Quem será o Campeão Brasileiro de 2010?

*Adriano K. Suzuki¹
Francisco Louzada²
José F. Francini³
Luis E. B. Salazar⁴*

Resumo

Neste artigo, apresentamos duas metodologias estatísticas aplicadas à previsão de resultados da Série A do Campeonato Brasileiro de futebol de 2010. A previsão da classificação final é obtida por meio de um procedimento de Monte Carlo. Antes do início de cada rodada do retorno, calculamos a probabilidade de cada time vencer o campeonato e ser rebaixado pra a Série B. Além disso, comparamos a qualidade preditiva dos dois métodos usando a medida de de Finetti também e pelo percentual de previsões corretas para três tipos de ponderação de observações passadas: peso total para todos os dados, pesos colocados somente para as 5 últimas rodadas e as 10 últimas rodadas.

Palavras-chaves: Distribuição de Poisson, Modelagem Estatística, Simulação, Futebol.

¹ ICMC - Universidade de São Paulo, São Carlos - SP

² ICMC - Universidade de São Paulo, São Carlos - SP

³ Des – Universidade Federal de São Carlos, São Carlos - SP

⁴ Des – Universidade Federal de São Carlos, São Carlos - SP

1. Introdução

O futebol é considerado o esporte mais popular do mundo, com cerca de 270 milhões de pessoas participando de suas várias competições nacionais e internacionais. A sua principal competição internacional, a Copa do Mundo FIFA, é o evento recordista mundial de espectadores, com o dobro de espectadores dos Jogos Olímpicos. Desde sua chegada ao Brasil no século retrasado, trazido por Charles Miller, se tornou o esporte nacional favorito da maior parte da população.

Este jogo coletivo consiste em uma disputa de 90 minutos entre duas equipes, sendo declarado o vencedor a equipe que marcar o maior número de gols e, pode haver empate quando o número de gols marcados por cada equipe for igual.

Assim como em outros esportes, tanto coletivos quanto individuais, sempre existe o interesse em determinar as chances de uma equipe sagrar-se campeã ou de atleta vencer uma determinada competição. Especificamente no futebol existem diversos trabalhos publicados na literatura abordando os diversos aspectos que podem influenciar o resultado de uma partida e o resultado final das competições. Por exemplo, Barnett e Hilditch (1993), a pedido da Liga Inglesa de Futebol durante os anos 1980-1990, verificaram se os quatro clubes ingleses que colocaram gramados artificiais em seus estádios teriam uma vantagem ao jogarem em casa. Mediram os desempenhos destes quatro times por meio de pontos, gols e resultados das partidas, e concluíram que estes times obtiveram vantagem ao jogarem em casa. Dessa forma, a Liga Inglesa de Futebol proibiu gramados artificiais depois da publicação do artigo de Barnett e Hilditch.

Previsão de resultados de jogos de futebol é de grande interesse, principalmente por parte dos torcedores e imprensa. Vários trabalhos podem ser encontrados na literatura. Dyte e Clarke (2000) apresenta um modelo de regressão Poisson log-linear para as previsões das partidas de futebol aplicadas à Copa do Mundo de futebol de 1998, levando em conta o ranking da FIFA como covariável. Neste artigo, os autores dão alguns resultados sobre o poder preditivo do modelo e apresentam os resultados de simulação para estimar as probabilidades de campeão. Suzuki et al. (2010) propõem uma metodologia Bayesiana para previsão de resultados de partida de futebol. A metodologia proposta é aplicada na Copa do Mundo de 2006. Para tentar suprir a carência de resultados recentes das seleções, os autores utilizam como informações prévias o ranking da FIFA e opiniões de especialistas. O método é aplicado para calcular as probabilidades

de vitória, empate e derrota como também simular uma competição inteira com o objetivo de estimar as probabilidades de classificação na fase de grupo e as chances de cada seleção ser campeã. Levando em consideração uma abordagem diferente, Brillinger (2008) propôs a modelagem direta das probabilidades de vitória, empate e derrota. Neste artigo, Brillinger empregou um modelo trinomial e aplicou-o ao Campeonato Brasileiro 2006 série A para estimar as probabilidades de um particular time ser campeão e ficar entre os quatro primeiros colocados, além das probabilidades para a pontuação final. Karlis e Ntzoufras (2009) aplicaram a distribuição de Skellam para modelar a diferença de gols entre equipes mandantes e visitantes. Os autores argumentam que essa abordagem não leva em consideração nem a suposição de independência, nem que a marginal siga uma distribuição de Poisson para o número de gols marcados pelas equipes. Uma análise Bayesiana para prever os resultados das partidas da Liga Inglesa 2006-2007 foi realizada utilizando uma função de ligação log-linear e distribuições a priori não informativas para os parâmetros. Lee (1997) assume que o número de gols marcados pelo time mandante e visitante são independentes e seguem uma distribuição de Poisson em que a média reflète a força do ataque, defesa e casa.

Neste artigo, seguindo a ideia de Lee (1997), estimamos a média de gols marcados por cada time assumindo que o número de gols marcado por cada time em uma partida segue uma distribuição Poisson univariada, para isto, utilizarmos os métodos de previsão sugeridos por Arruda (2000) e modificados por Suzuki (2007) adicionando uma covariável indicadora de crise. Apresentamos duas metodologias para calcular as probabilidades de interesses em um campeonato de futebol.

Como podemos observar o resultado de uma partida de futebol como um vetor aleatório bivariado (cujas variáveis são: número de gols marcados pelo time mandante e número de gols marcados pelo time visitante) e assim argumentar que o número de gols marcados por uma equipe segue uma distribuição de Poisson univariada. Assim, são estudadas as distribuições de Poisson Bivariadas, com destaque para a classe “de Holgate”. A partir destas distribuições, considerando um procedimento de simulação calculamos a probabilidade de um determinado time se sagrar campeão como também a probabilidade de rebaixamento para a série B.

As metodologias desenvolvidas foram aplicadas no Campeonato Brasileiro Série A de 2010. O Campeonato Brasileiro Série A, popularmente conhecido como Brasileirão, é o principal torneio entre clubes de futebol do Brasil. Atualmente, é organizado pela Confederação Brasileira de Futebol (CBF) e dá acesso aos quatro primeiros colocados à Taça Libertadores da América, que é a competição sulamericana mais importante. A primeira divisão do Brasileirão 2010 foi disputado por 20 clubes em dois turnos. Em cada turno, todos os times jogam entre si uma única vez. Os jogos do segundo turno serão realizados na mesma ordem do primeiro, apenas com o mando de campo invertido. Não há campeões por turnos, sendo declarado campeão brasileiro o time que obtiver o maior número de pontos após as 38 rodadas. Caso haja empate de pontos entre dois clubes, os critérios de desempates serão aplicados na seguinte ordem: número de vitórias, saldo de gols, gols marcados, confronto direto, número de cartões vermelhos e número de cartões amarelos.

Além disso, comparamos a qualidade preditiva dos dois métodos por meio da medida de de Finetti e pela quantidade de acertos considerando como conjunto de dados três tipos de históricos: completo, as 5 últimas rodadas e as 10 últimas rodadas.

O artigo está organizado da seguinte forma: na Seção 2 apresentamos os dois modelos probabilísticos, na Seção 3 apresentamos o procedimento para obter as previsões, na Seção 4 apresentamos os resultados das aplicações no Campeonato Brasileiro 2010 e por fim, na Seção 5 apresentamos alguns comentários finais.

2. Modelos probabilísticos

Para uma dada partida de futebol, considere X e Y o número de gols marcados pelo time mandante e pelo time visitante, respectivamente. Dessa forma, assumimos que X e Y são variáveis aleatórias que seguem distribuição Poisson com médias λ_X e λ_Y , respectivamente.

2.1 Método 1

Este método, definido por Arruda (2000) como Método SDO e modificado por Suzuki (2007) ao adicionar uma covariável indicadora de crise, assume que X e Y são independentes. Pela propriedade linear da esperanças, temos que

$$\lambda_x = E[X] = \frac{E[X+Y] + E[X-Y]}{2} \quad (1)$$

$$\lambda_y = E[Y] = \frac{E[X+Y] - E[X-Y]}{2} \quad (2)$$

Ou seja, λ_X e λ_Y podem ser expressadas como funções das médias das variáveis $X+Y$ e $X-Y$. Baseado nas identidades (1) e (2), consideramos o vetor aleatório (U,V) definido por $U=X+Y$ e $X-Y$ como variável resposta ao invés do vetor (X,Y) de número de gols marcados por cada time. Portanto, a estimação dos parâmetros λ_X e λ_Y podem ser obtidas baseada na estimação para as médias $\mu = E[U]$ e $v = E[V]$ usando $\widehat{\lambda}_X = (\hat{\mu} + \hat{v})/2$ e $\widehat{\lambda}_Y = (\hat{\mu} - \hat{v})/2$. Para cada uma das futuras partidas da competição, $\hat{\mu}$ e $\hat{v}$ serão obtidas por meio de dois modelos lineares ajustados definidos a seguir.

Suponha que os times são categorizados como $1, 2, \dots, T$ e que o conjunto de dados disponível consiste das n primeiras partidas da competição. Para a i -ésima partida entre o time mandante $h[i]$ contra o time visitante $a[i]$, considere que U_i e V_i são a soma e a diferença de gols entre o time mandante e o visitante, respectivamente, em que $h[i]$ e $a[i] \in \{1, 2, \dots, T\}$, $i=1, \dots, n$. Os dois modelos lineares são definidos por

$$U_i = \mu_i + \varepsilon_i \quad (3)$$

$$V_i = v_i + \varepsilon'_i, \quad (4)$$

$i=1, \dots, n$ em que $\varepsilon_1, \dots, \varepsilon_n$ e $\varepsilon'_1, \dots, \varepsilon'_n$ são erros independentes com média 0.

As médias μ_i e v_i de U_i e V_i são dadas pelas funções de ligação

$$\mu_i = aI_{casa[i]} + \beta_{h[i]} + \psi_{a[i]} + \eta I_{crise[i]} \quad (5)$$

$$v_i = a'I_{casa[i]} + \beta'_{h[i]} + \psi'_{a[i]} + \eta' I_{crise[i]} \quad (6)$$

em que a e a' se referem ao efeito mando de campo, $\beta_{h[i]}$ e $\beta'_{h[i]}$ o poder ofensivo do time $h[i]$, $\psi_{a[i]}$ e $\psi'_{a[i]}$ o poder defensivo do time $a[i]$, η e η' o efeito crise, $I_{casa[i]}$ é uma variável indicadora que assume 1 caso um dos times tem o mando de campo e 0 caso contrário e $I_{crise[i]}$ é uma variável indicadora que assume 1 caso um dos times está em crise e 0 caso contrário, $i = 1, \dots, n$. Note que cada um dos modelos lineares (5) e (6) tem 42 parâmetros (20 efeitos ofensivos, 20 efeitos defensivos, 1 efeito casa e 1 efeito crise) a serem estimados.

Os modelos definidos por (3) e (4) são ajustados separadamente via estimação por mínimos quadrados assumindo pesos iguais usando a matriz inversa generalizada de Moore-Penrose (Penrose, 1995). A partir das estimativas de mínimos quadrados (a, β, ψ, η) e $(a', \beta', \psi', \eta')$ obtida, em que $\beta = (\beta_1, \beta_2, \dots, \beta_{20})$, $\psi = (\psi_1, \psi_2, \dots, \psi_{20})$, $\beta' = (\beta'_1, \beta'_2, \dots, \beta'_{20})$ e $\psi' = (\psi'_1, \psi'_2, \dots, \psi'_{20})$, obtemos as estimativas de $\hat{\mu}$ e $\hat{v}$ para cada partida futura e estimamos λ_X e λ_Y o número médio de gols para cada uma destas partidas.

2.2 Método 2

Definido por Arruda (2000) como Método SD1 e modificado por Suzuki (2007) ao adicionar uma covariável indicadora de crise, este método considera que a covariância de X e Y seja não nula, dessa forma temos outro método de previsões análogo ao visto anteriormente cujas variáveis X e Y seguem distribuição Poisson com médias λ_X e λ_Y , respectivamente.

Segue da propriedade linear das esperanças que:

$$\lambda_X = \frac{E[X - Y] + 2E[X + Y] - \{E[(X + Y)^2] - (E[X + Y])^2\}}{2} \quad (7)$$

$$\lambda_Y = \frac{2E[X + Y] - E[X - Y] - \{E[(X + Y)^2] - (E[X + Y])^2\}}{2} \quad (8)$$

$$\lambda_{XY} = \frac{\{E[(X + Y)^2] - (E[X + Y])^2\} - E(X + Y)}{2} \quad (9)$$

em que λ_X e λ_Y continuam a ser expressadas como funções das médias das variáveis $X+Y$ e $X-Y$, e λ_{XY} é função da média da variável $(X + Y)^2$. As estimativas das médias, podem ser obtidas de forma análoga ao método anterior.

3. Previsão

Nesta seção descrevemos os procedimentos realizados para o cálculo das probabilidades de interesse (campeão e rebaixamento) para o campeonato inteiro. As previsões foram realizadas para os jogos do retorno (a partir da rodada 20). Note que podemos também calcular as probabilidades de vitória (P_V), empate (P_E) e derrota (P_D) a partir dos parâmetros estimados por meio de um dos métodos, usando as seguintes fórmulas

$$P_V = P(X_{AB} > X_{BA}) = \sum_{i=1}^{\infty} \sum_{j=0}^{i-1} P(X_{AB} = i)P(X_{BA} = j) \quad (10)$$

$$P_E = P(X_{AB} = X_{BA}) = \sum_{i=0}^{\infty} P(X_{AB} = i)P(X_{BA} = i) \quad (11)$$

$$P_D = P(X_{AB} < X_{BA}) = \sum_{j=1}^{\infty} \sum_{i=0}^{j-1} P(X_{AB} = i)P(X_{BA} = j) \quad (12)$$

Para uma dada rodada fixa i , $i=20,\dots,38$, temos a atual tabela de classificação dos times (jogos, número de pontos, número de vitórias, número de empate, número de derrota, número de gols marcados, número de gols sofridos e saldo de gols). A previsão para a classificação final foi realizada a partir de um algoritmo de simulação que envolve as seguintes etapas

- 1) Considere r o número de campeonatos;
- 2) Para o r -ésimo campeonato, simule o placar (X_{AB}, X_{BA}) para cada um dos n jogos entre os times mandante (A) e visitante (B) a serem jogados, $n=[38-(r-1)]*10$;
- 3) Para cada um dos n jogos preditos, verifica-se se houve vitória do mandante $(X_{AB} > X_{BA})$, empate $(X_{AB} = X_{BA})$ ou vitória do visitante $X_{AB} < X_{BA}$. Atribua 3 pontos para o time vitorioso e 1 ponto para ambos os times se ocorreu empate;
- 4) A partir da atual tabela de classificação, construir a tabela final de classificação com os resultados simulados para o r -ésimo campeonato;

Nesse artigo consideramos $r=10.000$ campeonatos simulados. A partir daí podemos calcular a chance de um determinado time ser campeão e de ser rebaixado da seguinte forma:

$$P_{\text{campeão}} = \frac{\#(\text{time terminou em primeiro lugar})}{r},$$

$$P_{\text{rebaixamento}} = \frac{\#(\text{time terminou entre os últimos colocados})}{r},$$

em que $\#$ refere-se ao número de vezes obtidos na simulação. Uma observação importante que deve ser mencionada é que como estamos realizando um número finito de simulações, um evento com probabilidade não nula mas muito pequena (ou muito grande diferente de 1) de ocorrer pode-se acontecer que a frequência observada seja zero (ou um).

Analogamente à Suzuki et. al. (2010) e Bastos e Rosa (2013), utilizamos duas métricas para comparação das qualidades das previsões: a medida de de Finetti (DeFinetti, 1972) e a quantidade do número de acertos. A medida de de Finetti consiste na consideração de um simplex contido em R^3 como representação geométrica do conjunto de possíveis previsões probabilísticas. Assim, os vértices desse simplex correspondem às ocorrências dos resultados e os demais pontos a todas as outras possíveis previsões. Formalmente, $S = \{(P_V; P_E; P_D) \in R^3 \mid P_V + P_E + P_D = 1, P_V \geq 0, P_E \geq 0, P_D \geq 0\}$, em que P_V denota a probabilidade de vitória do time mandante, P_E a probabilidade de empate e P_D a probabilidade de derrota do time mandante. A medida de distância de de Finetti corresponde à distância euclidiana quadrática entre o ponto correspondente à probabilidade prevista e o vértice correspondente ao resultado efetivamente observado. Para mais de uma previsão, pode-se construir um índice dado pela média aritmética das distâncias de de Finetti chamado Medida de de Finetti.

Na nossa aplicação no futebol, associam-se os vértices $(0, 1, 0)$, $(0, 1, 0)$ e $(0, 0, 1)$ à vitória da equipe mandante, ao empate e à derrota da equipe mandante, respectivamente. Ao vetor de probabilidades atribuídas para uma determinada partida associa-se o ponto $(P_V, P_E, P_D) \in S$. Assim, a distância de de Finetti é igual a:

$(P_V - 1)^2 + (P_E - 0)^2 + (P_D - 0)^2$ se a equipe mandante vencer a partida;

$(P_V - 0)^2 + (P_E - 1)^2 + (P_D - 0)^2$ se a partida terminar empatada;

$(P_V - 0)^2 + (P_E - 0)^2 + (P_D - 1)^2$ se a equipe mandante perder a partida. Por exemplo, se a

previsão para um determinado jogo for $(0,2, 0,65, 0,15)$ e o resultado observado for um empate $(0, 1, 0)$, então a distância de de Finetti será

$$(0,2 - 0)^2 + (0,65 - 1)^2 + (0,15 - 0)^2 = 0,185.$$

Com relação à quantidade de número de acertos, consideramos que um método “acerta” o resultado de um determinado jogo quando o evento com maior probabilidade estimada é observado, por exemplo, se em um determinado jogo o time mandante vencer e a probabilidade estimada para a vitória do time mandante for maior que as probabilidades de empate e derrota, consideraremos um acerto para esta previsão, caso contrário um erro. Da mesma forma para a ocorrência de empate e vitória do time visitante. No entanto,

vale ressaltar que não é correto associar a eventos com alta probabilidade uma certeza de ocorrência ou a eventos com baixa probabilidade uma certeza de não ocorrência. Aqui, o objetivo não é afirmar que uma dada previsão é correta ou não e sim construir uma métrica para o conjunto das previsões. Como analogia considere um apostador que sempre aposta no evento com maior probabilidade. Ao final de cada partida, ele terá observado um acerto ou um erro para sua previsão.

4. Aplicação: Campeonato Brasileiro de 2010

Nesta seção apresentamos os resultados obtidos quando da aplicação da metodologia desenvolvida ao Campeonato Brasileiro Série A de 2010.

As previsões para partidas simples por meio dos Métodos 1 e 2 para todas as rodadas i , $i = 20, \dots, 38$, foram feitas utilizando como conjunto de dados três tipos de históricos: 1) completo, considerando todas as partidas realizadas no campeonato, 2) 5 últimas rodadas, considerando para as previsões somente os resultados das 5 rodadas anteriores e 3) 10 últimas rodadas, considerando apenas os resultados das 10 rodadas anteriores. A ideia é verificar se ao considerar apenas resultados mais recentes, os métodos aplicados poderiam apresentar uma melhor qualidade preditiva. Foi motivado pelo fato das mudanças nos times ao longo do campeonato que o modelo não leva em consideração, tais como: saída e chegada de jogadores dos times com a abertura da janela de transferência internacional, mudança de treinadores etc. Para exemplificar, nas Tabelas 1 e 2 apresentamos a previsão para a rodada 37 com histórico completo para os Métodos 1 e 2, respectivamente. Nesta rodada ambos os métodos acertaram 7 resultados e as medidas de de Finetti associadas foram 0,4976 e 0,5153, respectivamente.

Tabela 1: Tabela de previsão da 37ª rodada para o Método 1 com histórico completo.

Previsões para partidas simples (37ª Rodada)						
Time Mandante	Placar Observado	Time Visitante	Vitória do Mandante (%)	Empate (%)	Derrota do Mandante (%)	Acerto
Atlético-GO	1 – 1	São Paulo	50,71	22,28	27,00	Não
Internacional	1 – 1	Vitória	59,20	23,33	17,45	Não
Corinthians	2 – 0	Vasco	66,71	22,11	11,17	Sim
Palmeiras	1 – 2	Fluminense	22,14	26,73	51,12	Sim
Botafogo	3 – 1	Prudente	73,54	17,67	8,77	Sim
Avaí	3 – 2	Santos	40,94	22,50	36,54	Sim
Atlético-MG	3 – 1	Goiás	56,64	20,05	23,30	Sim
Ceará	1 – 1	Atlético-PR	43,91	31,82	24,26	Não
Guarani	0 – 3	Grêmio	32,32	25,12	42,54	Sim
Flamengo	1 – 2	Cruzeiro	22,71	28,98	48,30	Sim

Tabela 2: Tabela de previsão da 37ª rodada para o Método 2 com histórico completo.

Previsões para partidas simples (37ª Rodada)						
Time Mandante	Placar Observado	Time Visitante	Vitória do Mandante (%)	Empate (%)	Derrota do Mandante (%)	Acerto
Atlético-GO	1 – 1	São Paulo	50,82	17,72	31,46	Não
Internacional	1 – 1	Vitória	59,93	27,10	12,97	Não
Corinthians	2 – 0	Vasco	66,83	22,28	10,89	Sim
Palmeiras	1 – 2	Fluminense	24,19	24,51	51,30	Sim
Botafogo	3 – 1	Prudente	71,91	16,83	11,26	Sim
Avaí	3 – 2	Santos	39,86	25,07	35,07	Sim
Atlético-MG	3 – 1	Goiás	56,06	17,95	25,99	Sim
Ceará	1 – 1	Atlético-PR	45,48	26,13	28,39	Não
Guarani	0 – 3	Grêmio	33,44	23,45	43,11	Sim
Flamengo	1 – 2	Cruzeiro	23,26	28,31	48,43	Sim

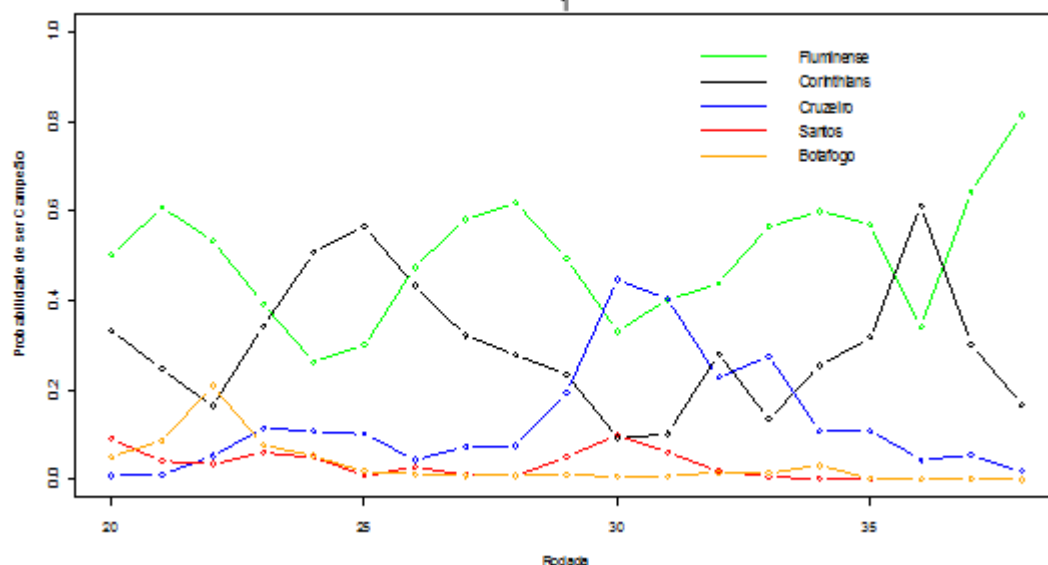
Verificamos na Tabela 3 a porcentagem de acertos para cada método, considerando os diferentes tipos de conjunto de dados históricos. Observamos que os métodos com maiores porcentagem de acertos são aqueles que consideraram o histórico completo, sendo que o Método 1 foi o melhor entre eles. As medidas de de Finetti associadas às previsões com histórico completo para os Métodos 1 e 2 foram 0,6271 e 0,6408, respectivamente. Como resultado obtivemos melhor resultado preditivo em ambas as métrica com o uso do Método 1 e considerando o histórico completo. Assim, os resultados apresentados a seguir são referentes ao Método 1 considerando o histórico completo.

