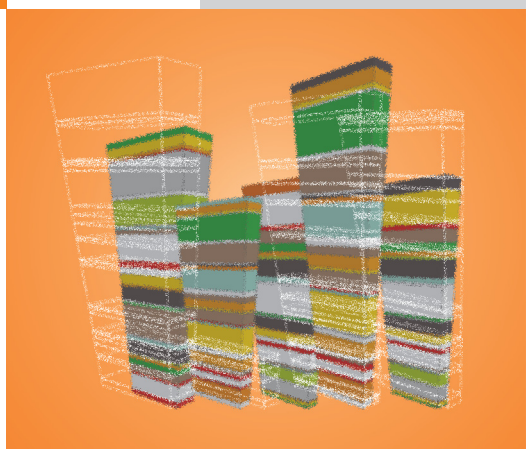
REVISTA BRASILEIRA DE ESTATÍSTICA

ISSN 0034-7175



volume 73
número 236
janeiro / junho 2012

IBGE
Instituto Brasileiro de Geografia e Estatística

Presidenta da República
Dilma Rousseff

Ministra do Planejamento, Orçamento e Gestão
Miriam Belchior

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidenta
Wasmália Bivar

Diretor-Executivo
Fernando J. Abrantes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Marcia Maria Melo Quintslr

Diretoria de Geociências
Wadih João Scandar Neto

Diretoria de Informática
Paulo César Moraes Simões

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Denise Britz do Nascimento Silva

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

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 73 número 236 janeiro/junho 2012

ISSN 0034-7175

R. Bras. Estat., Rio de Janeiro, v. 73, n. 236, p. 1-117, jan./jun. 2012

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

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.

Editoração
Marilene Pereira Piau Câmara - 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

Foi com enorme satisfação que assumi a posição de Editora Responsável da Revista Brasileira de Estatística a partir de 2012. Agradeço ao Instituto Brasileiro de Geografia e Estatística (IBGE) e à Associação Brasileira de Estatística (ABE) pelo convite e confiança.

O presente volume da RBEs reúne uma lista variada de contribuições. O artigo de Julio da Motta Singer, Carmen Diva Saldiva de André e Clóvis de Araújo Peres propõe um planejamento que minimiza a variância do estimador do coeficiente de correlação intraclasse. No artigo de Heitor Pinto de Moura Filho é proposta uma tipologia das tábuas de mortalidade enquanto que o artigo de Taiane Schaedler Prass e Silvia Regina Costa Lopes analisa o desempenho dos métodos do teste de estresse, da perda máxima e do valor de risco na presença de heteroscedasticidade e longa dependência na volatilidade. Por último, o artigo de Gustavo Silva Semaan, José André de Moura Brito e Luiz Satoru Ochi propõe um método de agrupamento hierárquico baseado no algoritmo Bisecting K-Means, com o objetivo de identificar a quantidade ideal de grupos.

A RBEs conta agora com dois Editores Executivos, Pedro Luis do Nascimento Silva (ENCE/IBGE), que continua de gestões anteriores, e Mário de Castro Andrade Filho (ICMC-USP). O Editor de Estatísticas Oficiais passou a ser José André de Moura Brito (ENCE/IBGE) e o Editor de Metodologias, Fernando Antonio da Silva Moura (UFRJ) permanece o mesmo. Alguns novos membros passaram a fazer parte do corpo de Editores Associados. Agradeço a todos por me acompanharem nesta tarefa e espero podermos, juntos, fazer um bom trabalho pela RBEs e pela Estatística.

Quero agradecer também aos autores que submeteram seus artigos, IBGE, ABE e a todos que atuaram como revisores, que anonimamente contribuíram para mais este volume da Revista Brasileira de Estatística. Estendo meus agradecimentos à equipe de editoração do periódico.

Saudações cordiais

Lúcia Pereira Barroso
Editora Responsável

Confiabilidade e Precisão na Estimação de Médias

Julio M. Singer [1]
Carmen Diva Saldiva de André [2]
Clóvis de Araújo Peres [3]

Resumo

Num estudo planejado para estimar a concentração média de dióxido de nitrogênio ao longo de um corredor de ônibus, *m* filtros passivos (réplicas) são colocados em *n* diferentes locais (repetições). A análise de problemas com essa estrutura pode ser realizada por intermédio de um modelo misto e a amplitude do intervalo de confiança construído nesse contexto depende da relação entre o número de repetições e réplicas além do coeficiente de correlação intraclasse que, em geral precisa ser estimado por meio de um estudo piloto. Mostramos que, para coeficientes de correlação intraclasse maiores que 0,2 e número de observações (*N = n x m*) da ordem de 20 ou mais, a escolha de quatro réplicas corresponde aproximadamente ao planejamento que minimiza a variância do estimador. Utilizando os dados do estudo supramencionado, mostramos como esse resultado pode ser empregado no planejamento de um experimento cujo objetivo é estimar a concentração média de dióxido de nitrogênio com uma precisão pré-fixada.