Tabela 3: Tabela de porcentagem de acertos dos Métodos.

Histórico	Método	Acerto
Completo	1	52%
Completo	2	49%
5 últimas rodadas	1	45%
5 últimas rodadas	2	37%
10 últimas rodadas	1	43,75%
10 últimas rodadas	2	45%

A Figura 1 ilustra as chances das 5 equipes (Fluminense, Corinthians, Cruzeiro, Santos e Botafogo) se tornarem campeãs rodada a rodada.

Figura 1: Chance de campeão.



Podemos notar na Figura 1 que, a partir da rodada 20, Fluminense e Corinthians apareciam como candidatos mais prováveis ao título de campeão brasileiro e que foi a partir da rodada 29 que o Cruzeiro também entrou nessa disputa. Dessa forma, nota-se que como tínhamos bons indicativos dos primeiros colocados no campeonato com um número razoável de rodadas de antecedência, onde pode-se observar a classificação final do Campeonato Brasileiro de 2010 na Tabela 4.

Tabela 4: Classificação final do Campeonato Brasileiro 2010 de futebol.

Posição	Times	PG	J	V	E	D	GP	GC	SG
1	Fluminense	71	38	20	11	7	62	36	26
2	Cruzeiro	69	38	20	9	9	53	38	15
3	Corinthians	68	38	19	11	8	65	41	24
4	Grêmio	63	38	17	12	9	68	43	25
5	Atlético-PR	60	38	17	9	12	43	45	-2
6	Botafogo	59	38	14	17	7	54	42	12
7	Internacional	58	38	16	10	12	48	41	7
8	Santos	56	38	15	11	12	63	50	13
9	São Paulo	55	38	15	10	13	54	54	0
10	Palmeiras	50	38	12	14	12	42	43	-1
11	Vasco	49	38	11	16	11	43	45	-2
12	Ceará	47	38	10	17	11	35	44	-9
13	Atlético-MG	45	38	13	6	19	52	64	-12
14	Flamengo	44	38	9	17	12	41	44	-3
15	Avaí	43	38	11	10	17	49	58	-9
16	Atlético-GO	42	38	11	9	18	51	57	-6
17	Vitória	42	38	9	15	14	42	48	-6
18	Guarani	37	38	8	13	17	33	53	-20
19	Goiás	33	38	8	9	21	41	68	-27
20	Grêmio Prudente*	28	38	7	10	21	39	64	-25

PG: Pontos Ganhos, J: Jogos, V: Vitórias, E: Empates, D: Derrotas, GP: Gols Pró, GC: Gols Contra, SG: Saldo de Gols.

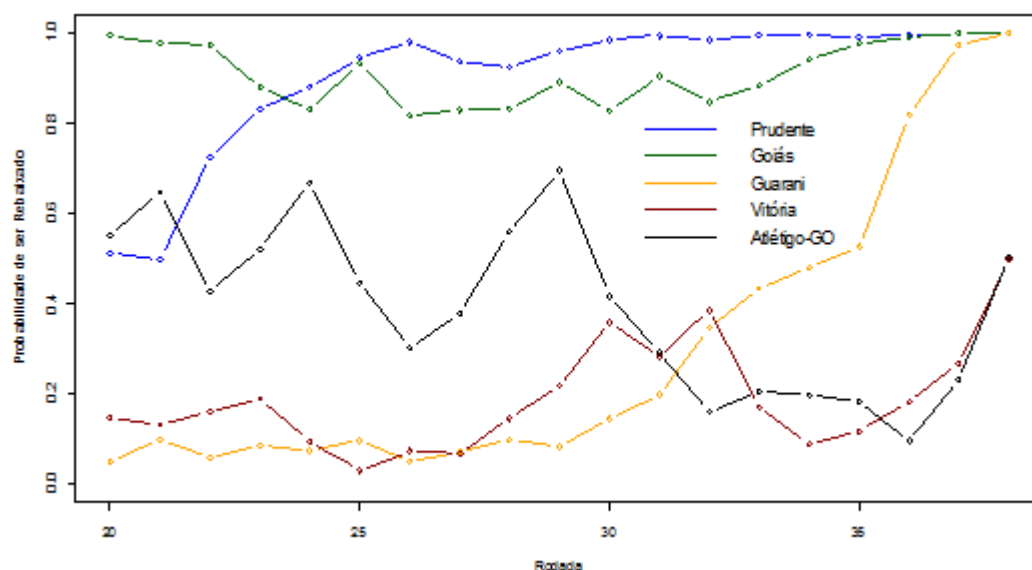
*** O Grêmio Prudente foi punido por escalar um jogador irregular e perdeu 3 pontos.**

Temos ainda ilustrado na Figura 2 a chance de rebaixamento rodada a rodada das 5 piores equipes, de acordo com a Tabela 4 (Grêmio Prudente, Goiás, Guarani, Vitória e Atlético-GO). Podemos notar uma visível queda de rendimento do Guarani a partir da rodada 30 e, a cada rodada, uma extensiva “luta” contra o rebaixamento do Vitória e do Atlético-GO, chegando à última rodada com chances de descenso próxima a 50%.

Podemos observar na Figura 3 que o Corinthians e o Fluminense são os times que, em média, obtiveram melhores pontuações finais. Nesta rodada, ainda faltando 14 rodadas para o término do campeonato, podemos notar que os times: Fluminense, Cruzeiro e Corinthians (os times que terminaram nas três primeiras colocações, ver Tabela 4) são os times que apresentam, para esta rodada, os melhores desempenhos em nossas previsões.

Na rodada 33, faltam 6 decisivas rodadas a serem jogadas, os times podem ser dividido em três grupos: no primeiro, os times: Fluminense, Corinthians, Cruzeiro, Botafogo, Santos, Internacional, São Paulo, Atlético-PR e Grêmio que estão “lutando” pela conquista do título; no segundo, os times: Grêmio Prudente, Goiás, Avaí, Atlético-MG, Guarani, Atlético-GO, Vitória e Flamengo que lutam pela permanência na elite do futebol brasileiro e no terceiro, os times: Palmeiras, Vasco e Ceará que lutam pela classificação para a Taça Libertadores de América ou para a Copa Sul-Americana.

Figura 2: Chance de rebaixamento.



À medida que o campeonato se aproxima do final ocorrem times que não pertencem mais a um dos três grupos, ou seja, ou o time já garantiu sua classificação e não tem mais chances de ganhar o campeonato, ou já garantiu a permanência na primeira divisão e não tem mais chance de classificação para os torneios sul-americanos ou o time já está rebaixado (nota-se que poderia ter a ocorrência de um time ter sido campeão antes da última rodada).

Para os times que disputam algo no campeonato, os jogos vão “se tornando mais decisivos”, “aumentando a importância” dos times para obter resultados positivos. Vale ressaltar que, principalmente na última rodada, ocorreram times que fizeram “corpo mole”, jogando com os times reservas, “prejudicando” as equipes que dependiam dos resultados positivos (vitória) destes times.

A Figura 4 refere-se ao box-plot da pontuação final previsto para a rodada 33. Nota-se o favoritismo do Fluminense (time que se sagrou-se campeão, ver Tabela 4) e, apesar que nesta rodada o Corinthians tem 3 pontos a mais que o Cruzeiro (uma boa vantagem se consideramos que faltam somente 18 pontos a serem disputados), ambos têm praticamente a mesma chance de ser campeão (Cruzeiro com um leve favoritismo). Na Figura 5, apresentamos o box-plot para os pontos finais previstos antes da rodada 37, no qual nota-se uma disputa acirrada pelo título entre Fluminense e Corinthians, e na outra ponta da tabela, Grêmio Prudente e Goiás estão praticamente rebaixados.

Figura 3: Box-plot para a pontuação final prevista para a rodada 25.

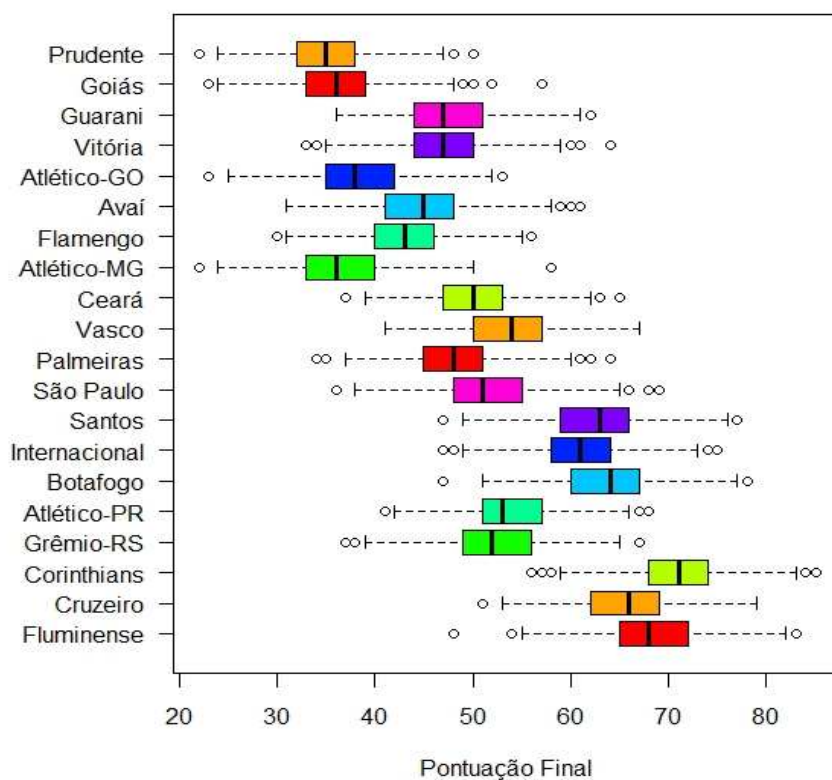


Figura 4: Box-plot para a pontuação final prevista para a rodada 33.

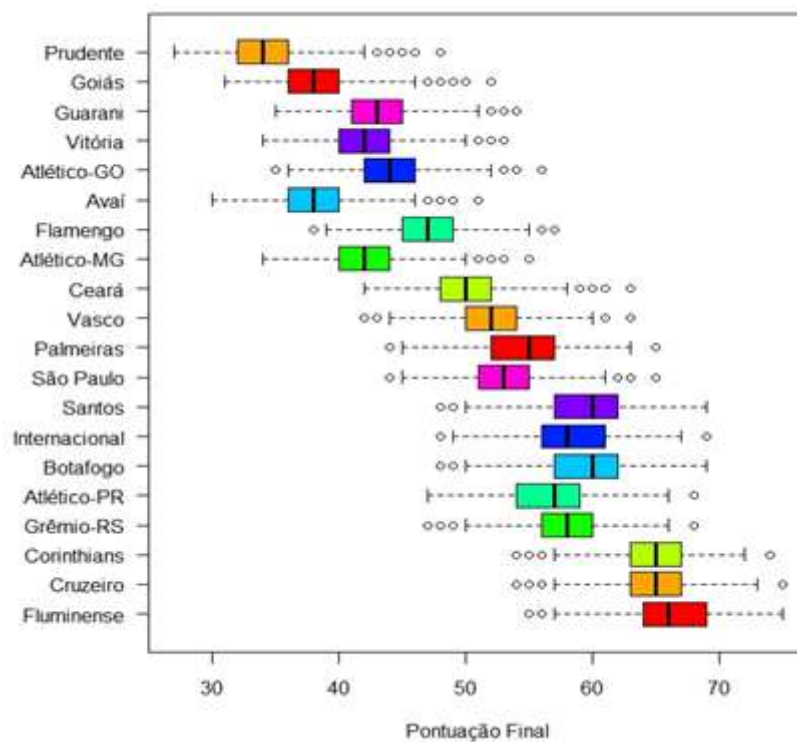


Figura 5: Box-plot para a pontuação final prevista para a rodada 37.

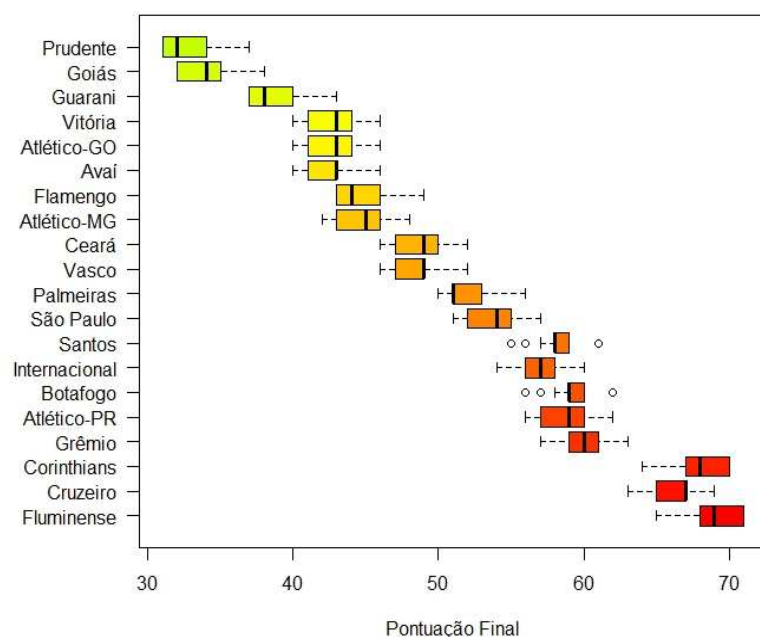


Figura 6: Box-plot para o número de vitórias final prevista para a rodada 33.

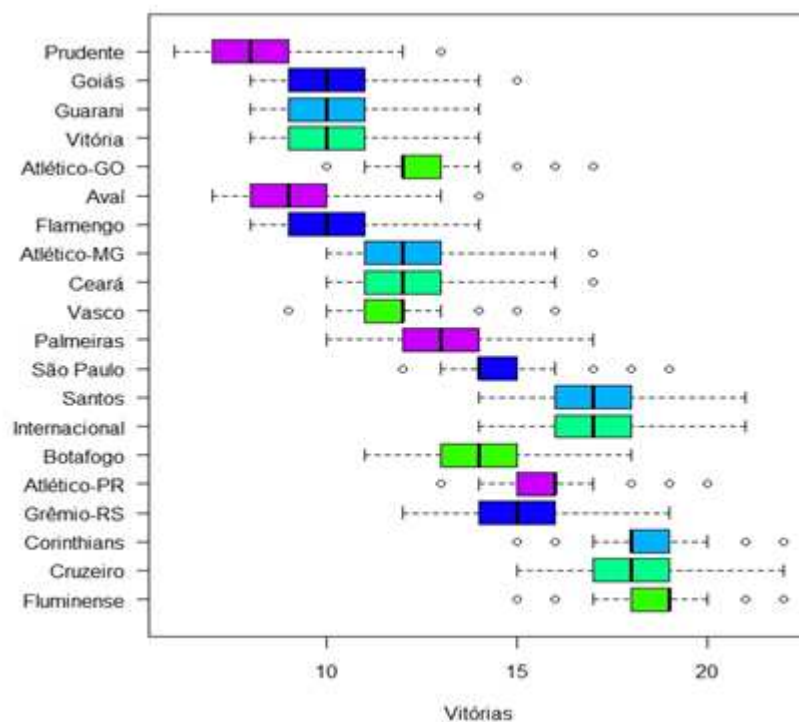
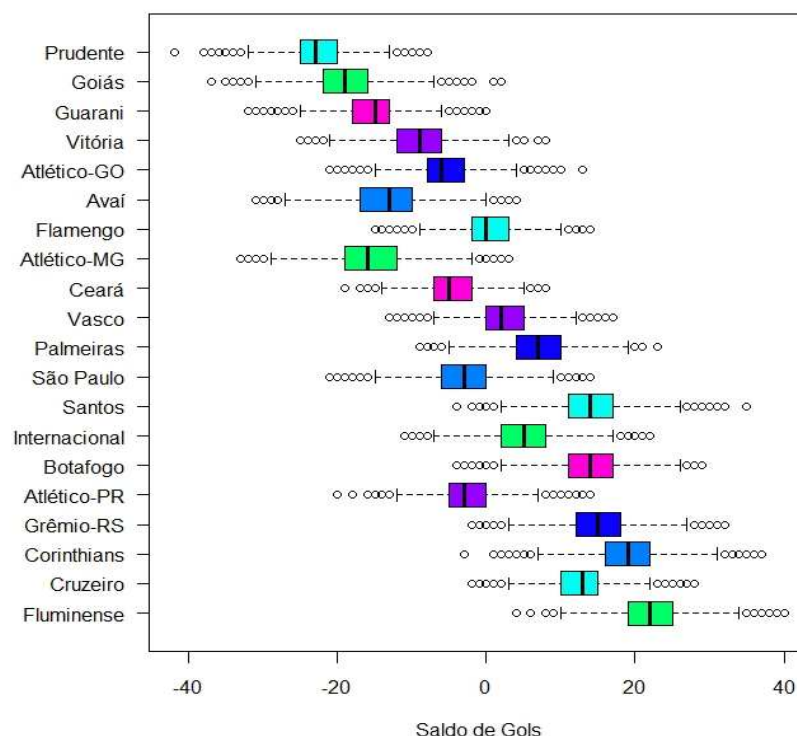


Figura 7: Box-plot para o saldo de gols final previsto para a rodada 33.



Outras informações de interesse, tais como o número de vitórias e o saldo de gols, que são os primeiros critérios de desempate caso dois times terminam com a mesma pontuação final. Nas figuras 6 e 7 apresentamos o box-plot para o número de vitórias final e saldo de gols final para a rodada 33, respectivamente. Em comparação com o resultado final apresentado na Tabela 4 pode-se observar, por exemplo, que na Figura 6, com exceção do último colocado Grêmio Prudente, há um equilíbrio entre o número de vitórias os demais times rebaixados. Já na Figura 7 observa-se um menor saldo de gols do Atlético-PR e São Paulo com relação aos demais concorrentes diretos.

5. Comentários Finais

Neste artigo, consideramos dois métodos simples, que apresentaram boas capacidades preditivas de acordo com a medida de de Finetti e a quantidade de acertos. Aliados a um procedimento de simulação, pode-se calcular a chance de uma equipe ser campeã ou ser rebaixada, bem como, a previsão do número de pontos e o saldo de gols ao término do campeonato a partir de uma rodada específica. Outras probabilidades de interesse também podem ser obtidas, tais como: Qual a chance de classificação para a Copa Libertadores da América ou a Copa Sul-Americana? Qual a chance de um determinado time ter o melhor ataque e a melhor defesa?, entre outras. As metodologias apresentadas foram implementadas no *software* R (Development Core Team, 2007).

Referências bibliográficas

- Arruda, M.L. (2000). Poisson, Bayes, Futebol e DeFinetti, Dissertação de Mestrado, IME-USP.
- Bastos, L.S.; Rosa, J. M. C. (2013). Predicting probabilities for the 2010 FIFA World Cup games using a Poisson-Gamma model. *Journal of Applied Statistics*, v. 40, p.1-12.
- Barnett, V. e Hilditch, S. (1993). The effect of an artificial pitch surface on home team performances in football (soccer). *Journal of the Royal Statistical Society A*, 1956, 39-50.
- Brillinger, D. R. (2008). Modeling games outcomes of the Brazilian 2006 series a championship as ordinal-valued. *Brazilian Journal of Probability and Statistics*, 22, 89 – 104.
- Definetti, B. (1072). *Probability, Induction and Statistics*. London, Wiley.
- Dyte, D. e Clarke, S.R. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society*, 51, 993 – 998.
- Karlis, D. e I. Ntzoufras (2009). Bayesian modeling of football outcomes: Using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20, 133 – 145.
- Lee, A. (1997). Modeling scores in the Premier League: Is Manchester United really the best? *Chance*, 10, 15 – 19.
- Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51, 406 – 413.
- R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Suzuki, A.K., (2007). *Modelagem Estatística para Determinação de Resultados de Dados Esportivos*, Dissertação de Mestrado, UFSCar.
- Suzuki, A.K., Salasar, L.E.B., Leite, J.G., Louzada-Neto, F. (2010). A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup. *Journal of the Operational Research Society*, 61(10), 1530-1539 (10).

Agradecimento

A pesquisa é financiada parcialmente pela CAPES e pelo CNPQ.

Abstract

In this paper we present two statistical methodologies directed to predict the outcomes of the 2010 Brazilian National Soccer League Series A. The predictions for the final standings are obtained by means of a Monte Carlo procedure. Just before each round of the second stage, we calculated for each team the probabilities of championship winning and relegation to Series B. Furthermore, we compared the predictive quality of the methods by using the de Finetti measure and the percentage of correct forecasts for three types of weighting of previous observations: full weight to all data, weight only for the last 5 rounds and for the last 10 rounds.

Keywords: Poisson Distribution, Statistical Modeling, Simulation, Football.

Comparação de modelos CAViaR e CARE para mensuração de risco

*Francielle de Lima e Silva¹
Clélia Maria de Castro Tolo²*

Resumo

O Valor em Risco (VaR) é uma medida de risco muito utilizada para gerenciamento e controle de risco de mercado. A mensuração pode ser feita por vários métodos como a abordagem econométrica e a Teoria dos Valores Extremos (TVE). Porém, esses métodos apresentam algumas fragilidades, como por exemplo a suposição inadequada da distribuição dos erros, resultando em medidas imprecisas. Sendo assim, neste artigo são apresentados, discutidos e comparados os modelos de Valor em Risco Autorregressivo Condicional (CAViaR) e os modelos de Expectil Autorregressivo Condicional (CARE). Os modelos foram ajustados para a mensuração do Valor em Risco (VaR) e Expected Shortfall (ES) feitos para a série do IBOVESPA.

Palavras-chaves: Valor em Risco, Expected Shortfall, Modelos CAViaR, Modelos CARE.

¹ Departamento de Estatística - Universidade de São Paulo, Brasil. E-mail: franciellelimasilva@gmail.com

² Departamento de Estatística - Universidade de São Paulo, Brasil. E-mail: clelia@ime.usp.br

1. Introdução

No mercado financeiro, existem três tipos de risco aos quais as empresas estão expostas: operacional, estratégico e financeiro. Este último pode ainda ser subdividido em: operacional, de crédito, de liquidez, legal e de mercado.

O artigo apresenta somente medidas de riscos financeiros de mercado, que estão associadas às possíveis perdas ou ganhos potenciais de investimentos (posições financeiras) de acordo com o comportamento do mercado.

Os investimentos citados são referências feitas às posições financeiras definidas como *vendidas* ou *compradas*.

Na seção 2 são detalhados e apresentados os modelos de Valor em Risco Autorregressivo Condicional (CAViaR) propostos por Engle e Manganelli (2004), que ao invés de modelar toda a distribuição dos retornos, modela diretamente seus quantis. Os modelos Expectil Autorregressivo Condicional (CARE), propostos por Taylor (2008), que também são utilizados para a estimação do VaR, podem ser vistos na seção 3. Esta seção também mostra as novas especificações que visam discutir melhor as propriedades dos modelos CARE, que foram propostas por Kuan, Yeh e Hsu (2009). Na seção 4 é mostrado o Expected Shortfall (ES), uma medida de risco coerente que apresenta informações sobre os retornos que estão acima ou abaixo do VaR. Na seção 5 são feitas aplicações a série de Índices da Bolsa de Valores de São Paulo (IBOVESPA), considerando todos os métodos descritos para a mensuração do VaR e do ES.

2. CAViaR - Valor em Risco Autorregressivo Condicional

Séries de retornos financeiros muitas vezes apresentam autocorrelação, logo o Valor em Risco (VaR) pode apresentar um comportamento semelhante. Assim, com uma estrutura similar aos modelos GARCH (Generalized auto-regressive conditional heteroscedasticity), os modelos de Valor em Risco Autorregressivo Condicional (CAViaR), através de termos autorregressivos, incorporam as autorrelações do VaR ao modelo.

De acordo com Kuan, Yeh e Hsu (2009), a ideia básica é modelar o θ – quantil condicional $Q_\theta(r_t|x_t) = VaR_t(\theta)$ como uma função $g(\cdot)$ das informações $x_t \in F_{t-1}$ (informações até o instante $(t-1)$), e do vetor de parâmetros β , ou seja,

$$VaR_t(\theta) \equiv g_\theta(x_t, \beta_\theta),$$

em que $r_t, \{t: 1, \dots, T\}$, são os log-retornos, θ a probabilidade associada ao VaR e β e o vetor de parâmetros dependente de θ .

Os modelos CAViaR possuem especificações distintas, que são utilizadas de acordo com o comportamento dos retornos. Esses modelos não fazem suposições sobre a distribuição dos retornos.

Uma forma mais geral da especificação do modelo é:

$$VaR_t(\theta) = \beta_0 + \sum_{i=1}^p \beta_i VaR_{t-i}(\theta) + g(\beta_{p+1}, \dots, \beta_{p+q}, F_{t-1})$$

Por simplicidade, assim como nos modelos GARCH, a especificação mais utilizada é a de ordem 1:

$$VaR_t(\theta) = \beta_0 + \beta_1 VaR_{t-1}(\theta) + g(\beta_2, r_{t-1}, VaR_{t-1}(\theta)),$$

em que $g(\dots)$ é a função de ligação entre $VaR_{t-1}(\theta)$ e r_{t-1} , que mede o quanto o VaR pode variar no tempo baseado em novas informações dos log-retornos, como as curvas de impacto dos modelos GARCH, com parâmetros apresentando valores entre $(-1, 1)$.

A seguir são apresentadas algumas formulações específicas:

2.1 Especificações do Modelo

- **Valor Absoluto Simétrico:**

$$VaR_t(\theta) = \beta_1 + \beta_2 VaR_{t-1}(\theta) + \beta_3 |r_{t-1}| \quad (1)$$

- **Inclinação Assimétrica:**

$$VaR_t(\theta) = \beta_1 + \beta_2 VaR_{t-1}(\theta) + \beta_3 r_{t-1} I(r_{t-1} > 0) + \beta_4 r_{t-1} I(r_{t-1} < 0), \quad (2)$$

em que $I(\cdot)$ representa uma função indicadora.

- **GARCH (1,1) Indireto:**

$$VaR_t(\theta) = (\beta_1 + \beta_2 VaR_{t-1}^2(\theta) + \beta_3 r_{t-1}^2)^{\frac{1}{2}} \quad (3)$$

- **Adaptativo:**

$$VaR_t(\theta) = VaR_{t-1}(\theta) + \beta_1 \left\{ \left[1 + \exp \left[k \left(r_{t-1} - VaR_{t-1}(\theta) \right) \right] \right]^{-1} - \theta \right\} \quad (4)$$

- **Adaptativo Simétrico Proporcional:**

$$VaR_t(\theta) = VaR_{t-1}(\theta) + \beta_1 \max \{ 0, |r_{t-1}| - VaR_{t-1}(\theta) \} - \beta_2 \min \{ 0, |r_{t-1}| - VaR_{t-1}(\theta) \}. \quad (5)$$

- **Valor Absoluto Assimétrico:**

$$VaR_t(\theta) = \beta_1 + \beta_2 VaR_{t-1}(\theta) + \beta_3 |r_{t-1} - \beta_4|. \quad (6)$$

- **AR(1)-GARCH (1,1) indireto:**

Pensando em um modelo que possa capturar a autocorrelação da média condicional da série de retornos, Kuester, Mittnik e Paoletta (2006) estenderam a estrutura do modelo **GARCH Indireto**, propondo:

$$VaR_t(\theta) = \phi_1 r_{t-1} + \left[\beta_1 + \beta_2 (VaR_{t-1}(\theta) - \phi_1 r_{t-2})^2 + \beta_3 (r_{t-1} - \phi_1 r_{t-2})^2 \right]^{\frac{1}{2}}. \quad (7)$$

2.2 Estimação do Modelo

Introduzida por Koenker e Bassett (1978), a técnica de Regressão Quantílica permite caracterizar toda a distribuição condicional de uma variável resposta a partir de um conjunto de regressores.

Seja a v.a. Y com função de distribuição $F(y)$ em que para algum $0 < \theta < 1$,

$$F^{-1}(\theta) = \inf \{y \in \mathfrak{R} : F(y) \geq \theta\} \quad (8)$$

é chamado θ – ésimo quantil de Y .

Agora considere um simples problema de teoria da decisão com a função de perda convexa descrita como a seguinte função linear:

$$p_{\theta}(u) = \begin{cases} u(\theta - 1), & \text{se } u < 0 \\ u\theta, & \text{se } u > 0 \end{cases}, \quad (9)$$

em que $u = (Y - b)$ é o desvio, $b \in \mathfrak{R}$.

Fixando um valor θ , é possível calcular a média da função e perda como sendo

$$E = \{\rho_{\theta}(u)\} = E \{\rho_{\theta}(Y - b)\} = (0 - 1) \int_{-\infty}^b (y - b) dF(y) + \theta \int_b^{+\infty} (y - b) dF(y).$$

Com o objetivo de encontrar o menor desvio, calcula-se o valor de b que minimiza a média da função de perda, dada em (9), ou seja, basta diferenciar $E \{\rho_{\theta}(Y - b)\}$ com relação a b e igualar a zero, assim:

$$(1 - \theta) \int_{-\infty}^b dF(y) - \theta \int_b^{+\infty} dF(y) = 0 \Rightarrow F(b) - \theta = 0 \Rightarrow F(b) = \theta.$$

Desde que F seja monótona, algum elemento de $\{y \in \Re : F(y) = \theta\}$ minimiza a perda esperada. Quando a solução é única,

$$b = F^{-1}(\theta), \quad (10)$$

caso contrário, tem-se um intervalo para b , no qual o menor elemento deve ser escolhido.

Note que por (8) e (10), a divisão da amostra em quantis pode ser feita através da minimização da função perda esperada com respeito à variável aleatória Y , ou seja, basta obter o valor de b que minimiza

$$CQ = \min_{b \in \Re} [E \{\rho_\theta(Y - b)\}],$$

em que a função CQ é chamada de Critério Quantílico, em que $b = Q(\theta)$ – é o θ – quantil da v.a. Y .

Para o cálculo do quantil amostral, através da função de perda para o caso discreto, considere $\{Y_1, Y_2, \dots, Y_T\}$ uma amostra aleatória da v.a. Y e F_n a função de distribuição empírica:

$$F_n(y) = n^{-1} \sum_{i=1}^n I_{(Y_i \leq y)}, \quad (11)$$

em que $i \in \{1, \dots, T\}$ e $n \in \{1, \dots, T\}$, com $i \leq n$.

Escolhendo b que minimiza a perda esperada, considerando $F_n(\cdot)$ dada em (11), faz-se

$$CQ = \min_{b \in \Re} \left[n^{-1} \sum_{i=1}^n \rho_{\theta}(Y_i - b) \right], \quad (12)$$

e considerando um modelo linear

$$Y_i = x_i' \beta + u_i \Rightarrow u_i = Y_i - x_i' \beta,$$

em que $x_i' \beta$ são os regressores, de acordo com (12), tem-se que o vetor $\hat{\beta}_{\theta}$ é o vetor β , tal que

$$CQ = \min_{\beta} \left[n^{-1} \sum_{i=1}^n \rho_{\theta}(Y_i - x_i' \beta) \right].$$

Assim,

$$CQ = \min_{\beta} \left(\frac{1}{n} \left\{ \sum_{i: Y_i \geq x_i' \beta} \theta |Y_i - x_i' \beta| + \sum_{i: Y_i < x_i' \beta} (1 - \theta) |Y_i - x_i' \beta| \right\} \right) \quad (13)$$

resulta no cálculo da estimativa do vetor β_{θ} de parâmetros, ou seja, $Q(\theta)$ é obtido a partir do vetor $\hat{\beta}_{\theta}$ que minimiza a função perda esperada.

Para estender o conceito de Regressão Quantílica a uma série de retornos financeiros, existem algumas suposições básicas a serem feitas para que o cálculo do VaR, de acordo com as especificações citadas anteriormente seja possível.

A suposição chave do modelo linear é que

$$r_t = x_t' \beta_\theta + u_{t,\theta} \text{ e } Q_\theta = (u_{t,\theta} | x_t) = 0 \quad (14)$$

em que r_t são os retornos, $u_{t,\theta}$ é a parte aleatória, x_t o vetor de regressores, com ordem $(k \times 1)$ para um t fixo, β_θ o vetor de k parâmetros a serem estimados e $Q_\theta = (u_{t,\theta} | x_t)$ o θ – quantil de $u_{t,\theta}$ condicional em x_t .

De fato, tomando $Q_\theta = (u_{t,\theta} | x_t) = 0$, o modelo supõe que os erros tenham valores próximos de zero, independentemente do quantil adotado, assim possibilitando que os quantis condicionais dos retornos sejam relativamente iguais a $x_t' \beta_\theta$, ou seja,

$$Q_\theta = (r_t | x_t) = x_t' \beta_\theta. \quad (15)$$

Logo o θ – quantil de r_t dadas as informações até o instante t , é a parte determinística do modelo, ou seja, o conjunto de informações disponíveis até o período t .

Portanto, de (13)

$$CQ = \min_{\beta} \left(\frac{1}{T} \left\{ \sum_{t: r_t \geq VaR_t(\theta)} \theta |r_t - VaR_t(\theta)| + \sum_{t: r_t < VaR_t(\theta)} (1 - \theta) |r_t - VaR_t(\theta)| \right\} \right)$$

resulta na estimativa $\hat{\beta}_\theta$, do vetor β de parâmetros, necessária para o cálculo do $VaR_t(\theta)$, que pode assumir as especificações mostradas na seção 2.1.

Por exemplo, considerando a especificação **GARCH(1,1) Indireto**, dada em (3), calculando o vetor $\hat{\beta}_\theta$ pelo CQ tem-se,

$$CQ = \min_{\beta} \left[\frac{1}{T} \left\{ \sum_t \left| I_{r_t < \left(\beta_0 + \beta_1 VaR_{t-1}^2(\theta) + \beta_2 r_{t-1}^2 \right)^{\frac{1}{2}}} \right|^{-\theta} \right\} \right],$$

e $\hat{\beta}_\theta = \{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}^T$.

A função CQ é considerada como uma alternativa a mais para a seleção e adequação de modelos, pois para cada especificação, considerando os parâmetros estimados, deseja-se o que resulta no menor valor de CQ.

Considerados como um caso particular de estimadores de mínimos absolutos, os estimadores por regressão quantílica possuem características que os definem como sendo mais robustos que os de mínimos quadrados, por serem menos sensíveis ao comportamento das caudas das distribuições dos dados.

Segundo Engle e Manganelli (2004), os estimadores obtidos por regressão quantílica são consistentes e assintoticamente normais, ou seja, sob certas condições e utilizando como base o trabalho de Weiss (1991), demonstram que

$$\hat{\beta}_\theta \xrightarrow{p} \beta_{\theta, T \rightarrow \infty},$$

em que $\hat{\beta}_\theta$ é a solução para:

$$CQ = \min_{\beta} \left(\frac{1}{T} \left\{ \sum_{t: r_t \geq VaR_t(\theta)} \theta |r_t - VaR_t(\theta)| + \sum_{t: r_t < VaR_t(\theta)} (1 - \theta) |r_t - VaR_t(\theta)| \right\} \right)$$

Sob certas suposições, Engle e Manganelli (2004) provam que

$$\sqrt{T} A_T^{-1/2} D_T (\hat{\beta}_{\theta} - \beta_{\theta}) \xrightarrow{d} N(O, I),$$

em que

$$A_T \equiv E \left[T^{-1} \theta (1 - \theta) \sum_{t=1}^T \nabla' VaR_t(\theta) \nabla VaR_t(\theta) \right],$$

$$D_T \equiv E \left[T^{-1} \sum_{t=1}^T h_t(0|F_t) \nabla' VaR_t(\theta) \nabla VaR_t(\theta) \right],$$

em que $\nabla VaR_t(\theta)$ é gradiente de $VaR_t(\theta)$ e $h_t(0|F_t)$ a densidade condicional dos erros calculada em zero.

As matrizes A_T e D_T podem ser estimadas por

$$\begin{aligned} \hat{A}_T &= T^{-1} \theta (1 - \theta) \nabla' \widehat{VaR}_t(\theta) \nabla \widehat{VaR}_t(\theta) \\ \hat{D}_T &= (2T\hat{c}_T) \sum_{t=1}^T I(|r_t - \widehat{VaR}_t(\theta)| < \hat{c}_T) \nabla' \widehat{VaR}_t(\theta) \nabla \widehat{VaR}_t(\theta), \end{aligned}$$

em que

$$\hat{c}_T = \frac{1}{T} \sum_{t=1}^T \frac{\nabla \widehat{VaR}_t(\theta) \nabla' \widehat{VaR}_t(\theta)}{\widehat{VaR}_t(\theta)}.$$

2.3 Teste de Ajuste dos Modelos

Considerando uma série de retornos r_t , para uma posição comprada, para determinar se o modelo está corretamente especificado, tem-se o teste:

$$\begin{cases} H_0 : P\{r_t < VaR_t(\theta)\} = \theta, \forall t, \\ H_1 : P\{r_t < VaR_t(\theta)\} \neq \theta. \end{cases}$$

É possível obter, também, uma sequência de funções indicadoras

$$I(r_t < VaR_t(\theta)), t = 1, \dots, T$$

Assim, uma forma de testar se o modelo é válido, é verificar se a sequência de funções indicadoras é i.i.d.. Logo, uma propriedade que qualquer estimativa de VaR deve satisfazer é se comportar como um filtro, capaz de transformar uma série autocorrelacionada e heteroscedástica em uma série independente de funções indicadoras.

Esta é uma condição necessária, mas não suficiente para avaliar o comportamento e desempenho do modelo quantílico, uma vez que esta não considera e nem incorpora a influência de outras variáveis ao modelo.

Sendo assim, Engle e Manganelli (2004) propuseram um novo teste que pode ser estendido para incorporar essas variáveis ao modelo.

Seja

$$Hit_t(\beta_\theta) = I(r_t < VaR_t(\theta)) - \theta,$$

uma função que assume os valores $(1-\theta)$ quando r_t é menor que o θ -quantil e o valor $-\theta$ caso contrário.

Considerando $f_{t-1} \in F_{t-1}$,

$$E\{(f_{t-1}Hit_t)|f_{t-1}\} = f_{t-1}E\{Hit_t|f_{t-1}\},$$

o que mostra que Hit_t é não correlacionada com qualquer subconjunto pertencente ao conjunto de informações passadas F_{t-1} , ou seja, não há correlação entre Hit_t e qualquer Hit_{t-l} , para todo l .

Segundo os autores, uma maneira natural de fazer o teste é verificar se a estatística do teste

$$\frac{1}{T} X'(\hat{\beta})Hit(\hat{\beta})$$

é significativamente diferente de zero, em que $x_t(\hat{\beta}), t=1, \dots, T$ é um k -vetor, sendo uma linha de $X'(\hat{\beta})$, possivelmente dependendo de $\hat{\beta}$, mensurável em F_{t-1} , e

$$Hit(\hat{\beta}) \equiv [Hit_1(\hat{\beta}), Hit_2(\hat{\beta}), \dots, Hit_T(\hat{\beta})]'.$$

Visando testes para fins de seleção de modelos e que também possam avaliar quando uma estimativa de VaR , apresentada a uma instituição financeira, satisfaz alguns requisitos básicos, Engle e Manganelli (2004) desenvolveram os testes de *Quantis Dinâmicos (QD)*, que podem ser definidos como amostrais (QD_{am}) e pos-amostrais (QD_{pam}).

Inicialmente, detalha-se o teste QD_{am} , que verifica a qualidade do ajuste, permitindo a avaliação do modelo e possibilitando a escolha daquele mais adequado aos dados.

Sob certas suposições, assumindo algumas condições de regularidade,

$$QD_{am} \equiv \frac{Hit'(\hat{\beta})X(\hat{\beta})(\hat{M}_T\hat{M}_T')^{-1}X'(\hat{\beta})Hit'(\hat{\beta})}{\theta(1-\theta)} \stackrel{d}{\sim} \chi_{(k)}, \text{ quando } T \rightarrow \infty,$$

em que

$$\hat{M}_T \equiv X'(\hat{\beta}) - \left\{ (2T\hat{c}_T)^{-1} \sum_{t=1}^T I(|r_t - \widehat{VaR}_t(\theta)| < \hat{c}_T) \times x'_t(\hat{\beta}) \nabla \widehat{VaR}_t(\theta) \right\} \hat{D}_T^{-1} \nabla' \widehat{VaR}_t(\theta).$$

Já o teste de QD_{pam} serve como ferramenta para verificar se as estimativas dos parâmetros de interesse satisfazem alguns requisitos que toda boa estimativa quantílica deve satisfazer. Uma das grandes vantagens do teste é que pode ser considerado como simples, além do fato de não depender do procedimento de estimação. Para implementá-lo, é necessário a sequência de valores de $VaRs$ e dos correspondentes valores dos retornos.

Sendo assim, seja T_R o número de observações amostrais e N_R o número de observações pós-amostrais, ambos dependendo de T , em que $T \in \mathbb{Z}$.

Sob certas suposições,

$$QD_{pam} \equiv N_R^{-1} \text{Hit}'(\hat{\beta}_{T_R}) X(\hat{\beta}_{T_R}) [X'(\hat{\beta}_{T_R}) X(\hat{\beta}_{T_R})]^{-1} X \frac{X'(\hat{\beta}_{T_R}) \text{Hit}'(\hat{\beta}_{T_R})^d}{\theta(1-\theta)} \sim X_{(k)},$$

quando $R \rightarrow \infty$.

Todos os detalhes podem ser vistos em Engle e Manganelli (2004).

3. CARE – Expectil Autorregressivo Condicional

Uma das fragilidades dos modelos CAViaR com parâmetros estimados por Regressão Quantílica é que a magnitude dos dados não tem influência no processo de estimação, pois o que se considera de fato são os quantis dos dados.

Assim, ao considerar por exemplo, dois conjuntos de dados com caudas se comportando de formas diferentes, eles podem apresentar o mesmo valor para um θ -quantil.

Com o objetivo de obter medidas de risco calculadas por um método mais adequado, Taylor (2008) introduz os modelos Expectil Autorregressivo Condicional (CARE) com as mesmas estruturas dos modelos CAViaR, porém com parâmetros estimados por Mínimos

Quadrados Assimétricos. Esses modelos receberam algumas críticas por apresentarem as mesmas estruturas dos modelos CAViaR, sem considerar as propriedades assintóticas do estimador de Mínimos Quadrados Assimétricos, além de não definir um teste no contexto dinâmico.

A seguir, Kuan, Yeh e Hsu (2009) propõem novas especificações, com propriedades assintóticas e testes.

Antes de detalhar as especificações do modelo, é necessário que se defina expectil.

3.1 Expectil

Considere a função de perda definida anteriormente $\rho_\tau(u)$, com $\tau \in (0,1)$ fixo. O τ – expectil, denotado por $\mu(\tau)$ é o parâmetro m que minimiza a função de perda calculada em $(Y - m)^2$, com relação à v.a. Y , ou seja, é um mínimo quadrado assimétrico (MQA), em que

$$MQA = \min_{m \in \mathfrak{R}} [E\{\rho_\tau(Y - m)^2\}]. \quad (16)$$

De fato, para o cálculo de $\mu(\tau)$ de acordo com (16), tem-se:

$$E\{\rho_\tau(Y - m)^2\} = |\tau - 1| \int_{-\infty}^m (Y - m)^2 dF_Y(y) + |\tau| \int_m^{+\infty} (Y - m)^2 dF_Y(y).$$

Minimizando em relação a m tem-se:

$$\begin{aligned} (1 - \tau) \int_{-\infty}^{\mu(\tau)} (Y - \mu(\tau)) dF_Y(y) + \tau \int_{\mu(\tau)}^{+\infty} (Y - \mu(\tau)) dF_Y(y) &= 0 \Rightarrow \\ \Rightarrow \mu(\tau) - E(Y) &= \left(\frac{1 - 2\tau}{\tau} \right) \int_{-\infty}^{\mu(\tau)} (Y - \mu(\tau)) dF_Y(y). \end{aligned} \quad (17)$$

Logo, pode-se concluir que o resultado dessa minimização é o valor $\mu(\tau)$ que satisfaz a equação (17).

Ao comparar o θ – quantil e o τ – expectil, Koenker (2005) observa que o expectil tem uma maior dependência global da forma da distribuição. Por exemplo, alterando a forma da distribuição nas caudas, o quantil não é alterado, mas há um impacto no expectil.

De acordo com a definição de expectil, inicialmente é necessário determinar o valor mais adequado para τ , dado o valor de θ .

Seja $\mu(\tau)$ tal que de acordo com a equação (17):

$$\frac{1}{\theta}(\mu(\tau) - E(Y)) = \left(\frac{1-2\tau}{\tau} \right) \frac{1}{\theta} \left\{ E(YI_{Y < \mu(\tau)}) - \mu(\tau) E(I_{Y < \mu(\tau)}) \right\}.$$

Fazendo $E(I_{Y < \mu(\tau)}) = F(\mu(\tau)) = \theta$, em que F é a função de distribuição acumulada da v.a. Y , tem-se:

$$\frac{1-2\tau}{\tau} = \frac{\mu(\tau) - E(Y)}{\theta \left\{ \frac{E(YI_{Y < \mu(\tau)})}{\theta} - \mu(\tau) \right\}}. \quad (18)$$

Estendendo o conceito para uma série de retornos, inicialmente, fixa-se o valor de θ , logo $\hat{\mu}(\tau) = q_\theta(r_t | F_{t-1})$, em que $q_\theta(r_t | F_{t-1})$ é o θ – quantil empírico da série de retornos dadas as informações passadas, $E(Y) = E(r_t | F_{t-1})$, e

$$\frac{E(YI_{Y < \mu(\tau)})}{\theta} = E\left((r_t | F_{t-1}) | (r_t | F_{t-1}) < q_\theta(r_t | F_{t-1})\right),$$

ou seja, $\frac{E(YI_{Y < \hat{\mu}(\tau)})}{\theta}$ é obtido pela média dos retornos abaixo de θ – quantil estimado.