Palavras-chave: confiabilidade, coeficiente de correlação intraclasse, modelos com fatores aleatórios, planejamento ótimo, repetições, réplicas.

[1] *Departamento de Estatística - Universidade de São Paulo, Brasil. Correspondência: jmsinger@ime.usp.br*
[2] *Departamento de Estatística - Universidade de São Paulo, Brasil.*
[3] *UNIFESP e Departamento de Estatística – Universidade de São Paulo, Brasil*.

1. Introdução

O objetivo de um estudo realizado pelo Laboratório de Poluição Atmosférica Experimental da Universidade de São Paulo (LPAE) é estimar a concentração média de dióxido de nitrogênio *(NO_2)* ao longo de um corredor de ônibus. Em um experimento piloto, esse poluente foi monitorado em 5 locais dessa via; em cada local, foram colocados 4 filtros passivos que geram medidas da concentração de *NO_2* por meio de um processo químico.

Os dados estão dispostos na Tabela 1.

Tabela 1: Concentração de dióxido de nitrogênio ($\mu g/m^3$)

Local	Concentração	Local	Concentração
1	170,6	3	128,4
1	154,7	3	118,1
1	136,4	4	153,9
1	153,1	4	149,1
2	68,0	4	147,5
2	66,4	4	103,8
2	70,3	5	83,9
2	71,1	5	101,4
3	151,5	5	117,3
3	138,0	5	114,1

A análise de problemas em que várias (m) réplicas de uma mesma característica são obtidas em n unidades amostrais (repetições) pode ser concretizada por meio de um modelo de análise de variância com um fator aleatório da forma

$$y_{ij} = \mu + a_i + e_{ij} \qquad (1)$$

em que ai $a_i \sim N(0, \sigma_a^2)$ e $e_{ij} \sim N(0, \sigma_e^2)$ são independentes, i = 1, ..., n e j = 1, ..., m. Para detalhes, o leitor pode consultar Neter et al. (2005, cap 25), entre outros.

No estudo descrito acima, as repetições correspondem aos locais de coleta $(n = 5)$ e as réplicas ($m = 4$) às medidas realizadas no diferentes filtros. O coeficiente de correlação intraclasse sob o modelo (1), dado por

$$\rho = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2) \qquad (2)$$

é uma medida de confiabilidade de uma observação da resposta conforme indicam Bartko (1966), Fleiss (1986) ou McGraw and Wong (1996). Fleiss (1986) também utiliza *coeficiente de correlação intraclasse de confiabilidade* ou simplesmente *confiabilidade* para se referir a ρ.

Quando a variância das observações intra-unidades amostrais, σ_e^2, é pequena relativamente à variância das observações inter-unidades amostrais, σ_e^2 , o valor de ρ se aproxima de 1; por outro lado, o valor de ρ diminui, quando σ_e^2 aumenta com σ_e^2 fixo. Portanto, quando as observações intraunidades amostrais são mais homogêneas, o valor de ρ se aproxima de 1. Segundo Fleiss (1986), em ensaios clínicos, valores de ρ maiores que 0,75 indicam alta confiabilidade e valores menores que 0,4, confiabilidade baixa. A confiabilidade da resposta de um estudo pode ser aumentada, se cada um de seus valores corresponder à média de m réplicas. Conforme comenta Fleiss (1986), assim como a média de $m > 1$ observações constitui um estimador mais preciso da média populacional do que uma única observação, pode-se dizer que a média de m réplicas é uma medida mais confiável do que uma única delas. Nesse contexto, sob o modelo

$$\bar{y}_i = \mu + a_i + \bar{e}_i \qquad (3)$$

Em que $\bar{y}_i = m^{-1}\Sigma_{j=1}^{m} y_{ij}$ e $\bar{e}_i = m^{-1}\Sigma_{j=1}^{m} e_{ij}$, o coeficiente de correlação intraclasse, que corresponde à confiabilidade da média das m réplicas, é dado por

$$\rho_m = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2 / m) \qquad (4)$$

Como

$$\rho_m = (m\rho)/[1+(m-1)\rho] \qquad (5)$$

A confiabilidade da média de m réplicas depende do número m de réplicas e de ρ e aumenta com o número de réplicas. De (5) conclui-se que o valor de m necessário para se atingir uma confiabilidade ρ_m fixada previamente, é

$$m = \rho_m(1-\rho)/[\rho(1-\rho_m)] \qquad (6)$$

Estudos nos quais as médias de observações obtidas em duplicata, triplicata, ou, mais geralmente, em m-plicata, constituem os valores da variável resposta são bastante freqüentes e exemplos podem ser encontrados em Singer et al. (2007) entre outros. No experimento piloto realizado pelo LPAE, as observações foram realizadas em quadruplicatas e um de seus objetivos era avaliar que número (m) de filtros deveria ser colocado em cada um dos n locais de coleta para se estimar a concentração média de NO_2 com uma confiabilidade fixada previamente ou equivalentemente, para obter um intervalo de confiança com uma amplitude pré-fixada num estudo subsequente.