Logo, de acordo com (18), tem-se:

$$\frac{1-2\tau}{\tau} = \frac{\hat{\mu}(\tau) - E(r_t | F_{t-1})}{\theta \left\{ E\left((r_t | F_{t-1}) | (r_t | F_{t-1}) < q_\theta(r_t | F_{t-1})\right) - \hat{\mu}(\tau) \right\}}. \quad (19)$$

Fazendo o segundo lado da igualdade igual a K, tem-se:

$$\hat{\tau} = \frac{\hat{K}}{1 + 2\hat{K}}. \quad (20)$$

3.2 Especificações dos modelos CARE

Seguindo as mesmas estruturas dos modelos CAViaR, e utilizando o τ – expectil como um estimador o θ – quantil, em que não necessariamente lado τ seja igual a θ , Taylor (2008) define as mesmas especificações apresentadas nas expressões (1)–(7).

Para a estimação dos modelos CARE, utiliza-se o mesmo procedimento que o feito para a estimação dos modelos CAViaR, porém ao invés de calcular a função de CQ é calculada a função e MQA.

Com o objetivo de selecionar e verificar a adequação do modelo, como feito para os modelos CAViaR, Taylor (2008) propõe os teste de Hit_{am} , Hit_{pam} e QD_{am} e QD_{pam} .

3.3 Novas especificações propostas para modelos CARE

Kuan, Yeh e Hsu (2009) propõem novas especificações, mencionadas a seguir:

Considerando:

- i) Uma coleção de k variáveis, X o conjunto de informações, tais que $X \in F_t$ e β_τ o vetor $(k \times 1)$ de parâmetros dependendo de τ ;
- ii) (r_t, x'_t) o conjunto de dados disponível;
- iii) o modelo:

$$r_t = x'_t \beta_\tau + e_{t,\tau}, \quad (21)$$

em que $x'_t \in X$, $t = \{1, \dots, T\}$ e $e_{t,\tau}$ é o erro;

v) $\mu(\tau, X)$ o τ -expectil do conjunto de informações condicional em F ;

vi) que o modelo está corretamente especificado para $\mu(\tau, X)$ se existe β_τ^0 tal que $X'\beta_\tau^0 = \mu(\tau, X)$ com probabilidade um;

Sob a suposição de que o modelo é corretamente especificado, tem-se:

$$r_t = x_t' \beta_\tau^0 + \varepsilon_{t,\tau}. \quad (22)$$

No contexto dinâmico, para modelar o τ -expectil de r_t , são consideradas informações até o instante $(t-1)$, ou seja $x_t \in F_{t-1}$.

Sabendo que o expectil é também um quantil, se uma série é autocorrelacionada, então $\mu(\tau)$, para os valores fixos τ , também será autocorrelacionada. Visando ter uma medida de risco que não apresente este comportamento, retornos defasados no tempo são incluídos nos modelos.

De acordo com a definição de expectil, visando incorporar o impacto que a magnitude dos dados exercem no modelo, o quadrado ou o valor absoluto dos retornos também podem ser incluídos nos modelos.

Sejam

$$\begin{aligned} r_{t-1} &= r_{t-1}^+ - r_{t-1}^- \\ |r_{t-1}| &= r_{t-1}^+ + r_{t-1}^- \\ r_{t-1}^2 &= (r_{t-1}^+)^2 + (r_{t-1}^-)^2, \end{aligned}$$

em que $r_{t-1}^+ = \max(r_{t-1}, 0)$ e $r_{t-1}^- = \max(-r_{t-1}, 0)$.

Logo, de acordo com (21) e (22), as especificações CARE são mostradas a seguir:

- Especificação [1]

Para $x_t = (1, r_{t-1}, (r_{t-1}^+)^2, (r_{t-1}^-)^2)'$,

$$\begin{aligned} r_t &= a_0(\tau) + a_1(\tau)r_{t-1} + b_1(\tau)r_{t-1}^2 + c_1(\tau)(r_{t-1}^-)^2 + e_{t,\tau} \Rightarrow \\ r_t &= a_0(\tau) + a_1(\tau)r_{t-1} + b_1(\tau)[(r_{t-1}^+)^2 + (r_{t-1}^-)^2] + c_1(\tau)(r_{t-1}^-)^2 + e_{t,\tau} \Rightarrow \\ r_t &= a_0(\tau) + a_1(\tau)r_{t-1} + b_1(\tau)(r_{t-1}^+)^2 + \gamma_1(\tau)(r_{t-1}^-)^2 + e_{t,\tau} \end{aligned} \quad (23)$$

em que $\gamma_1(\tau) = b_1(\tau) + c_1(\tau)$.

Neste caso,

$$\begin{aligned} \mu_t(\tau) &= a_0(\tau) + a_1(\tau)[\mu_{t-1}(\tau) + e_{t-1,\tau}] + b_1(\tau)(r_{t-1}^+)^2 + \gamma_1(\tau)(r_{t-1}^-)^2 \Rightarrow \\ \mu_t(\tau) &= a_0(\tau) + a_1(\tau)\mu_{t-1}(\tau) + b_1(\tau)(r_{t-1}^+)^2 + \gamma_1(\tau)(r_{t-1}^-)^2 + a_1(\tau)e_{t-1,\tau}. \end{aligned}$$

Neste modelo, a magnitude e o erro defasado no tempo são responsáveis pelo comportamento do expectil condicional do modelo.

Note que, $\beta_\tau = (a_0(\tau), a_1(\tau), b_1(\tau), \gamma_1(\tau))'$.

- Especificação [2]

Uma outra alternativa de especificação, é o uso do $|r_{t-1}|$ para representar o efeito dos dados no modelo. Para este caso, $x_t = (1, r_{t-1}^+, r_{t-1}^-)'$.

Assim,

$$\begin{aligned} r_t &= a_0(\tau) + a_1(\tau)r_{t-1} + b_1(\tau)|r_{t-1}| + e_{t,\tau} \Rightarrow \\ r_t &= a_0(\tau) + \gamma_1(\tau)r_{t-1}^+ + \delta_1(\tau)r_{t-1}^- + e_{t,\tau}, \end{aligned} \quad (24)$$

em que $\gamma_1(\tau) = a_1(\tau) = b_1(\tau)$ e $\delta_1(\tau) = b_1(\tau) - a_1(\tau)$.

Neste caso, $\beta_\tau = (a_0(\tau), \gamma_1(\tau), \delta_1(\tau))$.

Esta especificação é similar a especificação do modelo CAViaR inclinação assimétrica, porém no modelo CARE o τ -expectil condicional defasado no tempo não está inserido no modelo.

- Especificação [3]

Esta especificação, é na realidade uma extensão natural do modelo definido em (23).

Seja

$$r_t = a_0(\tau) + \sum_{i=1}^p a_i(\tau) r_{t-i} + \sum_{j=1}^q \left[b_j(\tau) (r_{t-j}^+)^2 + \gamma_j(\tau) (r_{t-j}^-)^2 \right] + e_{t,\tau}. \quad (25)$$

Neste caso,

$$\mu_t(\tau) = a_0(\tau) + \sum_{i=1}^p a_i(\tau) \mu_{t-i}(\tau) + \sum_{j=1}^q \left[b_j(\tau) (\mu_{t-j}^+)^2 + \gamma_j(\tau) (\mu_{t-j}^-)^2 \right] + \sum_{i=1}^p e_{t-i,\tau},$$

em que de forma análoga ao modelo (23), a magnitude dos dados e os erros defasados no tempo eu determinam o comportamento do τ -expectil, em que:

$$\beta_\tau = (a_0(\tau), a_1(\tau), \dots, a_p(\tau), b_1(\tau), \dots, b_q(\tau), \gamma_1(\tau), \dots, \gamma_q(\tau)).$$

- Especificação [4]

Como no caso anterior, esta é uma extensão do modelo (24).

Assim:

$$r_t = a_0(\tau) + \sum_{i=1}^p \left[b_i(\tau) r_{t-i}^+ + \gamma_i(\tau) r_{t-i}^- \right] + e_{t,\tau}, \quad (26)$$

com $\beta_\tau = (a_0(\tau), b_1(\tau), \dots, b_p(\tau), \gamma_1(\tau), \dots, \gamma_p(\tau))$

3.4 Estimação do Modelo

As quatro especificações detalhadas anteriormente podem ser estimadas utilizando o método de Mínimos Quadrados Assimétricos (MQA), descrito e proposto por Newey e Powell (1987).

Neste caso a média da função de perda ρ_τ é calculada utilizando uma diferença quadrática $(y - m)^2$, atribuindo pesos $(1 - \tau)$ e τ de acordo com os sinais das diferenças, calculando o seu mínimo, como visto em (16). Devido a tais pesos o método é chamado de assimétrico.

Assim, o estimador $\hat{\beta}_\tau$ é o vetor β que satisfaz.

$$MQA = \min_{\beta} \left\{ T^{-1} \sum_{t=1}^T \rho_\tau(r_t - x_t' \beta)^2 \right\}. \quad (27)$$

Este estimador satisfaz

$$\hat{\beta}_\tau = \left(\sum_{t=1}^T \left[\tau - I(r_t \leq x_t' \hat{\beta}_\tau) \right] (x_t r_t) \right) \left(\sum_{t=1}^T \left[\tau - I(r_t \leq x_t' \hat{\beta}_\tau) \right] (x_t x_t') \right)^{-1} \quad (28)$$

Embora (28) não seja uma solução fechada, pode ser calculada como um estimador de mínimos quadrados ponderados iterado.

Newey e Powell (1987) estabelecem a consistência e normalidade assintótica do estimador dado em (27), sob as condições de que os dados são i.i.d. De fato, mostram que

$$\hat{\beta}_\tau \xrightarrow{p} \beta_\tau, T \rightarrow \infty,$$

e também,

$$\sqrt{T}(\hat{\beta}_\tau - \beta_\tau) \xrightarrow{d} N(0, \Sigma(\tau)),$$

em que

$$\Sigma(\tau) = A(\tau)^{-1}V(\tau)A(\tau)^{-1},$$

com

$$A(\tau) = E\left[\left|\tau - I(r_t \leq x'_t \beta_\tau)\right| x'_t x'_t\right],$$

$$V(\tau) = \lim_{T \rightarrow \infty} V_T(\tau) := \lim_{T \rightarrow \infty} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T \left| \tau - I(r_t \leq x'_t \beta_\tau) \right| x'_t (r_t - x'_t \beta_\tau) \right).$$

De acordo com (21) e (22), se o modelo estiver corretamente especificado, $\beta_\tau = \beta_\tau^0$, em que

$$\beta_\tau^0 = E\left[\rho_\tau(r_t - x'_t \beta_\tau)^2 \middle| F_{t-1}\right].$$

Considerando todas as condições de regularidade e as propriedades de diferenças martingais, Newey e Powell (1987) mostram que

$$\sqrt{T}(\hat{\beta}_\tau - \beta_\tau^0) \xrightarrow{d} N(0, \Sigma(\tau)),$$

em que

$$\Sigma(\tau) = A(\tau)^{-1}V(\tau)A(\tau)^{-1},$$

com

$$V(\tau) = \text{Var}\left(\left|\tau - I(r_t \leq x'_t \beta_\tau^0)\right| x'_t (r_t - x'_t \beta_\tau^0)\right).$$

A matriz de covariâncias assintótica $\Sigma(\tau)$ pode ser estimada por

$$\hat{\Sigma}(\tau) = \hat{A}(\tau)^{-1} \hat{V}(\tau) \hat{A}(\tau)^{-1},$$

em que

$$\hat{A}(\tau) = \frac{1}{T} \sum_{t=1}^T \left(\left| \tau - I(r_t \leq x_t' \hat{\beta}_\tau) \right| x_t x_t' \right) \xrightarrow{P} A(\tau),$$

$$\hat{V}(\tau) = \frac{1}{T} \sum_{t=1}^T \left(\left| \tau - I(r_t \leq x_t' \hat{\beta}_\tau) \right|^2 (r_t - x_t' \hat{\beta}_\tau)^2 x_t x_t' \right) \xrightarrow{P} V(\tau),$$

com

$$V(\tau) = \text{Var} \left\{ \left| \tau - I(r_t \leq x_t' \beta_\tau^0) \right| x_t (r_t - x_t' \beta_\tau^0) \right\}$$

3.5 Teste de Ajuste dos Modelos

Pelas especificações escolhidas de acordo com os autores, para determinar qual o modelo mais apropriado, foi desenvolvido um teste com as seguintes hipóteses:

$$\begin{cases} H_0 : x_t' \beta_\tau^0 = \mu(\tau, F_{t-1}), \text{ com probabilidade } 1, \\ H_1 : \zeta_t' \beta_\tau^0 = \mu(\tau, F_{t-1}), \text{ com probabilidade } 1, \end{cases}$$

em que x_t tem ordem $(k \times T)$ e ζ_t' com ordem $(m \times T)$, ambos contidos em F_{t-1} , em que $k=3q+1$ e $m=2q+1$.

Note que

$$x_t' = (1, r_{t-i}, (r_{t-i}^+)^2, (r_{t-i}^-)^2), i=1, \dots, q \text{ e}$$

$$\zeta_t' = (1, r_{t-i}^+, r_{t-i}^-), i=1, \dots, q.$$

De acordo com os autores, uma forma de testar o modelo sob a hipótese nula, é checar se os erros ponderados são não correlacionados com as variáveis do modelo sob a hipótese alternativa:

Então, sob certas suposições,

$$\frac{1}{T} \left(\sum_{t=1}^T \left| \tau - I(r_t \leq x'_t \hat{\beta}_\tau) \right| \tilde{\zeta}_t \hat{e}_{t,\tau} \right)' (\tilde{\Omega}(\tau))^{-1} \left(\sum_{t=1}^T \left| \tau - I(r_t \leq x'_t \hat{\beta}_\tau) \right| \tilde{\zeta}_t \hat{e}_{t,\tau} \right) \xrightarrow{D} \chi^2_{(q)},$$

em que $\tilde{\Omega}(\tau)$ tem ordem q , e é estimado, substituindo ζ_t por $\tilde{\zeta}_t$.

Mais detalhes sobre a estatística do teste, assim como a distribuição assintótica podem ser visto em Kuan, Yeh e Hsu (2009).

4. ES – Expected Shortfall

O VaR, como comentado e definido anteriormente é uma medida de risco muito utilizada, mas que apresenta algumas fragilidades, como por exemplo não ser uma medida de risco coerente e não apresentar informações sobre os retornos que estão acima ou abaixo do Valor em Risco, dependendo da posição.

Assim, considerando uma posição vendida, de acordo com Artzner, Delbaen, Eber e Heath (1999), e McNeil, Frey e Embrechts (2005), Expected Shortfall (ES) é uma medida de risco definida como a esperança condicional dos retornos, dado que excedem o valor do VaR. Já para posições compradas, ES é a esperança condicional dos retornos que estão abaixo do valor do VaR estimado. Dessa forma, considerando L a v.a. que representa a variação percentual de preço e fixando θ para uma posição vendida, estima-se $\text{VaR}(\theta)$ definido anteriormente e calcula-se.

$$\widehat{ES}_L(\theta) = E\{L | L \geq \widehat{\text{VaR}}_L(\theta)\},$$

com

$$ES_L(\theta) = \frac{1}{1-\theta} \int_{\theta}^1 VaR_L(\theta) d\theta. \quad (29)$$

Para posição comprada,

$$ES_L(\theta) = \frac{1}{\theta} \int_0^{\theta} VaR_L(\theta) d\theta. \quad (30)$$

4.1 Estimação do ES utilizando Expectil

Um outro método, proposto por Taylor (2008), é o uso do τ -expectil como um estimador para o θ -quantil, possibilitando uma nova alternativa para a estimação de ES.

Considere MQA dado por

$$MQA = \min_{m \in \mathcal{R}} [E\{\rho_{\tau}(Y - m)^2\}] ,$$

para um τ fixo, como visto anteriormente em (17)

$$\begin{aligned} \frac{1-2\tau}{\tau} \int_{-\infty}^{\mu(\tau)} (Y - \mu(\tau)) dF_Y(y) &= \mu(\tau) - E(Y) \\ \Rightarrow E\{Y I_{(Y \leq \mu(\tau))}\} - \mu(\tau) E\{I_{(Y \leq \mu(\tau))}\} &= \left(\frac{\tau}{1-2\tau} \right) [\mu(\tau) - E(Y)]. \end{aligned} \quad (31)$$

Utilizando um $\mu(\tau)$ como estimador para o θ -quantil, tal que $F(\mu(\tau)) = \theta$, em que $F(\cdot)$ é a f.d. de Y , tem-se:

$$E\{I_{(Y \leq \mu(\tau))}\} = \theta \quad (32)$$

Substituindo (32) em (31), tem-se:

$$E\{Y I_{(Y \leq \mu(\tau))}\} - \mu(\tau)(\theta) = \left(\frac{\tau}{1-2\tau}\right) [\mu(\tau) - E(Y)].$$

Dividindo ambos os lados da igualdade por θ , tem-se:

$$ES(\theta) = \left(1 + \frac{\tau}{(1-2\tau)\theta}\right) \mu(\tau) - \frac{\tau}{(1-2\tau)\theta} E(Y).$$

Estendendo a estimação de ES para uma sequência r_t , com $t = (1, \dots, T)$, tem-se:

$$\widehat{ES}_t(\theta) = \left(1 + \frac{\tau}{(1-2\tau)\theta}\right) \hat{\mu}_t(\tau) - \frac{\tau}{(1-2\tau)\theta} E(r_t | \mathcal{F}_{t-1}). \quad (33)$$

Note que a estimativa do ES pelo expectil, é resultado da ligação entre a minimização da função de perda e a definição de ES, utilizando o τ -expectil como estimativa para o θ -quantil.

Sendo assim, de acordo com (33), pode-se ver que para o modelo condicional, o valor do τ -expectil é proporcional ao valor do ES.

Utilizando as estruturas do modelo CARE, pode-se estimar vários modelos autorregressivos para o ES, condicional a informações passadas. Por exemplo o modelo **CARE-Absoluto Simétrico**, tal que

$$\mu_t(\tau) = \beta_0 + \beta_1 \mu_{t-1}(\tau) + \beta_2 |r_{t-1}|, \quad (34)$$

em um modelo em que $E(r_t | F_{t-1}) = 0$, substituindo (34) em (33),

$$ES_t(\theta) = \gamma_0 + \gamma_1 ES_{t-1}(\theta) + \gamma_2 |r_{t-1}|,$$

em que $\gamma_1 = \beta_1$, e para $j = 0$ e 2 , $\gamma_j = \left(1 + \frac{\tau}{(1-2\tau)\theta}\right) \beta_j$.

Utilizando as outras especificações, utilizando a mesma ideia é possível estimar o ES.

4.2 Verificação dos Métodos e Modelos para o ES

A idéia principal, de acordo com McNeil e Frey (2000), é obter um teste que não faça suposições sobre a distribuição dos excessos, e nem sobre os resíduos dos modelos ajustados.

Inicialmente, o teste considera somente os retornos que excedem o valor do quantil condicional ajustado, o VaR.

Considere todos os retornos r_t tais que

$$r_t > VaR_t(\theta),$$

para posição vendida e

$$r_t < VaR_t(\theta),$$

para a posição comprada.

Após o ajuste do ES para cada instante de tempo, temos os excessos com as seguintes estruturas

$$Z_t = \frac{r_t - ES_t(\theta)}{\sigma_t},$$

em que Z_t são variáveis aleatórias i.i.d..

Note que para os modelos CARE propostos por Taylor (2008) e para os métodos utilizando a Teoria de Valores Extremos (TVE), em que a volatilidade não é estimada para o cálculo do ES, o teste bootstrap é adaptado. Ao invés de padronizar os excessos pela volatilidade condicional estimada, os excessos são padronizados pelo quantil condicional estimado, ou seja:

$$Z_t = \frac{r_t - ES_t(\theta)}{VaR_t(\theta)},$$

Para esses excessos, sob o a hipótese nula de que o modelo ou método estima corretamente o ES, espera-se que tenham média zero e para isso, constrói-se o teste bootstrap para testar, no caso de posição comprada,

$$\begin{cases} H_0 : \mu_z = 0 \\ H_1 : \mu_z < 0; \end{cases}$$

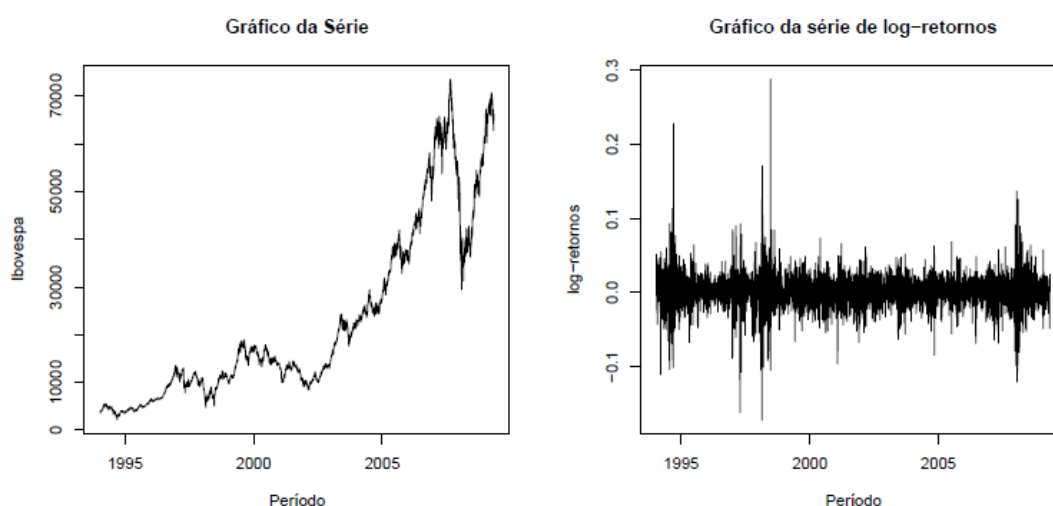
que é equivalente a testar se o modelo sistematicamente, superestima ou não o ES.

5. Aplicação

Para aplicação, considerou-se o Índice da Bolsa de Valores de São Paulo (IBOVESPA), série constituída de valores no período de 04/07/1994 a 11/02/2010, totalizando 3.862 observações. A Figura 1 apresenta o gráfico da série e dos log-retornos, respectivamente.

A Análise da Figura 1 revela que o IBOVESPA, assim como outros indicadores, sofreu impacto da crise econômica de 2008 a 2009, causada pela quebra de uma dos maiores bancos de investimentos do mundo, o Lehman Brothers e, apesar de órgãos do governo brasileiro afirmarem que a crise não afetaria o país, o índice da Bolsa de São Paulo mostrou como o período foi de grande instabilidade.

Figura 1: Gráficos da série de IBOVESPA e da série de log-retornos



5.1 Modelos CAViaR

Para o ajuste dos modelos CAViaR, de acordo com Engle e Manganelli (2004), inicialmente gera-se n vetores β entre 0 e 1, em que a ordem de cada vetor depende da especificação adotada.

De acordo com a posição desejada, fixa-se o valor de θ . Visando o cálculo das medidas de risco para posições compradas, adotou-se $\theta = 0,05$.

Para cada vetor β , aplica-se a função e Critério Quantílico (CQ). Obtém-se m vetores que produziram os menores valores para o CQ.

No trabalho, foram fixos os valores $n = [10^4, 10^5, 10^4, 10^4, 10^4, 10^4, 10^4]$ e $m = [10, 15, 10, 5, 10, 10, 10]$ para as 7 especificações.