Singer et al. (2007) utilizaram o modelo (1) para avaliar a influência do número m de réplicas na amplitude do intervalo de confiança para a média μ da variável resposta. Esses autores mostraram que a diminuição dessa amplitude depende não só do valor de m, mas também do coeficiente de correlação intraclasse ρ. Como, em geral, o valor desse parâmetro é desconhecido, ele deve ser estimado a partir de algum estudo piloto.

O objetivo deste trabalho é propor um planejamento com a finalidade de estimar ρ. Na seção 2, resumimos os resultados obtidos por Singer et al. (2007) e expressamos a redução no comprimento do intervalo de confiança para a média em termos do acréscimo da confiabilidade da resposta. Determinamos o planejamento que minimiza a variância de um estimador de ρ na Seção 3. Os resultados apresentados nas Seções 2 e 3 são aplicados ao estudo realizado pelo LPAE na Seção 4. Conclusões e considerações finais são apresentadas na Seção 5.

2. Precisão versus confiabilidade na estimação da média

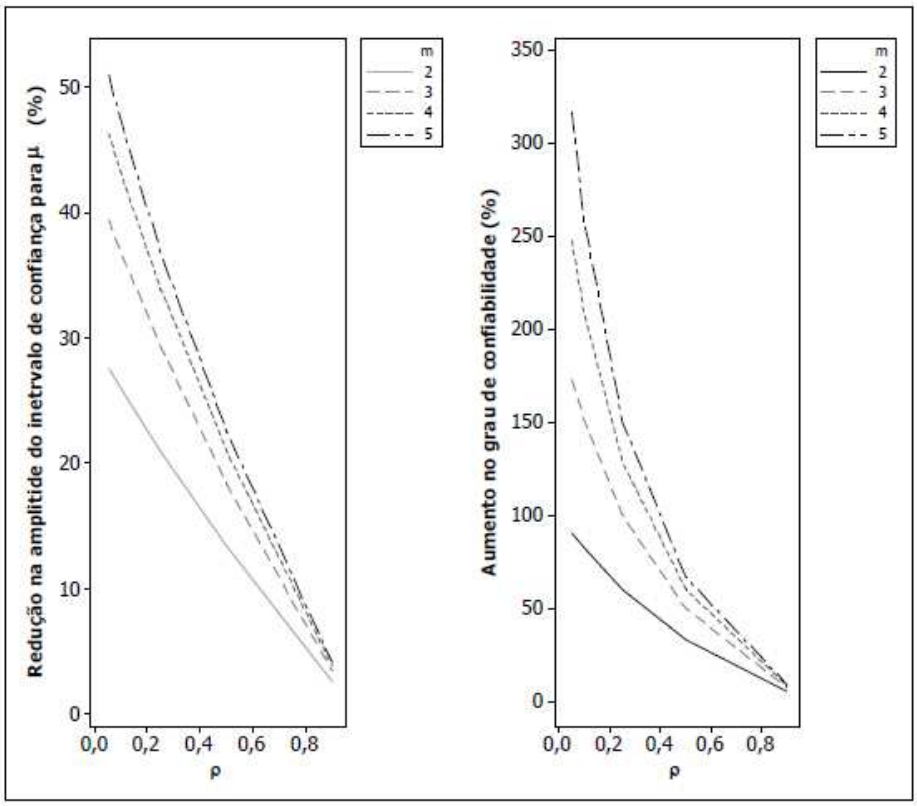
Singer et al. (2007) mostraram que, sob o modelo (1), a redução da amplitude do intervalo de confiança para média μ, baseada em n medidas obtidas em m-plicata ($m>1$) relativamente à amplitude do intervalo de confiança baseado em $m = 1$ é de

$$1-\sqrt{\rho+(1-\rho)/m}\times 100\% \qquad (7)$$

Como (7) é uma função decrescente de ρ, a redução será tanto menor, quanto maior for ρ para um valor de m fixado. Por outro lado, para um mesmo valor de ρ, a redução será maior quando m aumenta. De (5) conclui-se que a confiabilidade da média de m observações é $m/[1 + (m - 1)\ \rho]$ vezes a confiabilidade de uma única observação. Para um valor de m fixado esse fator é uma função decrescente de ρ, e para ρ fixado, a confiabilidade é uma função crescente de m.

O aumento da confiabilidade da resposta corresponde a uma redução da amplitude do intervalo de confiança para a média, ou seja, um aumento na precisão da estimativa da média. Para ilustrar este resultado, apresentamos na Figura 1, reduções percentuais na amplitude do intervalo de confiança e correspondentes aumentos da confiabilidade para diferentes valores de ρ, e de m.

Figura 1: Redução na amplitude do intervalo de confiança para μ e aumento no grau de confiabilidade da média de m-plicatas relativamente àquele obtido com uma única medida (m = 1), para diferentes valores de ρ e m



3. Alocação ótima para estimação de ρ

Seja $N = m$ x n o número total de observações, fixado a partir de considerações (custo, por exemplo) sobre o fenômeno sob investigação. A questão que vamos responder é: quais devem ser os valores de m e n para que a variância do estimados de ρ seja mínima? Consideramos o seguinte estimador baseado na análise de variância para o modelo (1):

$$\hat{\rho} = \frac{QMA - QMR}{QMA + (m-1)QMR} \qquad (8)$$

em que QMA = (n – 1)$^{-1}$ $\sum_{i=1}^{n} m(\bar{\gamma}_{i\bullet} - \bar{\gamma}_{\bullet\bullet})^2$, e $QMR = (N-n)^{-1} \sum_{j=1}^{n} (\gamma_{ij-} \bar{\gamma}_{i\bullet})^2$, com $\bar{y}_{i.\bullet} = m^{-1}\Sigma_{j=1}^{m} y_{ij}$ e $\bar{y}_{\bullet\bullet} = N^{-1}\Sigma_{i=1}^{n}\Sigma_{j=1}^{m} y_{ij}$.. Donner and Koval (1980) mostram que a variância de $\hat{\rho}$ é

$$\mathrm{V}\,\hat{\rho} = \frac{2(1-\rho)^2\left[n+(N-n)\rho\right]^2}{N(N-n)(n-1)} \qquad (9)$$

Para ρ fixo, $\mathrm{V}(\hat{\rho})$ tende ao infinito quando n tende a 1 ou a N. Podemos então afirmar que, dado ρ, existe pelo menos um valor de n inteiro no intervalo (1, N) que minimiza (9). Ching (1995) mostra que esse valor é dado por

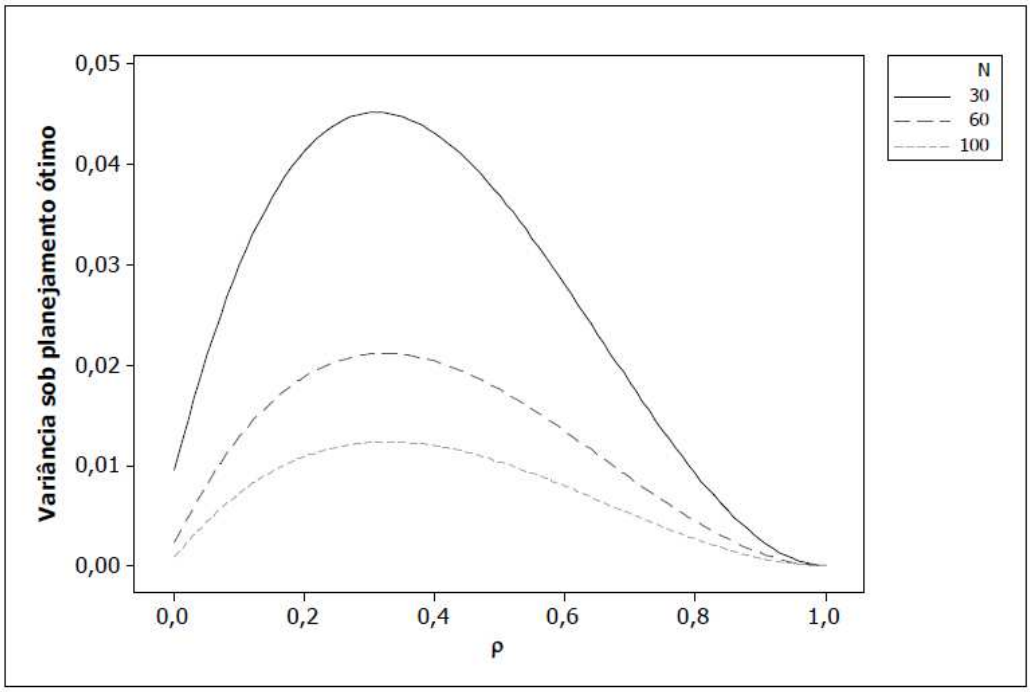
$$\mathrm{n_o} = \frac{N(N-1)\rho+2}{\left[(N+1)+(N-1)\rho\right]} \qquad (10)$$