Esses m vetores são utilizados como valores iniciais para o algoritmo Quasi-Newton. Este procedimento é realizado até a convergência do algoritmo. O nível de tolerância para a função e os valores dos parâmetros é 10^{-10} .

Para a estimação do modelo foi utilizado o software *Matlab*, com as funções de otimização *fminunc* e *fminsearch*.

Como estimativa do $VaR_1(\theta)$ toma-se θ – quantil das observações amostrais.

A figura 2 apresenta os gráficos dos valores estimados para o VaR de acordo com as especificações do modelo.

As Tabelas 1 e 2 apresentam o valor estimado, *P-valor* e *Erro Padrão* pra cada um dos parâmetros, o valor dos Hit_{am} , Hit_{pam} , dos teste de QD_{am} e QD_{pam} , além do valor do CQ para cada um dos modelos adotados.

As análises de todos os modelos foram feitas a um nível de significância de 5% para todos os testes.

De modo geral, ao se falar dos modelos CAViaR, definir qual é o melhor modelo depende de a qual cenário se está referindo.

No caso da série do IBOVESPA, de acordo com os modelos ajustados, o mais indicados para estimar o valor do VaR seria o modelo Inclinação Assimétrica, em que:

$$\widehat{VaR}_t(0,05) = 0,2096 + 0,8749\widehat{VaR}_{t-1}(0,05) - 0,0349r_{t-1}I(r_{t-1} > 0) + 0,3268r_{t-1}I(r_{t-1} < 0),$$

pois apresenta o menor valor para o CQ e o melhor valor para o teste de QD_{am} . As proporções de Hits amostrais e pós-amostrais também indicam que o modelo é adequado.

Figura 2: Gráficos de VaR estimados através dos modelos CAViaR

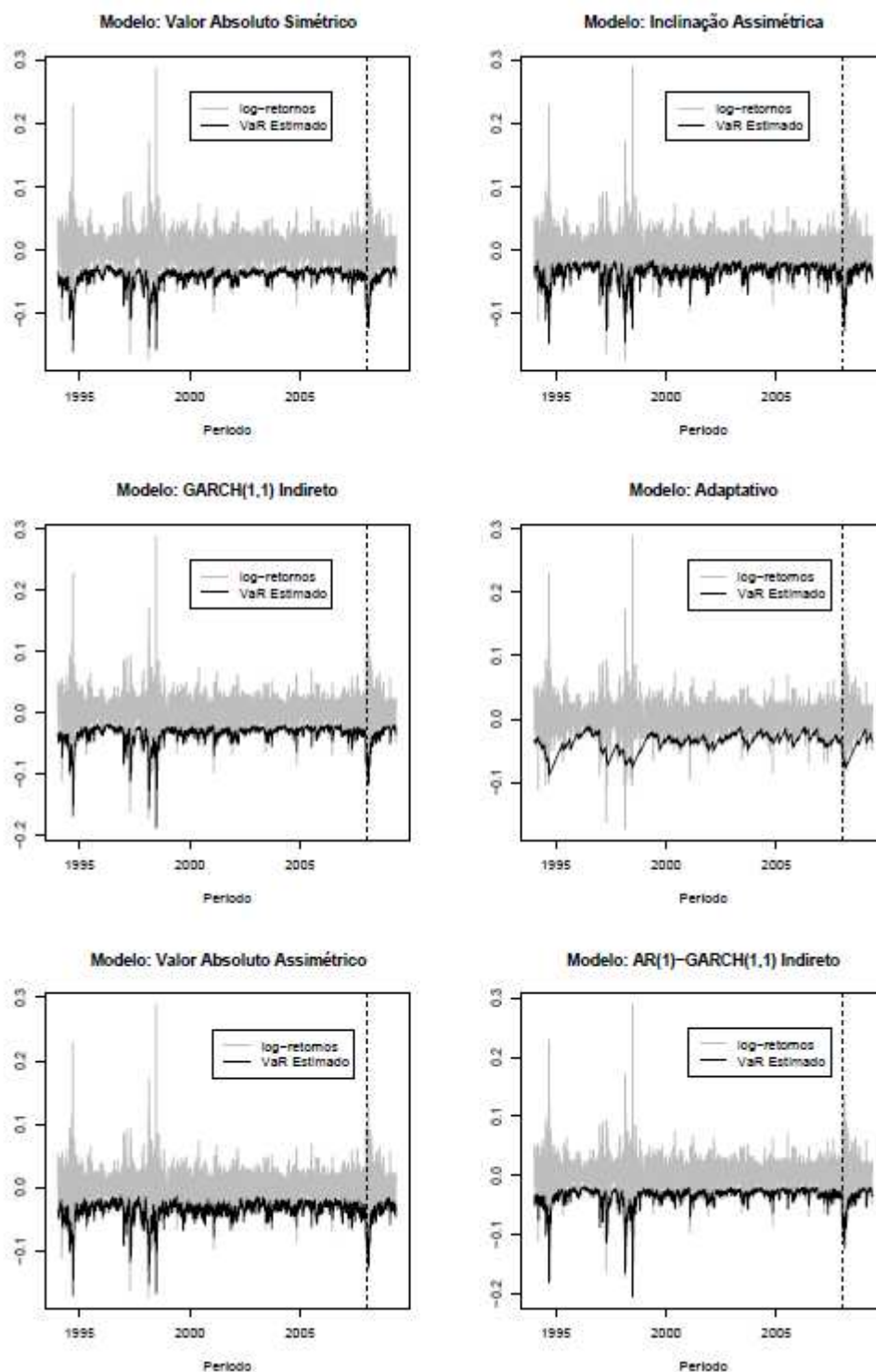


Tabela 1: Estimativas e Estatísticas dos Parâmetros Estimados utilizando as Especificações do Modelo CAViaR

Estimativas e Testes				
Modelo	Valor Absoluto Simétrico	Inclinação Assimétrica	GARCH(1,1) Indireto	Adaptativo
β_1	0,2485	0,2096	0,4596	0,6494
Erro Padrão	0,0875	0,0582	0,2522	0,0868
<i>P-valor</i>	0,0022	0,0002	0,0342	$3,6871.10^{-14}$
β_2	0,8433	0,8749	0,8359	-
Erro Padrão	0,0532	0,0175	0,0262	-
<i>P-valor</i>	0	0	0	-
β_3	0,2475	-0,0349	0,3454	-
Erro Padrão	0,0878	0,0310	0,1311	-
<i>P-valor</i>	0,0024	0,13	0,0042	-
β_4	-	0,3268	-	-
Erro Padrão	-	0,0267	-	-
<i>P-valor</i>	-	0	-	-
Hit_{am}	4,9688%	4,99%	5,0580%	4,8795%
Hit_{pam}	6,8%	6,4%	6,4%	4,8%
QD_{am}	0,0782	0,9327	0,0078	0,0184
QD_{pam}	0,5413	0,3419	0,7022	0,4344
<i>CQ</i>	835,4547	801,2110	836,2239	871,3505

Tabela 2: Estimativas e Estatísticas dos Parâmetros Estimados utilizando as Especificações do Modelo CAViaR

Estimativas e Testes			
Modelo	Adaptativo Sim.Prop.	Valor Absoluto Assimétrico	AR(1)-GARCH(1,1) Indireto
β_1	-	-0,0361	0,5175
Erro Padrão	-	0,1437	0,2428
<i>P-valor</i>	-	0,4007	0,0165
β_2	-	0,8314	0,8259
Erro Padrão	-	0,0393	0,0269
<i>P-valor</i>	-	0	0
β_3	-	0,2874	0,3697
Erro Padrão	-	0,0387	0,1141
<i>P-valor</i>	-	0	0,0006
β_4	-	0,0177	-
Erro Padrão	-	0,4154	-
<i>P-valor</i>	-	0	-
Hit_{am}	-	4,9985%	4,9985%
Hit_{pam}	-	5%	6,4%
QD_{am}	-	0,4020	0,0022
QD_{pam}	-	0,7906	0,5863
<i>CQ</i>	-	813,3368	842,6946

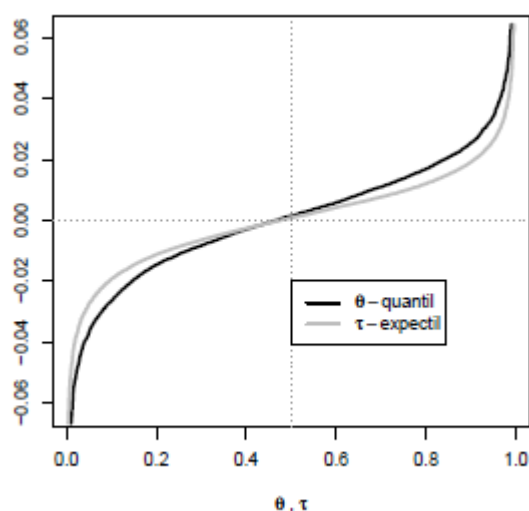
5.2 Modelos CARE

Considerando as especificações definidas por Taylor (2008), primeiramente é necessário que se estime o valor do θ – quantil pelo τ – expectil de acordo com as idéias de Efron (1991).

Este processo de estimação só é possível devido ao fato de que para cada θ – expectil existe um correspondente θ – quantil na amostra, em que não necessariamente τ seja igual a θ . Todo o suporte teórico é feito e demonstrado por Jones (1994).

A figura 3 apresenta o gráfico do θ – quantil e τ – expectil de acordo com os valores de θ e τ para a série de log-retornos.

Figura 3: Gráfico do θ – quantil e τ – expectil obtidos para a série de log-retornos.



Pelo gráfico é possível notar que τ apresenta valores mais extremos que os correspondentes valores de θ .

Considerando a série de retornos do IBOVESPA, assim como nos modelos CAViaR, foram utilizadas para a estimação dos modelos os primeiros 3361 log-retornos, e de acordo com o procedimento, utilizando a expressão (20), para $\theta = 0,05$, tem-se

$$\hat{\tau} = 0,02436630.$$

Logo, seguindo o mesmo procedimento para a estimação dos modelos CAViaR, estima-se os parâmetros dos modelos, utilizando Mínimos Quadrados Assimétricos (MQA), ao invés do Critério Quantílico.

A Figura 4 apresenta os gráficos dos valores estimados para o VaR de acordo com as especificações do modelo CARE.

Os gráficos indicam que não há muita diferença entre as estimativas utilizando modelos CARE e CAViaR, a não ser para o modelo Adaptativo.

As tabelas 3 e 4 apresentam o valor estimado, Erro Padrão, P-valor para cada um dos parâmetros, a proporção de Hit_{am} , Hit_{pam} , QD_{am} e QD_{pam} , e também o valor do MQA e CQ para cada modelo.

Assim como nos modelos CAViaR, todas as análises foram feitas considerando um nível de significância de 5%.

Tabela 3: Estimativas e Estatísticas dos Parâmetros Estimados utilizando as especificações do Modelo CARE

Estimativas e Testes				
Modelo	Valor Absoluto Simétrico	Inclinação Assimétrica	GARCH(1,1) Indireto	Adaptativo
β_1	0,3352	0,2612	0,7336	-0,0049
Erro Padrão	0,0719	0,0634	0,1971	0,0025
P-valor	0,1560	0	0,0001	0,0244
β_2	0,8040	0,8484	0,7973	-
Erro Padrão	0,0432	0,0287	0,0192	-
P-valor	0	0	0	-
β_3	0,2441	-0,0483	0,2984	-
Erro Padrão	0,0537	0,0290	0,1497	-
P-valor	0,2698	0,0481	0,0231	-
β_4	-	0,3197	-	-
Erro Padrão	-	0,0637	-	-
P-valor	-	0	-	-
Hit_{am}	6,3374%	6,5457%	6,57%	5,3853%
Hit_{pam}	8,4%	8,2%	8%	8%
QD_{am}	$3,18.10^{-8}$	0,6733	$8,9.10^{-6}$	0
QD_{pam}	0,0151	0,0267	0,07	$1,11.10^{-6}$
MQA	2.439,8	2.252,7	2.456,6	2.850,7
CQ	979,6795	954,7573	981,1122	1.092,6

Figura 4: Gráficos de VaR estimados através dos modelos CARE.

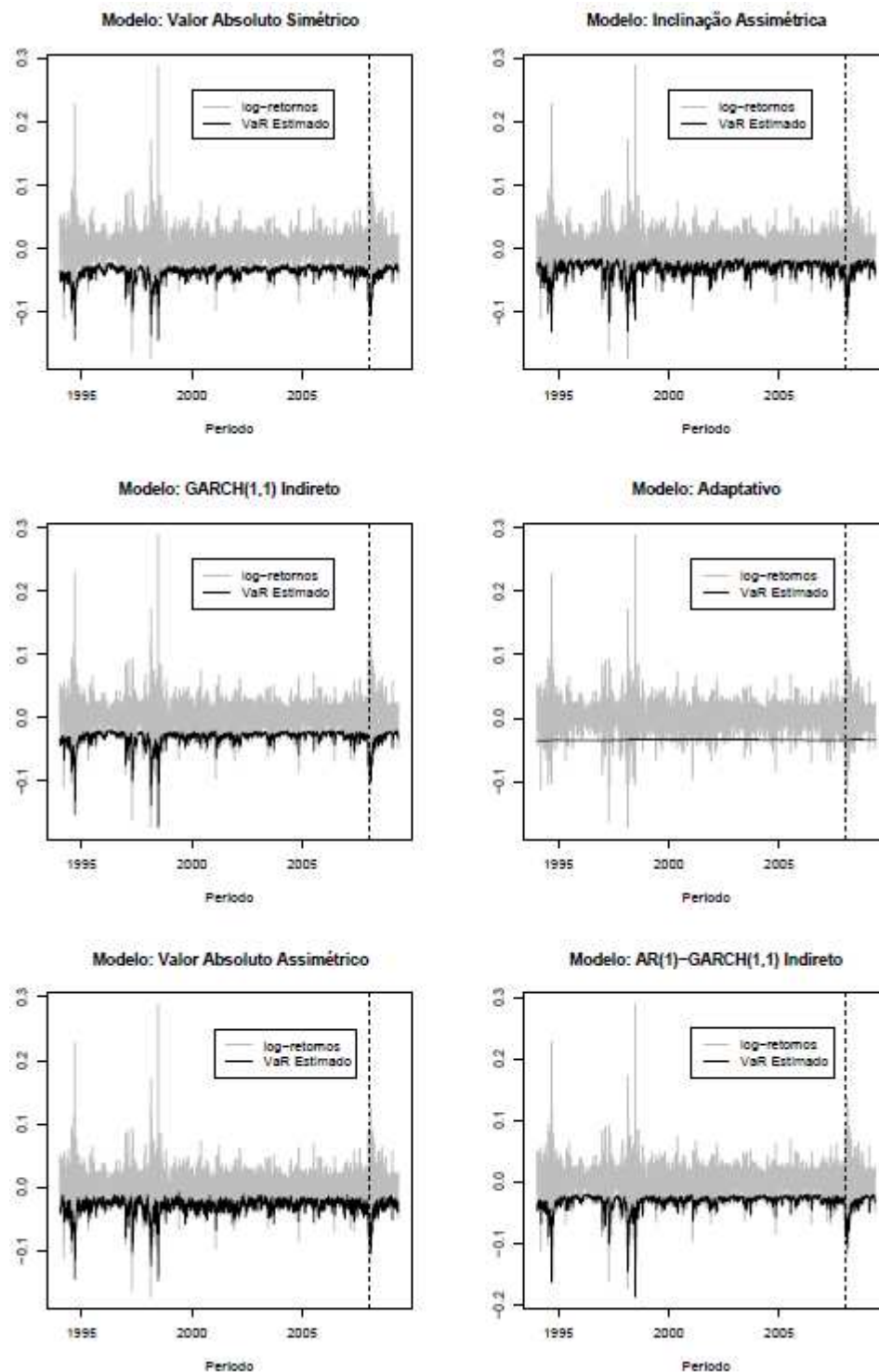


Tabela 4: Estimativas e Estatísticas dos Parâmetros Estimados Utilizando as Especificações do Modelo CARE

Estimativas e Testes			
Modelo	Adaptativo Sim.Prop.	Valor Absoluto Assimétrico	AR(1)-GARCH(1,1) Indireto
$\hat{\beta}_1$	-	-0,0055	0,7166
Erro Padrão	-	0,0672	0,2538
<i>P-valor</i>	-	0,4674	0,0024
$\hat{\beta}_2$	-	0,7905	0,7995
Erro Padrão	-	0,0295	0,0258
<i>P-valor</i>	-	0	0
$\hat{\beta}_3$	-	0,2743	0,3042
Erro Padrão	-	0,0304	0,1157
<i>P-valor</i>	-	0	0,0043
$\hat{\beta}_4$	-	0,0211	-
Erro Padrão	-	0,2021	-
<i>P-valor</i>	-	0	-
Hit_{am}	-	6,3076%	6,4267%
Hit_{pam}	-	8%	7,8%
QD_{am}	-	0,3946	$7,08.10^{-6}$
QD_{pam}	-	0,0164	0,0494
<i>MQA</i>	-	2.330,8	2.496,4
<i>CQ</i>	-	964,4457	987,6513

Todas as especificações do modelo CARE apresentam porcentagem de Hits pós-amostrais com valores maiores que os valores apresentados para os modelos CAViaR. De fato, essas observações não fazem parte do processo de estimação, além de pertencerem a um período de crise. Pelo fato da magnitude dos dados influenciarem mais nos modelos CARE, há um aumento da porcentagem dos Hit_{pam} . O valor dos teste de Quantis Dinâmicos pós-amostrais também indicam que, para períodos de instabilidade, como os identificados na série do IBOVESPA, os modelos CARE, aparentemente são menos indicados que os modelos CAViaR, para o cálculo do VaR.

5.3 Novas Especificações Propostas para os Modelos CARE

Para estas especificações, de forma mais prática, considerou-se a **Especificação 3**, com $p=5$ e $q=5$, e a **Especificação 4**, com $p=5$, em que $\theta=0,05$. Da mesma forma que os modelos CARE propostos por Taylor (2008), tem-se $\tau=0,024$.

A Figura 5 apresenta as estimativas para o VaR, de acordo com a **Especificação 3** e a **Especificação 4**, ambas utilizando os lags de ordem 5 para os log-retornos defasados no tempo.

Os gráficos mostram que os valores estimados por essa nova proposta, aparentemente não diferem muito dos valores estimados segundo as idéias de Taylor (2008). Porém, para obter uma análise mais detalhada, são apresentados os parâmetros estimados nas Tabelas 5 e 6, além do *Erro Padrão* e do *P-valor* de cada um dos parâmetros.

Figura 5: Estimativas para o VaR utilizando modelos CARE

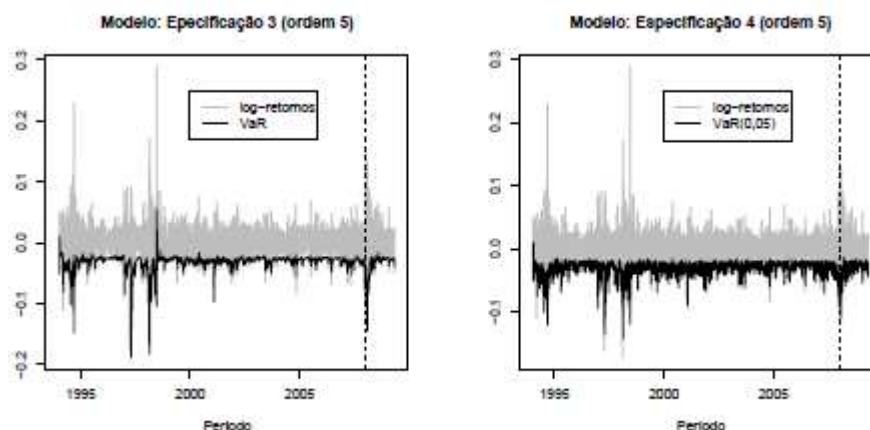


Tabela 5: Estimativas e Estatísticas dos Parâmetros Estimados Utilizando a Especificação 3 o Modelo CARE

Especificação 3			
$\hat{a}_0(\tau)$	0,2344	-	-
Erro Padrão	0,0081	-	-
<i>P – valor</i>	0	-	-
$\hat{a}_1(\tau)$	0,9050	-	-
Erro Padrão	0,0029	-	-
<i>P – valor</i>	0	-	-
$\hat{b}_1(\tau)$	-0,0225	$\hat{\gamma}_1(\tau)$	0,0198
Erro Padrão	0,0001	Erro Padrão	0,0002
<i>P – valor</i>	0	<i>P – valor</i>	0
$\hat{b}_2(\tau)$	0,0055	$\hat{\gamma}_2(\tau)$	0,0305
Erro Padrão	0,0001	Erro Padrão	0,0002
<i>P – valor</i>	0	<i>P – valor</i>	0
$\hat{b}_3(\tau)$	-0,0014	$\hat{\gamma}_3(\tau)$	-0,0240
Erro Padrão	0,0001	Erro Padrão	0,0002
<i>P – valor</i>	0	<i>P – valor</i>	0
$\hat{b}_4(\tau)$	0,0099	$\hat{\gamma}_4(\tau)$	0,0244
Erro Padrão	0,0001	Erro Padrão	0,0002
<i>P – valor</i>	0	<i>P – valor</i>	0
$\hat{b}_5(\tau)$	-0,0013	$\hat{\gamma}_5(\tau)$	-0,0165
Erro Padrão	0,0001	Erro Padrão	0,0002
<i>P – valor</i>	0	<i>P – valor</i>	0

Tabela 6: Estimativas e Estatísticas dos Parâmetros Estimados Utilizando a Especificação 4 do Modelo CARE.

Especificação 4			
$\hat{a}_0(\tau)$	-1,8544	-	-
Erro Padrão	0,0032	-	-
$P - valor$	0	-	-
$\hat{b}_1(\tau)$	-0,0984	$\hat{\gamma}_1(\tau)$	0,4421
Erro Padrão	0,0014	Erro Padrão	0,0014
$P - valor$	0	$P - valor$	0
$\hat{b}_2(\tau)$	0,0133	$\hat{\gamma}_2(\tau)$	0,2443
Erro Padrão	0,0014	Erro Padrão	0,0014
$P - valor$	0	$P - valor$	0
$\hat{b}_3(\tau)$	0,0291	$\hat{\gamma}_3(\tau)$	0,2369
Erro Padrão	0,0013	Erro Padrão	0,0014
$P - valor$	0	$P - valor$	0
$\hat{b}_4(\tau)$	0,1100	$\hat{\gamma}_4(\tau)$	0,2966
Erro Padrão	0,0013	Erro Padrão	0,0014
$P - valor$	0	$P - valor$	0
$\hat{b}_5(\tau)$	0,1097	$\hat{\gamma}_5(\tau)$	0,2418
Erro Padrão	0,0013	Erro Padrão	0,0014
$P - valor$	0	$P - valor$	0

Tabela 7: Hits Amostrais e Pós-amostrais e Valor do MQA

Modelo	Especificação 3	Especificação 4
Hit_{am}	6,5754%	6,5457%
Hit_{pam}	9%	7,8%
MQA	2.730,6	2.757,2

A um nível de significância de 5%, o teste apresentou a estatística

$$Q = 0,6783 \sim X^2_{(11)},$$

com $P - valor \approx 1$.

Já testando a **Especificação 4** com o modelo sob H_0 contra a **Especificação 3**, obteve-se a estatística:

$$Q_2 = 1,1843 \sim X^2_{(12)},$$

com $P - valor \approx 1$.