Os valores ótimos de n e de m, dependem do parâmetro ρ que queremos estimar. Vamos mostrar que é possível obter um planejamento próximo do ótimo que independe de ρ em parte do seu campo de variação. Em outras palavras, vamos determinar n tal que $\mathrm{V}(\hat{\rho})$ obtida com esse valor seja próxima daquela obtida quando $n = n_o$, independa de ρ. Substituindo n por n_o em (9), a variância de $\hat{\rho}$ sob o planejamento ótimo é

$$\mathrm{V}(\hat{\rho}\,|\,n = n_o) = \frac{8(1-\rho)^2\left[(N-1)\rho+1\right]}{(N-1)^2} \qquad (11)$$

Na figura 2, representamos a função dada em (1), para diferentes valores de N.

Figura 2: Variância do estimador $\hat{\rho}$ sob o planejamento ótimo para N = 30, 60 e 100



Dado N, (11) atinge seu valor máximo quando

$$\rho = \frac{N-3}{3(N-1)} \qquad (12)$$

e utilizando (10), concluímos que o correspondente valor ótimo é $n_o = (N+3)/4$. Para valores de N próximos daqueles usualmente empregados na prática (e.g., N = 40), o valor n_o pode ser aproximado por $N/4$. Conseqüentemente, quando $\rho = (N\text{-}3)/[3(N-1)]$, o planejamento com $n = N/4$ e $m = 4$ está próximo do ótimo. Para encontrar um planejamento próximo ao ótimo para outros valores de $0 < \rho < 1$, notemos primeiramente que, quando $n = N/4$ e $m = 4$, podemos concluir de (9), que

$$V(\hat{\rho} \setminus n = N/4) = \frac{2(1-\rho)^2(1+3\rho)^2}{3(N-4)} \qquad (13)$$

Na Figura 3, representamos a expressões (11) e (13) como função de ρ, para $N=60$ e $N=100$. Essa figura sugere que, para $\rho > 0{,}2$, a variância de $\hat{\rho}$ obtida sob o planejamento ótimo é próxima daquela obtida com $n=N/4$. Para fazermos uma comparação analítica das variâncias obtidas sob os dois planejamentos, consideremos a função

$$Rel(\rho) = \frac{V(\hat{\rho} \setminus n = N/4)}{V(\hat{\rho} \setminus n = n_o)}$$

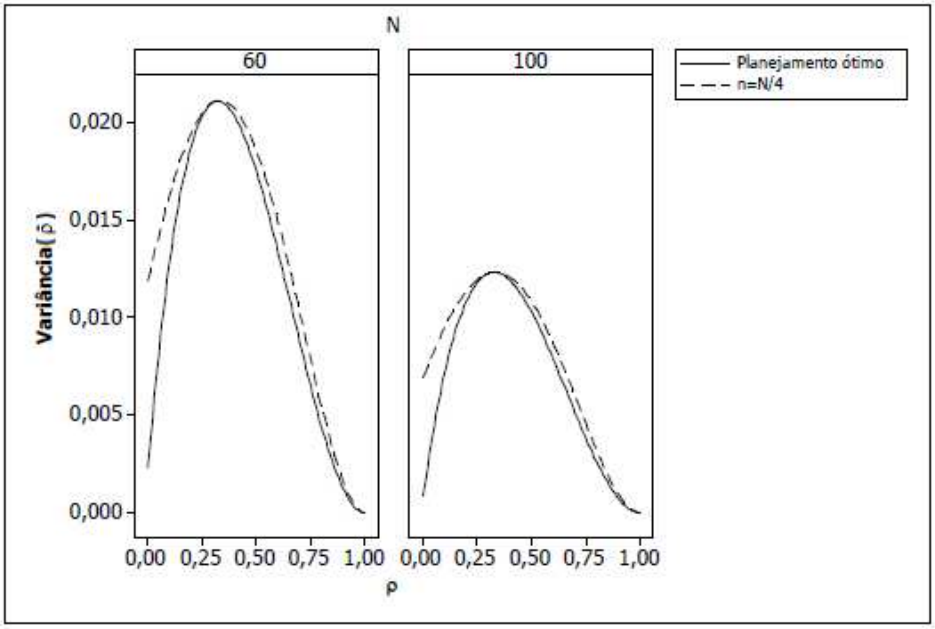
Que tem as seguintes propriedades:

a) $Rel(0) = \dfrac{(N-1)^2}{12(N-4)} \to \infty$ com $N \to \infty$

a) $Rel(1) = \dfrac{4(N-1)^2}{3N(N-4)} \to 1{,}33$ com $N \to \infty$

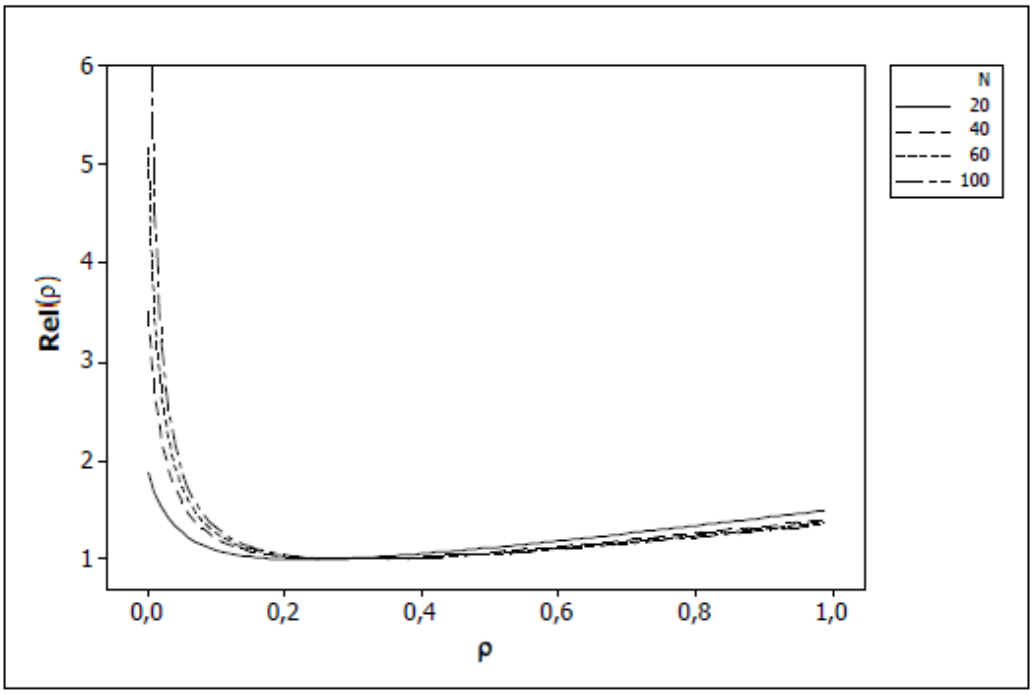
c) Para $N \geq 7$, $\min Rel(\rho) = 1$ e para $N < 7$, , $\min Rel(\rho) < 1$

Figura 3: Variância de $\hat{\rho}$ sob o planejamento ótimo e sob o planejamento com $n = N/4$ para $N = 60$ e 100



A função $Rel(\rho)$ está representada na Figura 4 para $N = 20$, 40, 60 e 100. Observamos, por exemplo, que, quando $N = 20$, $Rel(\rho) < 1{,}2$ para valores de ρ entre 0,07 e 0,62.

Figura 4 : Função *Rel* (ρ) para *N* = 20, 40, 60 e 100



4. Estudo da confiabilidade das medidas de Dióxido de Nitrogênio

No Estudo descrito no início deste trabalho foi adotado o planejamento quase ótimo para estimar ρ, sendo o número total de observações $N = 20$. Na Tabela 2 temos os valores de estatísticas descritivas para as concentrações de NO_2 nos 5 locais de coleta.

Tabela 2: Estatísticas descritivas para a concentração de dióxido de nitrogênio ($\mu g/m^3$) nos cinco locais de coleta

Local	Média	Desvio padrão	Mínimo	Mediana	Máximo
1	153,7	14,0	136,4	153,9	170,6
2	69,0	2,2	66,4	69,2	71,1
3	134,0	14,2	118,1	133,2	151,5
4	104,2	15,2	83,9	107,8	117,3

Da análise de variância obtém-se $QMA = 4533{,}2$ e $QMR = 235{,}4$, concluindo, por meio de (8), que a estimativa do coeficiente de correlação intraclasse é $\hat{\rho} = 0{,}82$. Substituindo este valor em (13), obtemos V ($\hat{\rho}$) = 0,016. Para que no estudo final a confiabilidade da resposta seja de 90% utilizamos (6) obtendo $m = 2$. Este resultado indica que os valores da resposta em cada local devem ser obtidos como médias das concentrações observadas em 2 filtros. Assim, se o número máximo de filtros disponíveis for 20, devem ser selecionados 10 locais de coleta, em cada qual 2 filtros devem ser colocados. A média das concentrações observadas nesses dois filtros será considerada como o valor observado da concentração de NO_2 nesse local. Quando este planejamento é adotado, a redução na amplitude do intervalo de confiança para μ é de 4,6,%.

Os dados da Tabela 2 indicam que tanto a média quanto a variância observadas são consideravelmente menores no local 2 do que nos demais. Em uma reanálise dos

dados realizada com a exclusão desses valores obtivemos $\hat{\rho} = 0{,}55$. Nesse caso, para que no estudo final a confiabilidade da resposta seja de 90%, deveríamos ter $m = 8$.

5. Considerações Finais

Utilizando resultados de Singer et al. (2007), concluímos que o aumento do grau de confiabilidade de uma resposta obtido a partir do aumento do número de réplicas pode também ser quantificado por meio do aumento da precisão da estimativa da média (expressa como amplitude do intervalo de confiança correspondente) e que o conhecimento do coeficiente de correlação intraclasse é indispensável para quantificar o ganho obtido. Para situações em que $\rho > 0{,}2$, o planejamento com $m = 4$ réplicas produz um estimador com variância próxima da ótima. Este resultado foi obtido considerando o estimador de análise de variância para ρ, que pode gerar estimativas negativas dos componentes de variância. Embora, na prática esse tipo de situação não seja comum (em geral, espera-se que a variabilidade entre unidades amostrais seja maior que aquela intraunidades amostrais), os resultados aqui obtidos se mantêm inalterados com a utilização do estimador de máxima verossimilhança desse parâmetro (ver Ching (1995), por exemplo, para detalhes).

Como no exemplo considerado, as análises sugerem que o número de réplicas esteja entre $m = 2$ e $m = 8$, o planejamento do estudo final poderia envolver os custos de obtenção das réplicas em um mesmo local e aqueles relacionados com a seleção dos locais de coleta. Resultados apresentados em Singer et al. (2007) podem ser utilizados para incorporar essa informação.

Referências bibliográficas

Bartko, J.J. (1966). The intra-class correlation coefficient as a measure of reliability. Psychological Reports, 19}, 3-11.

Ching, T.H. (1995). Coeficiente de correlação intraclasse: planejamento com alocação ótima e aplicação no estudo de confiabilidade de medidas. Dissertação de Mestrado. Departamento de Estatística, IME-USP

Donner, A. and Koval, J.J. (1980). The large sample variance of an intra-class correlation. Biometrika, 67, 719-722

Fleiss, J.L. (1986) The Design and Analysis of Clinical Experiments. New York: John Wiley and Sons.

McGraw, K.O. and Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1, 30-46.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Li, W. (2005). Applied Linear Statistical Models, 5th ed. Chicago: Irwin.

Singer, J.M., Lima, A.C.P., Tanaka, N.I. and González-López, V.A. (2007). To triplicate or not to triplicate? Chemometrics and Intelligent Laboratory Systems, 86, 82-85.

Agradecimento

Os autores agradecem o auxílio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) recebido durante a elaboração deste trabalho além da cuidadosa leitura e sugestões construtivas de um editor associado, que incorporadas no texto, melhoraram-no substancialmente.

Abstract

In a study designed to estimate the mean nitrogen dioxide concentration along a bus line, m of passive filters (replicates) are placed in n different spots (repetitions). The analysis of problems with this structure may be carried out via a mixed model and the range of a confidence interval constructed in this context depends on the relation between the number of repetitions and replicates as well as on the intraclass correlation coefficient, that, in general, must be estimated by means of a pilot study. We show that, for intraclass correlation coefficients larger than 0.2 and number of observations (N=n x m) of the order of 20 or more, the choice of four replicates corresponds approximately to the design that minimizes the variance of the estimator. Using data from the study mentioned above, we show how this result may be employed in the design of an experiment to estimate the mean nitrogen dioxide concentration with a given precision.

UMA TIPOLOGIA DE TÁBUAS DE MORTALIDADE

Heitor Pinto de Moura FIlho [1]

Resumo

Tábuas de mortalidade são instrumentos fundamentais para a representação de mecanismos demográficos. Propomos uma tipologia dessas tábuas a partir de seus critérios de construção. Fazemos uma primeira distinção entre tábuas "empíricas", que têm por referencial populações existentes ou que existiram no passado, e tábuas "matemáticas", modeladas sobre características empíricas, mas parametrizadas em função de variáveis demográficas, como o crescimento natural da população ou sua esperança de vida ao nascer. Estudos demográficos fazem uso crescente de combinações de ambos esses tipos, aproveitando as regularidades embutidas nas tábuas teóricas para suprir informações desconhecidas nos dados em estudo. As tábuas matemáticas são divididas em composições de tábuas empíricas (tábuas-modelo), tábuas derivadas da teoria de populações estáveis ou quase-estáveis, ou aplicações mistas dessas duas formas. Comentamos sucintamente algumas tábuas construídas para representar a mortalidade brasileira de 1870 a 1950.

Palavras-chave: demografia histórica; mortalidade; tábua de mortalidade.