Os valores dos testes indicam que os dois modelos podem ser considerados como adequados neste cenário. Porém, visando um modelo que possua menos parâmetros e uma melhor proporção de Hit_{pam} , o modelo mais indicado neste cenário, utilizando modelos CARE de acordo com a proposta de Kuan, Yeh e Hsu (2009), seria o modelo descrito como **Especificação 4**, isto é:

$$\begin{aligned}\mu_t(0,024) = & -1,8544 - 0,0984r_{t-1}^+ + 0,0133r_{t-2}^+ + 0,0291r_{t-3}^+ + \\ & + 0,11r_{t-4}^+ + 0,1097r_{t-5}^+ + 0,4221r_{t-1}^- + 0,2443r_{t-2}^- + \\ & + 0,2369r_{t-3}^- + 0,2966r_{t-4}^- + 0,2418r_{t-5}^-.\end{aligned}$$

5.4 Estimativas para ES

A estimação do ES, utilizando modelos CARE é possível devido à ligação entre Mínimos Quadrados Assimétricos e a definição de ES.

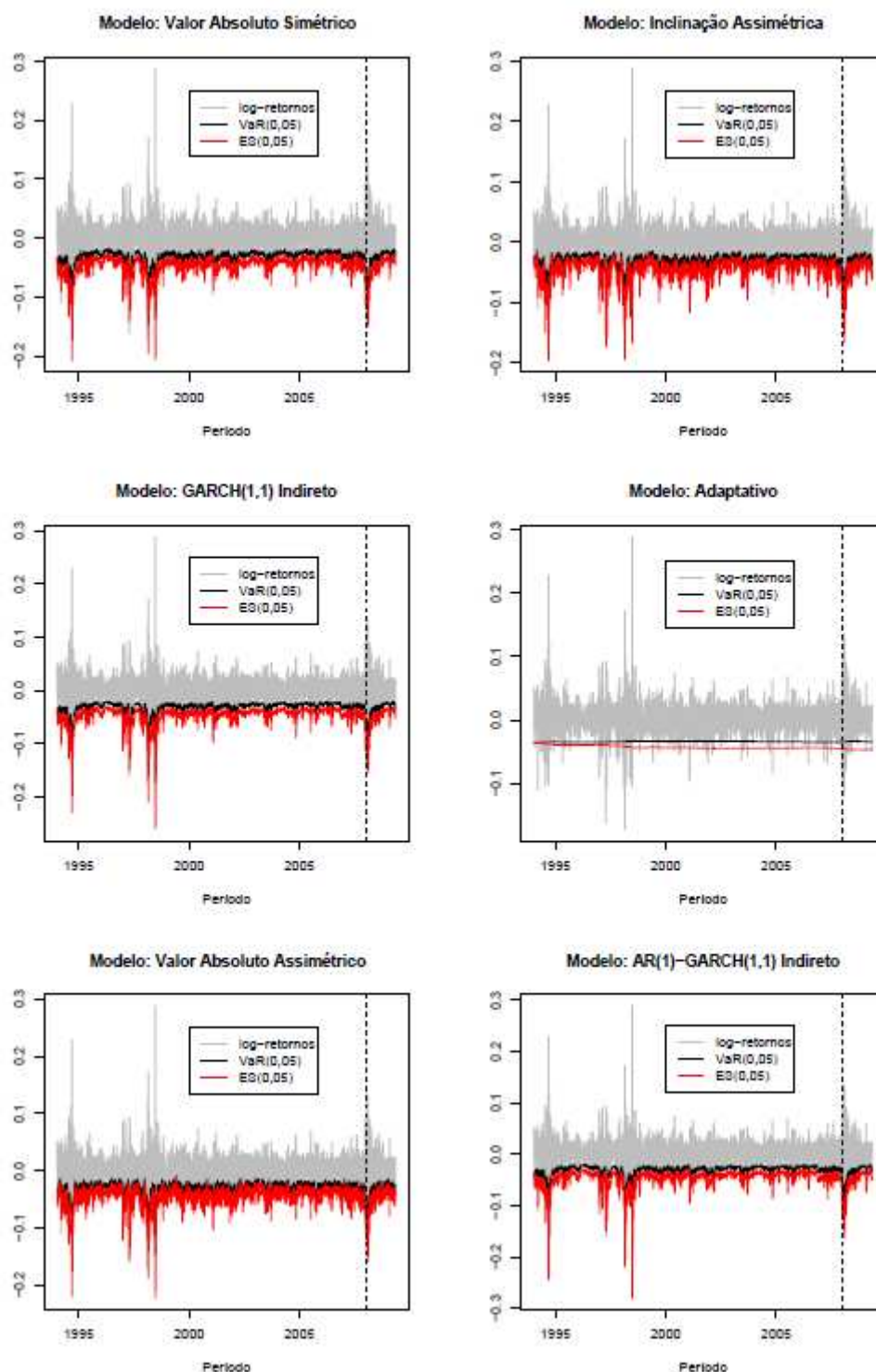
Assim, utilizando os modelos estimados para o τ -expectil de acordo com as especificações dos modelos CARE, conforme Taylor (2008), faz-se a estimação para o ES, cujas estimativas são apresentadas na Figura 6, em que se considerou $\theta = 0,05$ e $\hat{\tau} = 0,02436630$.

Tabela 8: Estatísticas do Teste Bootstrap para ES utilizando Abordagem Econométrica, TVE e modelos CARE

Teste Bootstrap para o ES			
Modelo	Estatística do Teste	Região Crítica	P-valor
Valor Absoluto Simétrico	-0,1062	$\leq -0,1449$	0,512
Inclinação Assimétrica	-0,0947	$\leq -0,1302$	0,534
GARCH(1,1) Indireto	-0,1119	$\leq -0,1470$	0,508
Adaptativo	0,3093	$\leq 0,2566$	0,493
Valor Absoluto Assimétrico	-0,09	$\leq -0,1261$	0,506
AR(1)-GARCH(1,1) Indireto	-0,0997	$\leq -0,1357$	0,502

Para esta série, todas as especificações são adequadas, pois todos os *P-valores* indicam a não rejeição de H_0 , ou seja, que todos os excessos padronizados tem média zero.

Figura 6: Gráficos de VaR e ES estimados através dos modelos CARE.



6. Conclusões

De forma geral, todos os métodos e modelos definidos, detalhados e discutidos são válidos para o cálculo do VaR e do ES. O objetivo principal do trabalho foi apresentar técnicas, não definidas como as melhores, mas as que apresentam resultados mais adequados dentro de determinados cenários, além de mostrar como modelos CAViaR e CARE podem se tornar tão precisos e adequados como os métodos mais simples e mais utilizados no mercado, porém sem fazer suposição alguma sobre a distribuição condicional das observações.

Referências bibliográficas

- Artzner, P., Delbaen, F., Eber, J. M. and Heath, D. (1999), Coherent Measures of Risk, *Mathematical Finance*, 9, 203-228.
- Efron, B. (1991), Regression Percentiles Using Asymmetric Squared Error Loss, *Estatística Sínica*, 1, 93-125.
- Engle, R. F. and Manganelli, S. (2004), CAViaR: Condicional Autoregressive Value at Risk by Regression Quantiles, *Journal of Business and Economic Statistics*, 22, 367-381.
- Jones, M. C. (1994), Expectiles and M-quantiles are Quantiles, *Statistics and Probability Letters*, 20, 149-153.
- Koenker, R. and Basset, G. (1978), Regression Quantiles, *Econometrica*, 46, 33-50.
- Koenker, R. W. (2005), *Quantile Regression*, Cambridge, UK: Cambridge University Press.
- Kuan, C., Yeh, J. and Hsu, Y. (2009), Assessing Value at Risk with CARE, the Conditional Autoregressive Expectile Models, *Journal of Econometrics*, 150, 261-270.
- Kuester, K., Mittnik, S. and Paolella, M., S. (2006), Value-at-Risk Prediction: A Comparison of Alternative Strategies, *Journal of Financial Econometrics*, 4, 53-89.
- McNeil, A. J. and Frey, R. (2000), Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach, *Journal of Empirical Finance*, 7, 271-300.
- McNeil, A. J., Frey, R. and Embrechts, P. (2005), *Quantitative Risk Management*, New Jersey; Princeton University Press.
- Newey, W. K. and Powell, J. L. (1987), Asymmetric Least Squares Estimation and Testing, *Econometrica*, 55, 819-847.
- Taylor, J. W. (2008), Estimating Value at Risk and Expected Shortfall Using Expectiles, *Journal of Financial Econometrics*, 6, 231-252.
- Weiss, A. (1991), Estimating Nonlinear Dynamic Models Using Least Absolute Error Estimation, *Econometric Theory*, 7, 46-68.

Agradecimento

Este trabalho foi financiado pela CAPES.

Abstract

Value at Risk (VaR) is a financial widely used risk measure to manage and control risk market. The measure risk estimation can be done by methods such as econometric approach and Theory of Extreme Values (TVE). However, these methods have some weaknesses such as inappropriate assumption of errors distribution, resulting in inaccurate measures. Therefore, in this article we discuss and compare Autoregressive Conditional Value at Risk (CAViaR) and Expected Autoregressive Conditional (CARE) models. We apply the proposed modeling approach to calculate the Value at Risk (VaR) and Expected Shortfall (ES) to the stock IBOVESOPA market index.

Keywords: Value at Risk, Expected Shortfall, CAViaR Models, CARE Models.

Aspectos socioeconômicos e ecológicos por escalonamento multidimensional e modelos de equações estruturais: Um Estudo de Caso no ZEE-MG

*Marcelo Angelo Cirillo¹
Lúcia Pereira Barroso²*

Resumo

O Zoneamento Ecológico Econômico do Estado de Minas Gerais – ZEE-MG é definido como um conjunto de normas e técnicas utilizadas para diagnosticar os meios geo-biofísico, socioeconômico, jurídico e institucional, gerando respectivamente duas cartas principais, a Carta de Vulnerabilidade Ambiental e a Carta de Potencialidade Social. Em particular, a Carta de Potencialidade Social é medida pelos componentes Produtivo, Natural, Humano e Institucional que, em síntese, caracteriza a capacidade de um município ou de uma microrregião alcançar o desenvolvimento sustentável. Este artigo tem por objetivo realizar um estudo das variáveis pertencentes a estes componentes, que possam ser sugeridas na construção de um modelo estrutural, plausível de explicar as relações de causas-efeitos. Assim, a técnica de escalonamento multidimensional é aplicada e posteriormente um modelo de equações estruturais é proposto, respeitando as diretrizes metodológicas propostas pelo Ministério do Meio Ambiente - MMA impostas na elaboração do ZEE, em conformidade com as diretrizes da Política e Legislação Ambiental do Estado de Minas Gerais.

Palavras-chave: equações estruturais, meio ambiente, potencialidade social.

¹ Universidade Federal de Lavras, Departamento de Ciências Exatas. E-mail: macufla@dex.ufla.br

² Universidade de São Paulo, Instituto de Matemática e Estatística. E-mail: lbarroso@ime.usp.br

1. Introdução

O Zoneamento Ecológico Econômico do Estado de Minas Gerais, nomeado por ZEE-MG, segundo Pereira (2008) representa um conjunto de informações capazes de fornecer uma perspectiva integrada e sintética da área estudada em inúmeros aspectos. Dentre eles, destaca-se a potencialidade social envolvendo setores nomeados por componentes: Produtivo, associado à infra-estrutura de transporte e atividades econômicas; Natural, associado à utilização das terras, estrutura fundiária e recursos minerais; Humano, associado à ocupação econômica, demografia e condições sociais e por fim, Institucional, envolvendo organizações jurídicas, financeiras, segurança pública, fiscalização e controle, ensino e pesquisa.

Este conjunto de informações, coletadas por municípios, permite compreender as principais tendências de uso do território, suas formas de produção e os modos e condições de vida a elas associados dentro do que preconiza a Agenda 21 brasileira: “que o desenvolvimento será construído sob uma ótica integradora que vê o território em estreita ligação com o capital humano” (AGENDA 21, 2002, p.26).

Além disso, a Carta de Potencialidade Social está em sintonia com as diretrizes metodológicas do Ministério do Meio Ambiente, que conceitua o ZEE-MG como “um instrumento político e técnico do planejamento, cuja finalidade última é otimizar o uso do espaço e as políticas públicas (MMA e SAE, 1997, p.12)”. Desta forma, essa carta tem a finalidade de compor uma grande base organizada e integrada de informações oficiais, fornecendo subsídios técnicos à definição de áreas prioritárias para a proteção e conservação da biodiversidade e para o desenvolvimento, segundo critérios de sustentabilidade econômica, social, ecológica e ambiental.

O ZEE-MG é de grande importância no planejamento e elaboração das políticas públicas e das ações em meio ambiente, orientando o governo e a sociedade civil na elaboração dos seus programas e em seus investimentos. Estes programas, ao serem planejados e implementados, respeitando-se as características de cada zona de desenvolvimento, podem promover com maior assertividade a melhoria na qualidade dos serviços prestados e na qualidade de vida de toda a população de Minas Gerais.

Neste trabalho ajustamos um modelo aos dados, considerando os componentes Natural, Humano e Institucional.

Segundo o projeto ZEE-MG, a formação do componente Produtivo considera indicadores relacionados ao transporte rodoviário, ferroviário e aeroviário, bem como outros indicadores, específicos dos setores energético e de comunicações. Devido à indisponibilidade dos dados, esse componente não foi considerado no modelo. Justificamos a sua ausência porque as interpretações poderiam ser incoerentes, uma vez que Vickerman (1995) menciona que a infraestrutura tem o caráter de um bem público, que reforça a produtividade dos fatores de produção do setor privado, podendo ser combinada ao capital privado a fim de elevar o potencial produtivo da localidade. Portanto, informações incompletas poderiam conduzir a interpretações incoerentes no modelo estrutural proposto.

Tendo por base essas informações, este artigo se propõe a realizar um estudo das variáveis pertencentes a estes componentes, que possam ser sugeridas na construção de um modelo estrutural, plausível de explicar as relações de causas-efeitos entre o conjunto de variáveis classificadas no componente Humano, Natural e Institucional por meio das técnicas de escalonamento multidimensional em conjunto com a modelagem de equações estruturais (MEE).

A realização dessas análises é importante porque os resultados estatísticos obtidos conjuntamente das variáveis classificadas por meio destes componentes e não de forma fragmentada poderá agregar conhecimentos de dada realidade municipal do Estado de Minas Gerais, proporcionando informações que possam subsidiar a tomada de decisões relativa à contribuição do desenvolvimento sustentável. Desta forma, as potencialidades sociais dos municípios e regiões, no sentido de identificar e apontar aqueles ambientes que estão fragilizados ou vulneráveis à ação do homem, e as capacidades de que o próprio homem dispõe sobre estes ambientes, poderão ser contemplados na adoção de políticas estratégicas para alocação de recursos.

2. Material e Métodos

2.1 Descrição das variáveis e legalidade do ZEE-MG

Em outubro de 2005 foi firmado convênio entre a Secretaria de Meio Ambiente e Desenvolvimento Sustentável (SEMAD), o Instituto Estadual de Florestas (IEF) e a Universidade Federal de Lavras (UFLA), com a interveniência da Fundação Estadual de

Meio Ambiente (FEAM), do Instituto Mineiro de Gestão das Águas (IGAM) e da Fundação de Apoio ao Ensino, Pesquisa e Extensão (FAEPE). Houve ainda uma expressiva atuação da Fundação João Pinheiro (FJP) e da Del Rey Engenharia, respectivamente Instituição Estadual e Empresa contratadas pela UFLA/FAEPE, parceiras importantes na construção do Zoneamento Ecológico-Econômico (ZEE). Sendo assim, o trabalho desenvolvido como objeto deste convênio, insere-se no conjunto de ações previstas e aprovadas no âmbito do referido Projeto Estruturador visando à obtenção dos resultados previstos no Acordo de Resultados firmado entre o Governo de Estado, Secretaria de Estado de Meio Ambiente e Desenvolvimento Sustentável (SEMAD) e órgãos vinculados.

Em relação à coleta das observações, segundo informações do projeto ZEE-MG (Scolforo et al., 2008), foram utilizados dados censitários, do Censo Demográfico de 2000, sobre as condições produtivas, humanas, naturais e institucionais de cada município do Estado de Minas Gerais, classificados em 5 regiões, conforme Tabela 1. Desta forma, todos os 853 municípios de Minas Gerais foram considerados neste estudo.

Tabela 1: Número de municípios por região do Estado de Minas Gerais.

Regiões	Descrição	Número de municípios por região
1	Região do Rio Doce	135
2	Região do Sul/Sudoeste	176
3	Região do Triângulo Mineiro / Alto Paranaíba	72
4	Região Central	60
5	Região da Zona da Mata	410

A caracterização das variáveis utilizadas encontra-se descrita nas Tabelas 2 a 4.

Tabela 2: Resumo descritivo das variáveis do componente Natural.

Codificação	Descrição	Definição
CN_I1	Densidade de ocupação econômica da terra	Em sua composição considera as variáveis: área cultivada, ocupada por unidades de preservação e aquelas inundadas para geração de energia elétrica. Este indicador além de permitir o acompanhamento indireto da produtividade, possibilita observar os prováveis riscos de contaminação dos solos pela utilização de fertilizantes químicos.
CN_I2	Nível tecnológico da agropecuária	Informa a respeito do nível tecnológico adotado pela agricultura nos diferentes municípios do estado, utilizando como referência o rendimento médio da cultura do milho.
CN_I3	Proporção entre o número total de agricultores em relação ao total do estado	Informa a proporção do número total de agricultores familiares do município em relação ao número total do estado.
CN_I4	Utilização das terras	Fornece informações referentes à intensidade de uso dos solos por atividades agrossilvopastoris e, também, à implementação de práticas e tecnologias que intensificam a produção e o rendimento das lavouras.
CN_I5	Estrutura fundiária	Mensura a concentração de terras entre os diversos segmentos, bem como a concentração de agricultores familiares, correlacionando-as com os dados referentes à distribuição das terras.
CN_I6	Recursos minerais	Calculado sobre o valor do faturamento líquido, obtido por ocasião da venda do produto mineral. Entende-se por faturamento líquido o valor da venda do produto mineral, deduzindo-se os tributos (ICMS, PIS, COFINS) que incidem na comercialização, como também as despesas com transporte e seguro.

Tabela 3: Resumo descritivo das variáveis do componente Institucional

Codificação	Descrição	Definição
CI_I1	Capacidade de aplicação da lei	Informa a respeito da capacidade do estado em efetivar a lei, através de informações sobre o número de operadores de cada agência que compõe o sistema de justiça criminal em relação ao número de moradores do município.
CI_I2	Gestão municipal	A composição desta variável considera a metodologia sugerida para o cálculo do Índice de Qualidade Institucional Municipal (Haddad, 2004) e a pesquisa de Informações Básicas Municipais do IBGE (IBGE, 2005).
CI_I3	Organizações de ensino e pesquisa	Esta variável fornece informações a respeito de instituições de ensino superior e de ensino profissionalizante, além de organizações de pós-graduação e de pesquisa em um município.
CI_I4	Organizações de segurança pública	Esta variável é composta por outras variáveis definidas por Unidades de Defesa Social, que avalia a presença, nos municípios, de organizações que promovem a segurança pública, e a Capacidade de Aplicação da Lei, que avalia a capacidade do município de reprimir atos contra a segurança pública.

Tabela 4: Resumo descritivo das variáveis do componente Humano

Codificação	Descrição	Definição
CH_TXOCP	Taxa de ocupação	Informa a respeito do mercado de trabalho relativo às tendências e variações no nível de ocupação, permitindo subsidiar a formulação de estratégias e políticas de geração de emprego e renda.
CH_PPU	Distribuição percentual da população urbana	Informa o percentual da população urbana dada pelo quociente entre o número da população urbana e o número total da população.
CH_I1	Percentual de pessoas que vivem em domicílios e terrenos próprios e quitados	Informa o percentual de pessoas que no ano 2000 viviam em domicílios que, juntamente com os terrenos onde se localizam, eram de propriedade total ou parcial de um dos moradores e já foram totalmente pagos. No caso de apartamentos, o critério estabelecido pelo ZEE-MG foi considerar a fração do terreno.
CH_I2	Percentual de pessoas que vivem em município com água encanada	Informa o percentual de pessoas que em 2000 viviam em domicílios com água canalizada para um ou mais cômodos proveniente de rede geral, de poço, de nascente ou de reservatório abastecido por água das chuvas ou carro-pipa.
CH_I3	Percentual de pessoas que vivem em município com banheiro e água encanada	Informa o percentual de pessoas que em 2000 viviam em domicílios com água encanada em pelo menos um de seus cômodos e com banheiro definido como cômodo que dispõe de chuveiro ou banheira e aparelho sanitário.
CH_I4	Percentual de pessoas que vivem em domicílios urbanos com serviço de coleta de lixo	Informa o percentual de pessoas que no ano 2000 viviam em domicílios localizados nas áreas urbanas, em que a coleta de lixo é realizada diretamente por empresa pública ou privada, e o lixo é depositado em local apropriado para coleta posterior.
CH_I5	Estado de saúde - taxa bruta de mortalidade - TBM ponderada	Informa o estado de saúde geral da população mensurado em 2004, considerando a Taxa Bruta de Mortalidade (TBM) definida pelo número total de óbitos por mil habitantes, buscando medir diretamente a saúde ou a sua falta.
CH_I6	Acesso e utilização dos serviços de saúde	Informa sobre os serviços de saúde prestados à população em 2004, em termos de acessibilidade e utilização, considerando a atenção primária à saúde, através do atendimento às gestantes e a assistência médico-hospitalar.
CH_EDU	Taxa de educação	Esta variável é composta por três índices: Escolaridade, Qualidade do Ensino e Acesso e Utilização dos Serviços. O peso de cada índice é de 33,33%.

Posteriormente, após obtidos os resultados sobre a similaridade das regiões, revelados com o uso do escalonamento multidimensional, apenas as observações dos municípios das regiões selecionadas foram consideradas para o ajuste do modelo de equações estruturais.

Considerando as variáveis descritas nas Tabelas 2 a 4, é importante ressaltar que elas apresentam escalas diferentes. Sendo assim, para que os resultados estatísticos não fossem comprometidos por diferenças de escalas, optou-se por fazer a transformação dada pela razão entre (o valor observado menos o menor valor) e (o maior valor menos o menor valor). Desta forma, todas as variáveis apresentaram valores em uma escala entre 0 e 1. Optou-se por esta transformação para garantir que todas as variáveis estejam em uma mesma escala. Neste contexto, a arbitrariedade na escolha da transformação é justificada, por facilidade na convergência dos procedimentos iterativos e não por razões teóricas.

Mantendo os dados transformados, realizou-se uma análise exploratória multidimensional com o propósito de identificar a similaridade entre as variáveis, de acordo com as regiões estudadas no ZEE-MG. Com este propósito utilizou-se a técnica de escalonamento multidimensional, descrita na Seção 2.2.1.

2.2 Metodologia estatística

2.2.1 Escalonamento multidimensional

Escalonamento multidimensional é uma técnica que permite representar um conjunto de pontos em um espaço de dimensão menor do que a original, mantendo as distâncias entre os pontos nesse espaço tão próximas quanto possível das dissimilaridades originais dos dados.

Inicialmente, considerou-se a matriz de dissimilaridades $\Delta = [\delta_{ij}]$ de ordem $(n \times n)$, em que cada elemento indica a dissimilaridade entre o *i*-ésimo e o *j*-ésimo objeto representado pelas regiões numeradas de acordo com a Tabela 1.

O processo de redução multidimensional se deu com a obtenção dos vetores de coordenadas, definidos por $x_i = (x_{i1}, \dots, x_{iq})$ de ordem $(n \times q)$. Tendo por base a matriz de distâncias euclidianas $D = [d_{ij}]$, de ordem $(n \times n)$, formada a partir de um conjunto de

dados, no qual cada elemento foi obtido calculando-se $d_{ij}^2 = \|x_i - x_j\|^2$ com $i = 1, \dots, n$ e $j = 1, \dots, n$, verificou-se a proximidade entre as matrizes D e Δ , isto é, se $d_{ij} \approx \delta_{ij}$.

A adequação desta aproximação foi confirmada. Como critério estatístico foi utilizada a função *Stress* (De Leeuw, 1988), com a finalidade de avaliar a diferença entre a matriz de dissimilaridades original e a matriz de distâncias calculadas.

Foi aplicado o escalonamento multidimensional não-métrico como definido por Jaworska e Anastasova (2009) e tendo por base o algoritmo que considera a escolha de uma função monótona para representar a relação entre as distâncias d_{ij} e δ_{ij} .

O algoritmo utilizado na minimização da função *Stress* (Kruskal, 1964) consiste em substituir a função $f(\delta_{ij}) = d_{ij}$ por uma função $g(\delta_{ij}, \delta_{ij}^*)$, fixado um valor representado por δ_{ij}^* . A escolha da função $g(\delta_{ij}, \delta_{ij}^*)$ deve ser feita obedecendo-se as restrições $f(\delta_{ij}) \leq g(\delta_{ij}^*, \delta_{ij}^*)$ e $f(\delta_{ij}^*) = g(\delta_{ij}^*, \delta_{ij}^*)$. Desta forma, o mínimo da função *Stress* é obtido por meio do algoritmo dado nos seguintes passos:

- 1) Definir, arbitrariamente, um valor para $\delta_{ij}^* = \delta_{ij0}^*$, considerado como um “chute” inicial do algoritmo;
- 2) Encontrar δ_{ij}' , para o qual $g(\delta_{ij}', \delta_{ij}^*) \leq g(\delta_{ij}^*, \delta_{ij}^*)$;
- 3) Fixar $f(\delta_{ij}^*) - f(\delta_{ij}') < \alpha$, sendo α a margem de erro do critério de convergência;
- 4) Encerrar o algoritmo se a desigualdade em 3) for satisfeita. Se isso não acontecer, redefinir δ_{ij}^* por $\delta_{ij}^* = \delta_{ij}'$ e repetir os passos 2 a 4 até que o algoritmo alcance a convergência no passo 3).

Convém ressaltar que a relação $f(\delta_{ij}) = d_{ij}$ pode ser verificada com base em análises descritivas, como por exemplo, um gráfico scatterplot entre as distâncias δ_{ij} e d_{ij} (diagrama de Shepard).

Seguindo as recomendações de Borg e Groenen (2005), a interpretação dos resultados relativos à função *Stress* é a de que quanto menor for o seu valor, melhor será o ajuste da matriz de distâncias reproduzidas (Δ) à matriz de distâncias observadas (D).

Após a execução desse procedimento, gráficos biplots foram construídos utilizando resultados algébricos da análise de componentes principais (maiores detalhes podem ser vistos em Gower e Hand (1996)). A identificação dos objetos a serem avaliados em relação à similaridade no gráfico dos biplots é feita por meio da técnica de

coordenadas principais, estimadas segundo o procedimento sugerido por Borg e Groenen (2005).

Após a identificação das regiões similares, utilizou-se uma análise exploratória para estimar as correlações entre as variáveis selecionadas com a aplicação do escalonamento multidimensional. Desta forma, tornou-se possível postular um modelo teórico para estudo das relações socioeconômicas e ter sua validade confirmada por meio da modelagem de equações estruturais, na qual considera-se um modelo teórico a ser validado (Seção 2.2.3).

2.2.2 Considerações para a imposição de um modelo teórico a ser utilizado na modelagem de equações estruturais

A principal suposição feita sobre o modelo estrutural é assumir que as relações entre as variáveis são lineares. Tal suposição é puramente empírica, uma vez que a questão da causa, segundo Pilati e Laros (2007), está diretamente relacionada ao tipo de delineamento utilizado pelo pesquisador em sua investigação e não à metodologia estatística.

A questão das relações entre as variáveis também é norteadas por princípios filosóficos, como mencionam Shadish et al. (2002), ao afirmarem que um simples experimento aleatório caracteriza a ocorrência de um efeito. Neste contexto, o uso da modelagem causal, sendo o modelo de equações estruturais uma alternativa, nos ajuda a explicar o efeito de possíveis relações entre as variáveis independentes mediadoras (latentes) e as variáveis independentes.

Hox e Bechger (1998) argumentam que os pesquisadores devem ter em mente que, se um modelo é corroborado pelos dados em estudos de MEE, isso não implica que esse modelo tenha sido provado como verdadeiro, mas sim que o modelo não foi rejeitado. Os autores ressaltam que isso não implica na ausência de outros modelos concorrentes que possam, também, ser não rejeitados.

Assim, podemos afirmar que o modelo estrutural proposto neste trabalho é empírico, bem como a especificação das relações lineares entre os efeitos. Entretanto, dada a confirmação pelos índices de qualidade de ajuste e a interpretação de cada construto, pode-se considerar o modelo estrutural adequado, porém, não único para explicar todas as relações do complexo projeto ZEE-MG, tendo em vista o expressivo número de variáveis e relações diretas e indiretas entre as mesmas.

Seguindo estas justificativas, na Seção 2.2.3 apresentamos a definição do modelo de equações estruturais, seguindo a notação encontrada em Kaplan (2000).

2.2.3 Definição do modelo de equações estruturais

O modelo de equações estruturais é dado pelo sistema de equações (1) a (3).

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta} \quad (1)$$

$$\mathbf{Y} = \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (2)$$

$$\mathbf{X} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\delta} \quad (3)$$

Em (1), $\boldsymbol{\eta}_{m \times 1}$ representa o vetor de variáveis latentes endógenas; $\boldsymbol{\xi}_{r \times 1}$ o vetor de variáveis latentes exógenas; $\boldsymbol{\zeta}_{m \times 1}$ o vetor dos erros estruturais; $\mathbf{B}_{m \times m}$ a matriz dos coeficientes que relaciona as variáveis latentes endógenas e por fim, $\boldsymbol{\Gamma}_{m \times r}$ a matriz dos coeficientes que relaciona as variáveis latentes exógenas com as variáveis latentes endógenas.

Especificamente sobre o modelo de mensuração $\mathbf{Y}$, conforme a expressão (2), a relação das variáveis latentes endógenas com as variáveis observadas y_j ($j=1, \dots, p$) é determinada através da matriz dos coeficientes $\boldsymbol{\Lambda}_{y(p \times m)}$, sendo o erro de medida representado por $\boldsymbol{\varepsilon}_{(p \times 1)}$.

A expressão (3) refere-se ao modelo de mensuração $\mathbf{X}$, utilizado para relacionar as variáveis latentes exógenas com as variáveis observadas x_k ($k=1, \dots, q$), especificado por meio da matriz dos coeficientes $\boldsymbol{\Lambda}_x(q \times r)$, sendo o erro de mensuração representado pelo vetor $\boldsymbol{\delta}_{(q \times 1)}$.

A estimação dos parâmetros do modelo representado pelas expressões (1) a (3) e a interpretação das relações entre as variáveis, tendo por base os resultados obtidos no escalonamento multidimensional, são apresentadas na Seção 3 estruturada em: 3.1- Biplot com escalonamento multidimensional e 3.2- Descrição do modelo de equações estruturais para estudo do ZEE-MG nas regiões Sul/Sudoeste e Central do Estado de Minas Gerais.

3. Resultados e Discussões

3.1 Biplot com escalonamento multidimensional

Sob a hipótese de que as variáveis pertencentes aos componentes Humano e Natural apresentam efeitos diferenciados nas várias regiões do estado, torna-se coerente buscar uma confirmação da possível existência da heterogeneidade entre as regiões, considerando simultaneamente todas as variáveis envolvidas.

Com este enfoque, inicialmente, construímos o gráfico biplot (Figura 1), utilizando técnicas de escalonamento multidimensional, conforme descrito na Seção 2.2.1. A princípio, sem a preocupação com a redução do espaço dimensional, convém salientar que as projeções ortogonais envolvidas na obtenção dos eixos preditivos dos biplots foram calibrados em intervalos de valores iguais. Desta forma, os eixos obtidos apresentam escalas iguais.

Para verificar a proximidade entre as distâncias euclidianas preditas e as dissimilaridades observadas, utilizou-se o diagrama de Shepard (Shepard, 1962), em que o eixo horizontal indica as dissimilaridades originais entre regiões e o eixo vertical as distâncias correspondentes. Desta forma, cada ponto marcado na Figura 2 representa um par de regiões, e os cinco que apresentaram maiores desvios em relação à reta de igualdade das medidas foram identificados nos gráficos como “quadrados”. Logo, há evidências para afirmar que, supostamente, a análise conjunta das variáveis, incluindo todas as regiões, poderá proporcionar resultados incoerentes.

Tendo por base os resultados discutidos anteriormente, novos gráficos biplots foram construídos, separadamente para as variáveis dos componentes Natural e Humano (Figuras 3 e 4, respectivamente). Especificamente ao componente Institucional, dado o número de variáveis inferior em relação aos demais componentes, julgou-se desnecessária a aplicação da técnica de escalonamento multidimensional.

Os resultados ilustrados por meio da Figura 3 permitem verificar que as variáveis CN_I3, CN_I2 e CN_I5, pertencentes ao componente Natural (Tabela 2 - Seção 2.2.1), foram as que apresentaram melhor desempenho na formação dos agrupamentos.

– Seção 2.2.1).

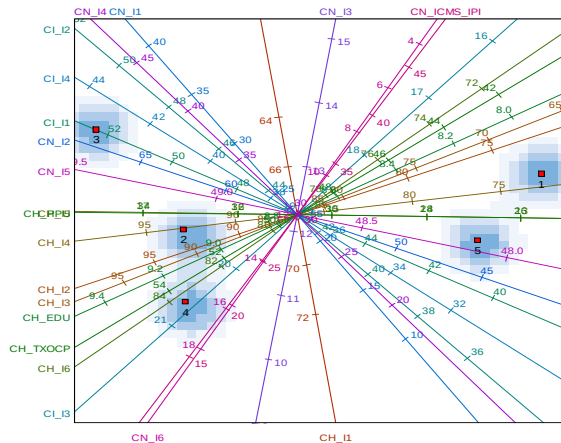


Figura 1– *Biplot* com escalonamento multidimensional considerando todas variáveis pertinentes aos componentes Humano, Institucional e Natural.

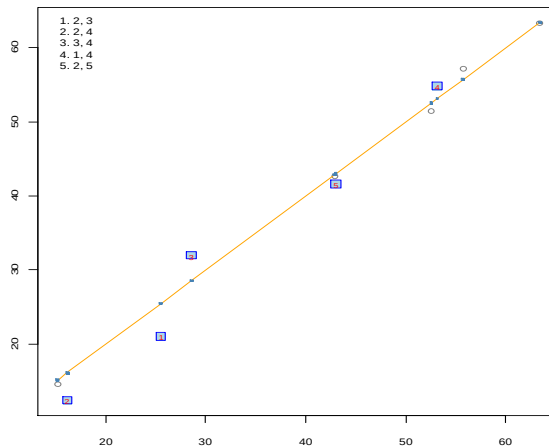


Figura 2 – Diagrama de Shepard para as regiões avaliadas no ZEE-MG, considerando a amostra multivariada formada pelas variáveis dos componentes Humano, Institucional e Natural.

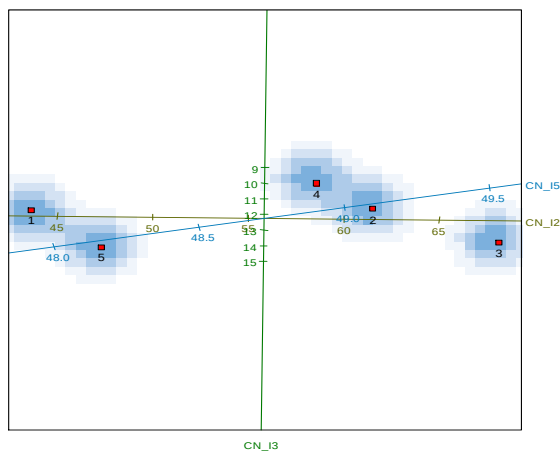


Figura 3 – Biplot por escalonamento multidimensional para os dados do componente Natural.

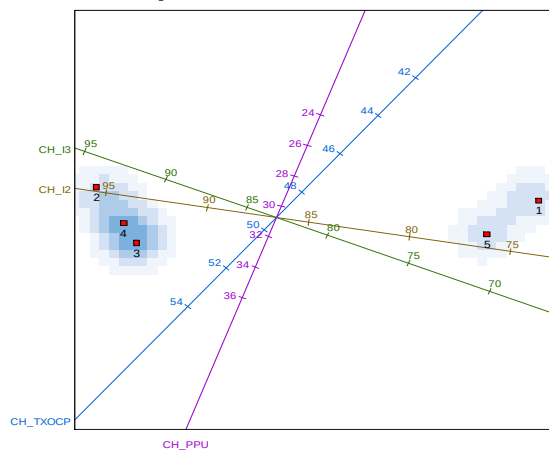


Figura 4 – Biplot por escalonamento multidimensional para os dados do componente Humano.

Conforme a descrição encontrada na Tabela 5, observou-se que tanto no componente Humano como no Natural, as regiões 2 (Região do Sul/Sudoeste) e 4 (Região Central) apresentaram similaridades, assim como as regiões 1 (Região do Rio Doce) e 5 (Região da Zona da Mata).

Tabela 5: Regiões agrupadas pelo escalonamento multidimensional

Componente	Regiões agrupadas
Humano	(1,5) ; (2,3,4)
Natural	(1, 5) ; (2, 4) e (3)

1: Rio Doce; 2: Sul/Sudoeste; 3: Triângulo Mineiro / Alto do Paranaíba; 4: Central e 5: Zona da Mata.

A avaliação da qualidade de ajuste da redução dimensional, proporcionada pelo escalonamento multidimensional realizado nos componentes Humano e Natural, foi feita por meio dos resultados da função Stress, mencionada na Seção 2.2.1, conforme ilustram as Figuras 5 e 6.

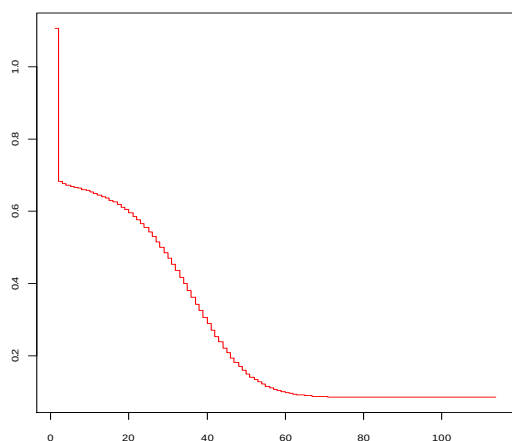


Figura 5: Gráfico da função Stress versus número de iterações aplicado na verificação da qualidade do ajuste do algoritmo MDS às variáveis do componente Humano.

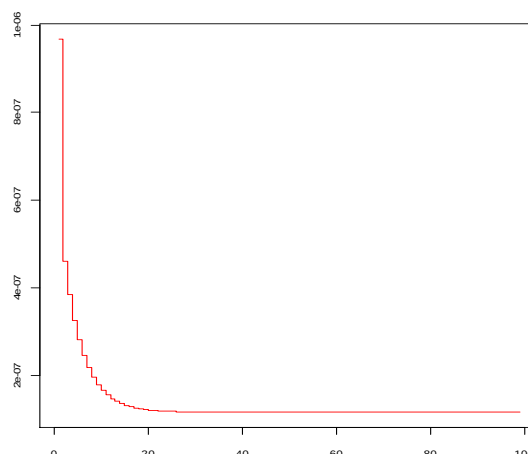


Figura 6: Gráfico da função Stress versus número de iterações aplicado na verificação da qualidade do ajuste do algoritmo MDS às variáveis do componente Natural.

Os gráficos das Figuras 5 e 6 são construídos considerando as iterações versus os valores obtidos na função *Stress*. Em relação ao componente Humano (Figura 5), percebe-se que após a 60ª iteração, os valores resultantes dessa função tendem a se estabilizar em um valor inferior a 0,10. No caso do componente Natural (Figura 6), essa convergência foi mais rápida, resultando em valores próximos a zero.

Decorrente dos resultados obtidos, pode-se afirmar que os agrupamentos formados (Tabela 5), resultantes da aplicação do método biplot com escalonamento multidimensional, foram bons, considerando os critérios estabelecidos por Borg e Groenen (2005), que descrevem a seguinte classificação para o ajuste segundo os valores da função Stress: maior ou igual a 0,20 (pobre); entre 0,20 e 0,10 (aceitável); entre 0,10 e 0,05 (bom) e menor ou igual a 0,025 (ótimo) e menor ou igual a 0,025 (excelente).

Em virtude do que foi mencionado, a análise inferencial apresentada neste artigo é dada apenas para as regiões 2 e 4. Entretanto, a modelagem de equações estruturais pode ser utilizada para construção de novos modelos para as demais regiões, com diferentes formulações do modelo teórico, mantendo-se a coerência, conforme as recomendações de Hox e Bechger (1998) de que o processo de especificação deve ser direcionado pela combinação de elementos teóricos e de evidências empíricas de pesquisas anteriores.

3.2 Descrição do modelo de equações estruturais para o estudo do ZEE-MG nas regiões Sul/Sudoeste e Central do Estado de Minas Gerais

Convém ressaltar que o MDS foi aplicado com o propósito de indicar agrupamentos entre as regiões em função de um número reduzido de variáveis. Procedeu-se, inicialmente, com uma análise descritiva das variáveis envolvidas.

A matriz de correlações entre as variáveis endógenas e exógenas observadas é apresentada na Tabela 6.

Tabela 6: Matriz de correlações das variáveis observadas possíveis de serem consideradas no ajuste do modelo.

Variáveis	CH_I2	CH_I3	CH_TXOCP	CH_PPU	CN_I2	CN_I3	CN_I5	CI_I1	CI_I2	CI_I3	CI_I4
CH_I2	1,000	0,999	0,968	0,922	0,923	-0,409	0,938	0,823	0,609	0,886	0,771
CH_I3	0,999	1,000	0,969	0,904	0,911	-0,418	0,928	0,797	0,583	0,883	0,743
CH_TXOCP	0,968	0,969	1,000	0,910	0,902	-0,316	0,928	0,792	0,669	0,963	0,746
CH_PPU	0,922	0,904	0,910	1,000	0,949	-0,268	0,955	0,972	0,817	0,896	0,949
CN_I2	0,923	0,911	0,902	0,949	1,000	-0,072	0,997	0,913	0,840	0,824	0,902
CN_I3	-0,409	-0,418	-0,316	-0,268	-0,072	1,000	-0,102	-0,181	0,313	-0,317	-0,079
CN_I5	0,938	0,928	0,928	0,955	0,997	-0,102	1,000	0,907	0,830	0,857	0,891
CI_I1	0,823	0,797	0,792	0,972	0,913	-0,181	0,907	1,000	0,857	0,788	0,993
CI_I2	0,609	0,583	0,669	0,817	0,840	0,313	0,830	0,857	1,000	0,696	0,897
CI_I3	0,886	0,883	0,963	0,896	0,824	-0,317	0,857	0,788	0,696	1,000	0,747
CI_I4	0,771	0,743	0,746	0,949	0,902	-0,079	0,891	0,993	0,897	0,747	1,000

Por meio dos resultados descritos na Tabela 6, nota-se que, de modo geral, praticamente todas as variáveis apresentam grau de correlação razoável para serem consideradas no ajuste, exceto a variável CN_I3. Decorrente deste resultado e por problemas de natureza numérica em relação à convergência do método de máxima verossimilhança por ocasionar variâncias negativas, conhecidos na literatura como casos Heywood, optou-se por não considerar esta variável no modelo.

Dada a proposta do programa ZEE-MG e respeitando os critérios socioeconômicos e ecológicos, a construção de um modelo de equações estruturais foi fundamentada no modelo teórico da Figura 7, em que as setas curvas bidirecionais são interpretadas como covariâncias e as setas unidirecionais como relações causais. A interpretação do modelo é feita por meio das setas que sugerem relações de linearidade entre as variáveis latentes e observadas.

Sob esse aspecto, Hershberger et al. (2003) mencionam que as relações são avaliadas pela magnitude do efeito (direto ou indireto) que as variáveis independentes (observadas ou latentes) têm sobre as variáveis dependentes (observadas ou latentes).

Contextualizando a confirmação destas relações “teóricas” de dependência estatística, por meio da significância das estimativas dos parâmetros do modelo ajustado para os dados do Zoneamento Ecológico Econômico do Estado de Minas Gerais, consideraram-se as seguintes variáveis latentes: construto latente exógeno (ξ) denominado “Questões Habitacionais” considerado como fonte causadora na relação com os demais construtos endógenos; construto latente endógeno (η_1) denominado “Questões Agrárias”; construto latente endógeno (η_2) denominado “Questões de Segurança” e por fim, construto latente endógeno (η_3) denominado “Questões de Ocupação Econômica e Cobertura da População Urbana”. A especificação do modelo e as relações causais encontram-se ilustradas na Figura 7.

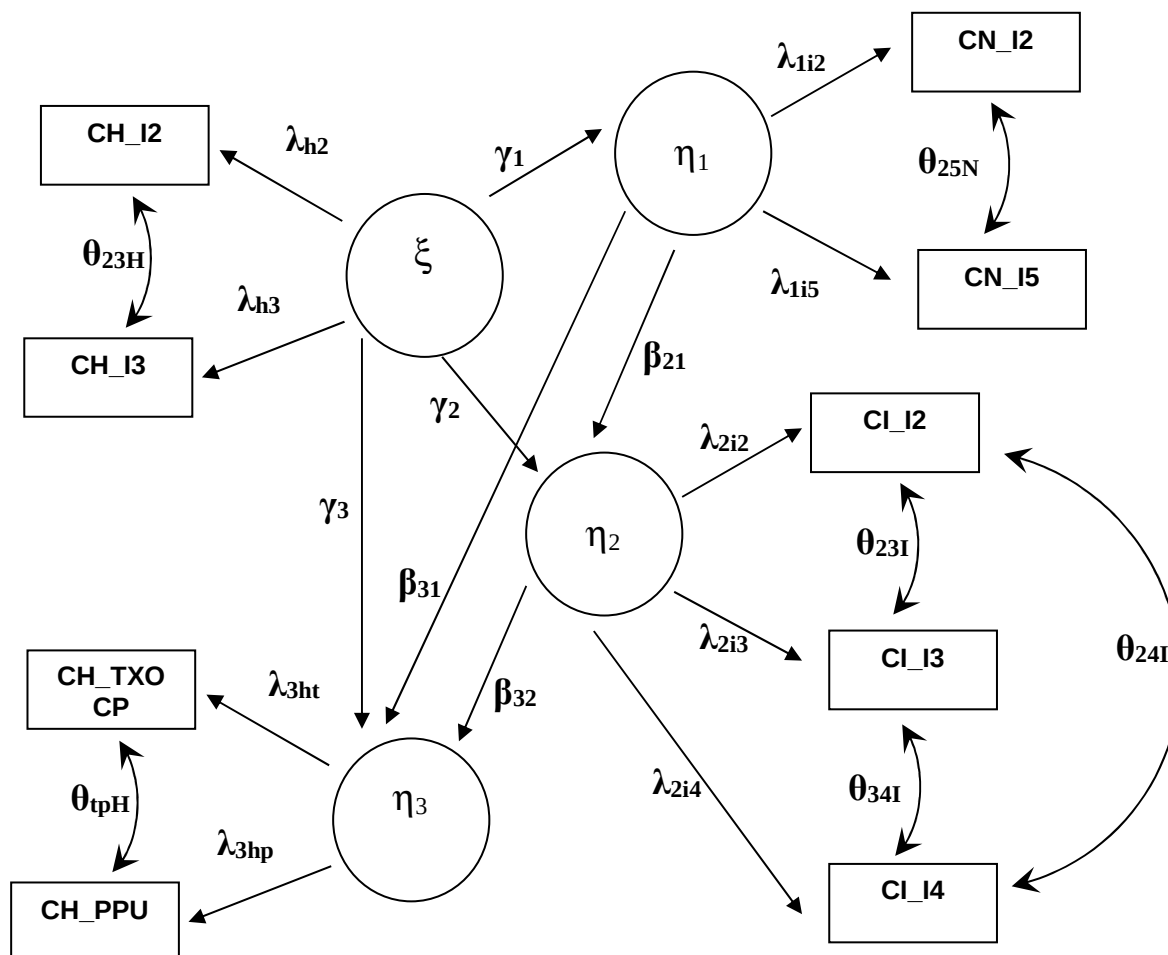


Figura 7: Diagrama do modelo de equações estruturais proposto para estudar as relações das variáveis dos componentes Humano, Natural e Institucional inseridas no ZEE-MG para as regiões Sul/Sudoeste e Central.

A validação do modelo proposto é confirmada pelos critérios estatísticos descritos na Tabela 7.

Tabela 7: Resultados dos índices de qualidade de ajuste de MEE

Índices de qualidade de ajuste	valor
Razão entre a estatística qui-quadrado e o número de graus de liberdade (χ^2/gl)	1,3867
Goodness-of-Fit Index (GFI)	0,9712
Adjusted Goodness-of-Fit Index (AGFI)	0,9451
Root Mean Square Error of Approximation (RMSEA)	0,0403
Normalized Fit Index (NFI)	0,9504
Non Normed Fit Index (NNFI)	0,9783
Comparative Fit Index (CFI)	0,9855
Standardized Root Mean Residual (SRMR)	0,0615

Mantendo as especificações dos valores de referência (Ullman, 2007) citados a seguir, na interpretação de cada índice de ajuste, em consonância com os resultados encontrados na Tabela 7, pode-se inferir que o modelo proposto (Figura 7) ajustou-se adequadamente aos dados amostrais.

A comprovação desta afirmativa é fundamentada nos resultados estatísticos proporcionados pelos índices GFI, AGFI, CFI, NFI próximos ao valor unitário, sendo este considerado como referência. Segundo Ullman (2007), os testes fundamentados nesses índices, que têm os valores de referência definidos entre 0,90 e 0,95, são indicadores de ajuste suficiente e valores acima de 0,95 são considerados bons ajustes.

Em relação ao resultado indicativo da qualidade de ajuste, apresentado pelo índice NNFI, a referência é que a estimativa seja inferior a 1 para que o modelo não seja caracterizado como “superajustado”, o que evidenciaria má especificação do modelo. Em se tratando do comportamento dos resíduos, o resultado dos índices SRMR e RMSEA foram próximos a zero, evidenciando boa aproximação da matriz de covariâncias imposta pelo modelo em relação à matriz de covariâncias amostral.

Finalizando a validação do modelo proposto, por meio do qui-quadrado ponderado, definido por Hair et al. (2005) como a razão entre a estatística qui-quadrado e o número de graus de liberdade, o valor obtido para o modelo ajustado foi 1,386. Os autores alertam que valores inferiores a 1 indicam ajuste “pobre” e valores entre 1 e 5 indicam que o modelo é aceitável.

Dada a concordância na validação do ajuste do modelo de equações estruturais sugerido, julga-se que há evidências estatísticas para afirmar que o modelo está bem ajustado e possível de ser considerado na interpretação das relações de causa e efeito postuladas na construção do modelo.

3.3 Interpretação dos parâmetros do modelo

Confirmada a validação, a interpretação do modelo, bem como as relações de causas e efeitos, foi discutida sob dois aspectos: Relação causas-efeitos entre as variáveis latentes (Seção 3.3.1) e relação das variáveis observadas com cada construto (Seção 3.3.2).

As estimativas dos parâmetros e dos respectivos erros padrões e valores-p do modelo ajustado são apresentadas na Tabela 8.

Há evidências estatísticas para afirmar que as estimativas que resultaram em valor- p superior ao nível de significância pré-fixado em 0,05 indicam que as relações lineares que especificam as causas e/ou efeitos são não significativas.

Contudo, convém ressaltar que por questões numéricas envolvendo identificabilidade do modelo e iterações, essas variáveis foram mantidas no modelo, confirmado pelos índices de qualidade de ajuste (Tabela 7).

Tabela 8: Estimativas dos parâmetros do modelo de equações estruturais ajustado para as regiões 2 e 4

Parâmetro	Estimativa	Erro padrão	valor-p
λ_{3ht}	6,459	0,3051	0,0000
λ_{3hp}	9,014	0,4187	0,0000
θ_{tpH}	-68,518	6,4029	0,0000
λ_{h2}	3,421	0,1714	0,0000
λ_{h3}	4,598	0,2223	0,0000
θ_{23h}	0,018	0,0945	0,8471
λ_{1i5}	3,752	0,1861	0,0000
λ_{1i2}	9,813	0,4557	0,0000
θ_{25N}	-41,613	3,9839	0,0000
λ_{2i3}	8,725	0,4064	0,0000
λ_{2i2}	15,206	0,7016	0,0000
λ_{2i4}	9,674	0,4510	0,0000
θ_{23I}	-68,977	6,4856	0,0000
θ_{24I}	-111,490	10,3679	0,0000
θ_{34I}	-60,095	5,6617	0,0000
β_{21}	0,072	0,0330	0,0315
β_{31}	0,004	0,0285	0,9007
β_{32}	-0,026	0,0316	0,4205
γ_1	-0,066	0,0448	0,1422
γ_2	-0,023	0,0494	0,6373
γ_3	0,267	0,0460	0,0000

3.3.1 Relação da variável latente (ξ) com as variáveis endógenas (η_3), (η_2) e (η_1)

Segundo o ZEE-MG, as políticas geradoras de investimento na área social são consideradas importantes indutoras do desenvolvimento econômico e social, ambientalmente sustentáveis, preservando os direitos econômicos e sociais, constitucionalmente garantidos. Nesta perspectiva, o modelo proposto neste artigo envolveu questões sociais relacionadas aos setores de habitação, agrário, segurança e ocupação econômica da população urbana, respectivamente representado por meio dos construtos (ξ), (η_1), (η_2) e (η_3).

Dada a amplitude e complexidade da relação funcional entre esses setores, o modelo estrutural (Figura 7) sugere que o setor de habitação seja uma das fontes causadora da melhoria da qualidade de vida do indivíduo, uma vez que as variáveis latentes endógenas são postuladas como indicadores reflexivos de ξ .

Scolforo et al. (2008) afirmaram que a moradia, na sociedade moderna, deverá obedecer a requisitos mínimos de segurança, conforto, bem-estar e higiene. No contexto urbano, a moradia deve oferecer ao indivíduo acesso a benfeitorias como urbanização, abastecimento de água e energia, esgotamento sanitário, coleta de lixo, drenagem das águas pluviais, escolas e postos médicos, transporte coletivo, lazer e segurança. No meio rural, a casa deve possibilitar a seus moradores o acesso aos requisitos essenciais de habitação, dentro de adequadas condições ambientais. Estatisticamente, essa relação não é confirmada pela probabilidade de significância observada nas estimativas do modelo estrutural ajustado (Tabela 8) em relação aos parâmetros γ_1 e γ_2 , sugerindo que o construto "Questões Habitacionais" (ξ) não apresente uma contribuição expressiva na dependência com os construtos endógenos η_1 e η_2 . Porém, ao analisar a probabilidade de significância obtida em relação ao parâmetro γ_3 , observa-se a confirmação da relação das questões habitacionais com a ocupação econômica e cobertura da população urbana.

Esta relação pode ser entendida, em uma realidade mais ampla envolvendo aspectos demográficos, na orientação da atividade produtiva para o setor industrial e de serviços, como fatores determinantes das alternativas de localização das atividades econômicas. Neste sentido, Oliveira et al. (2008) caracterizaram a região 4 (Central) como bastante heterogênea em relação ao potencial humano, porém por abranger uma

região metropolitana de grande porte com os consequentes problemas de saúde, renda, moradia e segurança, exige maior atenção no planejamento de ações e investimentos.

Em se tratando da região 2 (Sul/Sudoeste) há um número expressivo de municípios de pequeno porte populacional, no entanto, a maioria desses municípios está sob a influência de diversos municípios pólos que detêm condições sociais expressivas nas áreas de saúde, educação, renda e segurança.

Relacionando estas situações distintas observadas nestas regiões, verifica-se que a urbanização pode ser vista como resultado das mudanças nas formas históricas de apropriação do espaço e de orientação da atividade produtiva para os setores industrial e de serviços. Além do mais, a urbanização atua como fator relevante na determinação da localização das atividades econômicas e portanto, neste contexto, é razoável supor a dependência das questões habitacionais com a ocupação econômica e cobertura da população urbana.

Em se tratando da relação dos construtos endógenos, denominados por "Questões Agrárias" (η_1) e "Questões de Segurança" (η_2), respectivamente dos componentes Natural e Institucional, observou-se que a existência desta relação é notória dada a presença de diferentes órgãos públicos gerenciados pelo Ministério do Desenvolvimento Agrário¹, dos quais podemos mencionar as Promotorias Agrárias, atuando na emissão de pareceres nos processos referentes à disputa pela terra rural submetidos à apreciação do Poder Judiciário; Defensorias Públicas Agrárias, que atuam a favor dos trabalhadores rurais que não possuem condições financeiras de arcar com honorários de advogado e necessitam de orientação profissional na defesa de seus direitos, nas questões pertinentes a conflitos agrários ocorridos no campo; Ouvidorias Agrárias Regionais, exercendo sua função em parceria com a Ouvidoria Agrária Nacional, atuando na mediação e resolução de conflitos agrários nos estados; Delegacias de Polícia Civil Agrárias, que prestam atendimento referente às ocorrências registradas na zona rural, principalmente no que diz respeito aos conflitos coletivos pela posse da terra rural; Polícias Militares Agrárias, prestando segurança pública na zona rural, principalmente nas fazendas que se encontram ocupadas por trabalhadores rurais sem terras, que demandam providências dos poderes públicos para implantação

¹ Ministério do Desenvolvimento Agrário - Ouvidoria Agrária Nacional
<http://www.mda.gov.br/portal/ouvidoria/institucional/FuncaoDOSorgaosagrarios>, Acesso: 07/04/2011.

do programa de reforma agrária, assim como durante o cumprimento de mandados judiciais de reintegração de posse coletiva, buscando a resolução dos conflitos agrários de forma pacífica.

Mediante a diversidade de órgãos públicos de segurança direcionados a questões agrárias, torna-se plausível hipotetizar a dependência entre os construtos η_1 e η_2 , sendo esta comprovada pelo resultado significativo da estimativa do parâmetro β_{21} . Reportando à definição do fator condicionante imposta no ZEE-MG, entende-se que uma mudança (causa) em uma das variáveis utilizadas na formação de um desses construtos, certamente ocasionará (efeito) uma mudança na realidade de um determinado município pertencente às Regiões Sul/Sudoeste e Central. Desta forma, no que tange à administração pública, esta pode estar preparada e estruturada para acompanhar as demandas por segurança.

Em se tratando das relações dos construtos η_1 , η_2 e η_3 , ou seja, "Questões Agrárias" e "Questões de Segurança" em relação às "Questões de Ocupação Econômica e Cobertura da População Urbana", por meio das probabilidades de significância observadas em relação aos parâmetros β_{31} e β_{32} (Tabela 8), os resultados foram não significativos.

Uma suposta explicação para a ocorrência desse resultado é que na região 2 (Sul/Sudoeste) há um número expressivo de municípios de pequeno porte populacional, os quais estão sob a influência de diversos municípios pólos que detêm condições sociais expressivas nas áreas de saúde, educação, renda e segurança.

No que se refere aos aspectos demográficos, os municípios, principalmente os de pequeno contingente populacional, apresentam uma alta razão de dependência, o que pode limitar ações e investimentos.

No caso da região 4 (Central) verifica-se uma heterogeneidade que se refere ao potencial humano. O fato de abranger uma região metropolitana de grande porte com os consequentes problemas de saúde/renda/moradia/segurança indica restrições e cuidados ao se projetar ou planejar ações e investimentos. Por outro lado, no que se refere ao potencial humano, a região apresenta, em termos de educação e renda, uma potencialidade diferenciada que pode atrair investimentos e ações governamentais.

3.3.2 Relação das variáveis observadas com cada construto

A análise do construto "Questões Agrárias" (η_1) se dá no sentido de que a importância do componente Natural, no qual este fator está inserido, não se relaciona simplesmente aos aspectos relativos à qualidade e vulnerabilidade dos recursos naturais, mas sim aos aspectos econômicos dos mesmos. Portanto, busca-se analisar o quanto, economicamente, o recurso natural está sendo utilizado como fonte geradora e impulsionadora da economia. Além disso, as variáveis que formam esse construto retratam não apenas a estrutura fundiária, mas também o nível tecnológico empregado na agricultura e na produção familiar, importante produtora e fornecedora de alimentos básicos às populações locais e geradoras de ocupações e renda, no contexto da dinâmica agrária dos municípios estudados.

Assim sendo, os resultados encontrados na Tabela 8 confirmam de fato uma relação linear do construto η_1 com as demais variáveis observadas, uma vez que as mesmas apresentaram probabilidades de significância inferior ao nível nominal de significância definido em 0,05.

O construto "Questões de Segurança" (η_2) está envolvido com um fator condicionante do componente Institucional, portanto, as variáveis observadas na formação desse construto estão relacionadas a questões relativas a propriedades e patrimônio do bem público, gestão municipal e organizações de ensino e pesquisa.

As variáveis utilizadas na formação do construto "Questões de Ocupação Econômica e Cobertura da População Urbana" (η_3) foram significativas, portanto, pode-se afirmar que as variáveis consideradas apresentaram uma importante contribuição na formação desse construto.

O construto "Questões Habitacionais" (ξ) é formado por variáveis que expressam o desenvolvimento sustentável relacionado à ocupação populacional e condições básicas sanitárias.

Configurando o Zoneamento Ecológico Econômico do Estado de Minas Gerais, essas variáveis são classificadas como indicadores, portanto, é coerente considerar que o construto também seja considerado um indicador e, contextualizados no ZEE-MG, os indicadores são interpretados como causas e os fatores condicionantes como efeito. Sendo assim, as interpretações das estimativas dos parâmetros das variáveis utilizadas na formação dos construtos foram apenas confirmatórias, sendo sua contribuição relevante e justificada pelos valores significativos observados nas estimativas dos parâmetros.

4. Conclusões

As técnicas estatísticas empregadas neste artigo foram eficientes no estudo simultâneo dos componentes Humano, Natural e Institucional com a obtenção de resultados específicos sobre a formação de agrupamentos das regiões avaliadas. No estudo da relação de causas-efeitos apresentaram resultados coerentes com a perspectiva do Zoneamento Econômico Ecológico do Estado de Minas Gerais.

Especificamente, a aplicação do escalonamento multidimensional demonstrou uma diversidade socioeconômica e ecológica entre as regiões estudadas, de tal forma que as regiões Sul/Sudoeste e Central apresentaram resultados similares, formando um agrupamento. Analogamente, as regiões do Rio Doce e Zona da Mata apresentaram similaridades entre as variáveis estudadas.

A aplicação da modelagem de equações estruturais permitiu confirmar as hipóteses do modelo teórico, fundamentadas em um dos objetivos do Zoneamento Ecológico Econômico do Estado de Minas Gerais, consistindo justamente, em fornecer às regiões e municípios diagnósticos gerais e uma perspectiva global sobre a realidade do estado, bem como agregando conhecimento segundo as peculiaridades regionais.

Referências Bibliográficas

- AGENDA 21 (2002). Agenda 21 Brasileira: resultado da consulta nacional. (Org.) Maria do Carmo de Lima Bezerra, Márcia Maria Facchina e Otto Ribas. Brasília, Ministério do Meio Ambiente, Programa das Nações Unidas para o Desenvolvimento.
- BORG, I; GROENEN, P. (2005). *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer Series in Statistics.
- DE LEEUW, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, v.5, p.163-180.
- GOWER, J.C.; HAND, D. J. (1996). *Biplots*. London: Chapman & Hall, 277p.
- HADDAD, P.R. (2004). Força e fraqueza dos municípios de Minas Gerais. Cadernos BDMG, Nº 8. Belo Horizonte.
- HAIR JR, J.F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. (2005). *Análise multivariada de dados* (5a. ed.). Porto Alegre: Bookman.
- HERSHBERGER, S.L., MARCOULIDES, G.A.; PARRAMORE, M.M. (2003) Structural equation modeling: An introduction. In B.H. Pugesek, A. Tomer & A.V. Eye (Eds.), *Structural equation modeling: Applications in ecological and evolutionary biology*. Cambridge: Cambridge University Press.
- HOX, J. J.; BECHGER, T. M. (1998). An Introduction to Structural Equation Modeling. *Family Science Review*, 11, pp. 354-373.
- IBGE – Instituto Brasileiro de Geografia e Estatística (2005). *Pesquisa de Informações Básicas Municipais: Gestão Pública*. Rio de Janeiro.
- JAWORSKA, N.; ANASTASOVA, A.C. (2009). A Review of Multidimensional Scaling (MDS) and its Utility in Various Psychological Domains. *Tutorials in Quantitative Methods for Psychology*, .5, pp. 1-10.
- KAPLAN, D. (2000). *Structural Equation Modeling: Foundations and Extensions*. Sage: Thousand Oaks.
- KRUSKAL, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29, pp.115-129.
- MMA e SAE (1997). *Detalhamento da Metodologia para Execução do Zoneamento Ecológico-Econômico pelos Estados da Amazônia Legal*. Brasília: Ministério do Meio Ambiente, dos Recursos Hídricos e da Amazônia Legal. MMA, Secretaria de Coordenação da Amazônia. SCA, Secretaria de Assuntos Estratégicos da Presidência da República. SAE/PR, Responsáveis Técnicos Dra. Bertha K. Becker e Dr. Cláudio A. G. Egler, Laboratório de Gestão do Território da Universidade Federal do Rio de Janeiro.
- OLIVEIRA, L.C.F.S.; LEITE, E.T.; RIBEIRO, L.M.P.; RESENDE, J.B.; A.M.P (2008). Componente Humano. In: *Zoneamento Ecológico-Econômico do estado de Minas Gerais – Componentes Socioeconômico*. Editora UFLA. Lavras, p.77-100.
- PEREIRA, J.R. (2008). Construção dos indicadores de potencialidade social. In: *Zoneamento Ecológico-Econômico do estado de Minas Gerais – Componentes Socioeconômico*. Editora UFLA. Lavras, p.7-34.
- PILATI, R.; LAROS, J.A. (2007). Modelos de equações estruturais em psicologia: conceitos e aplicações. *Psicologia Teoria e Pesquisa*, 23, pp.205-216.
- SCOLFORO, J. R. S.; OLIVEIRA, A. D.; CARVALHO, L. M. T. (2008). *Zoneamento Ecológico-Econômico do estado de Minas Gerais – Componentes Socioeconômico*. Lavras: Editora UFLA.

- SHADISH, W. R., COOK, T. D.; CAMPBELL, D. T. (2002). Experimental and quasi-experimental design for generalized causal inference. Boston: Houghton Mifflin Company.
- SHEPARD, R. N. (1962). The analysis of proximities: multidimensional scaling with unknown distance function. *Psychometrika*, v.27, p.125-149.
- ULLMAN, J. B. (2007). Structural Equation Modeling. Em B. G. Tabachnick & L. S. Fidell (Orgs.), *Using multivariate statistics* (5ª ed.). Boston: Pearson Education.
- VICKERMAN, R. (1995). Location, accessibility and regional development: the appraisal of trans-European networks. *Butterworth Heinemann. Transport Policy*, v. 2, n.4, p. 225-234.

Agradecimentos

Ao apoio financeiro na realização do ZEE-MG em parceria com a UFLA - Universidade Federal de Lavras. A CAPES e IME-USP pelo financiamento da bolsa de Pós-doc como produtividade do projeto PROCAD. À FAPESP pelo Projeto Temático 2012/21788-2.

Abstract

The Ecological and Economic Zoning of the State of Minas Gerais (ZEE-MG) is a set of rules and techniques to diagnose geo-biophysical, socio-economic, legal, and institutional aspects, generating the Letter of Environmental Vulnerability and the Charter of Social Potential. In particular, the Charter of Social Potential is measured by Productive, Natural, Human, and Institutional components, which characterizes a municipality or micro-region ability to achieve sustainable development. Thus, we study the variables belonging to these components, that can be suggested in the construction of a structural model to explain plausible cause and effect relationship, using multidimensional scaling techniques and structural equation modeling. The trial follows methodological guidelines by the Brazilian Department of Environment-MMA for the ZEE-MG development, in accordance with guidelines by the Environmental Law and Policy of the State of Minas Gerais.

Keywords: structural equation, environment, social potential.

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística Aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação, citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará, também, artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográfico e econométrico.

Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores seja brasileiro e nem resida no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo 'duplo cego', isto é, os artigos são avaliados sem a identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBES

O processo editorial da RBES é eletrônico. Os artigos devem ser submetidos para o site <http://rbes.submitcentral.com.br/login.php>

Secretaria da RBES

Revista Brasileira de Estatística – RBES

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - IBGE

Rua André Cavalcanti, 106, sala 503-A

Centro, Rio de Janeiro – RJ

CEP: 20031-050

Tels.: 55 21 2142-3596 (Marilene Pereira Piau Câmara – Secretária)

55 21 2142-4957 (Pedro Luis do Nascimento Silva – Editor-Executivo)

Fax: 55 21 2142-0501

INSTRUÇÕES PARA PREPARO DOS ORIGINAIS

Os originais enviados para publicação devem obedecer às normas seguintes:

1. Podem ser submetidos originais processados pelo editor de texto *Word for Windows* ou originais processados em LaTeX (ou equivalente) desde que estes últimos sejam encaminhados e acompanhados de versões em pdf, conforme descrito no item 3, a seguir;
2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos, se for o caso de constarem no documento, também devem figurar nesta página;
3. No caso de a submissão não ser em *Word for Windows*, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, LaTeX). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página, como descrito no item 2, e outro contendo apenas o título, sem a identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;
4. A segunda página do original deve conter resumos em português e inglês (*abstract*), destacando os pontos relevantes do artigo. Cada resumo deve ser digitado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com, no máximo, 150 palavras;

5. O artigo deve ser dividido em seções, numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado;
6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;
7. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;
8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas sequencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções de trabalho;
9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos;
10. Não serão permitidos itens que identifiquem os autores do artigo dentro do texto, tais como: número de projetos de órgãos de fomento, endereço, *e-mail*, etc. Caso ocorra, a responsabilidade será inteiramente dos autores; e
11. No caso de o artigo ser aceito para a publicação após a avaliação dos pareceristas, serão encaminhadas as sugestões/comentários aos autores sem a sua identificação. Uma vez nesta condição, é de responsabilidade única dos autores fazer o *download* da formatação padrão da revista (em doc ou em LaTeX) para o envio da versão corrigida.