[1] Uniconsult Serviços Técnicos Ltda. (Rio de Janeiro) End. Rua Paulo César de Andrade, 450/101 Laranjeiras-RJ, 22221-090 – email: moura@uniconsult-rj.com.br

1. Introdução

...homens não morrem em Proporções exatas, nem em Frações...
John Graunt, 1662[2]

Em que pese ser indubitavelmente verdadeira a afirmação de Graunt, tábuas de mortalidade (ou de sobrevivência) são instrumentos fundamentais para a representação dos mecanismos demográficos.

Uma tábua de mortalidade agrupa um conjunto de indicadores que quantificam, sob vários aspectos, a evolução da mortalidade de um grupo, à medida em que seus componentes envelhecem: probabilidade de morrer no ano seguinte ao completar certa idade, número de sobreviventes que completam cada idade, número de óbitos em cada idade, entre outras medidas. Do ponto de vista da fundamentação metodológica desta evolução, podemos fazer uma primeira distinção, entre tábuas "empíricas", que têm por referencial populações existentes ou que existiram no passado, e tábuas "matemáticas", modeladas sobre características empíricas, mas parametrizadas em função de variáveis demográficas, como a variação natural da população ou sua esperança de vida ao nascer. Estudos demográficos fazem uso crescente de combinações de ambos esses tipos, aproveitando as regularidades embutidas nas tábuas teóricas para suprir informações desconhecidas nos dados em estudo. Embora possa eventualmente não haver maiores diferenças numéricas entre duas tábuas, uma originada de levantamentos empíricos e outra calculada exclusivamente a partir de fórmulas matemáticas, cremos importante que o usuário desse instrumento ou dos resultados obtidos por sua aplicação possa apreciar os alicerces epistemológico metodológico de uma e de outra.

Qualquer tábua de mortalidade – empírica ou matemática – pretende representar a evolução da mortalidade de certo grupo. No caso das tábuas empíricas, esta seria um registro da mortalidade efetivamente ocorrida numa população real, em algum período passado. O principal critério de sua avaliação vem a ser a precisão com que traduz a mortalidade ocorrida. No caso das tábuas matemáticas, o grupo retratado não passa de uma população abstrata que tenha certas características definidas, tais como

[2] Em *Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of*

composição por faixas etárias em certas proporções ou taxa de crescimento natural conhecida. Não há sentido em se considerar a precisão de representação como critério de análise de uma tábua matemática, pois, por definição, esta sempre satisfaz exatamente as condições desejadas. Ao compararmos tábuas empíricas com tábuas matemáticas, no entanto, surge a questão da semelhança entre as duas curvas de mortalidade. Para efetuar esta comparação, torna-se preciso entendermos como uma população empírica qualquer poderia atingir – e eventualmente manter – as características da população teórica. Como reverso da medalha, há a questão de se saber qual a curva matemática que melhor representaria certo momento na história da mortalidade de uma população existente. Ao serem estudadas as propriedades matemáticas de populações parametrizadas por sua fecundidade, sua mortalidade ou sua taxa de variação natural, foram desenvolvidos os conceitos de populações estacionárias, estáveis e quase-estáveis, que se tornaram o principal instrumento teórico para fazer a ponte entre populações virtuais, matemáticas, e populações empíricas, estatísticas[3].

Mortality, transcrito em Smith & Keyfitz (1977), nossa tradução.

[3] Desde o início do século XX, Alfred J.Lotka introduziu o conceito de população estável. Smith & Keyfitz (1977) transcrevem três de seus textos, entre 12 trabalhos sobre teoria de populações estáveis. Outras fontes sobre o assunto são: Nathan Keyfitz (1977:77-111) que expõe a teoria e analisa diversos tópicos pontuais; Ansley Coale (1972) faz um estudo sistemático e formal do tema; S.Halli e K.V.Rao (1992:30-34) resumem criticamente seus resultados; o *Manual IV* das Nações Unidas expõe a teoria de populações estáveis e propõe diversos exemplos detalhados de cálculo (United Nations 1967). Noutro relatório, *The Concept of Stable Population. Application to the Study of Populations of Countries with Incomplete Demographic Statistics*, há um tratamento abrangente da teoria e sua aplicação (United Nations 1968). Livros-texto, como Shryock e Siegel (1976:316-20, 487-91) e Keyfitz e Beeckman (1984:55-92), formalizam a exposição desses conceitos. Hervé Le Bras faz a crítica do uso do modelo de população estável com referência a populações históricas, sujeitas a picos de mortalidade, e desenvolve uma análise formal da aproximação à estabilidade (Le Bras, 1971).

Em termos de dados, as estimativas de óbitos e de população em risco, necessárias para o cálculo de tábuas empíricas, dependem da conjugação dos três principais registros demográficos: censos, estatísticas vitais e controle de migrações. Sabemos das dificuldades, mesmo nos dias de hoje, de se obter dados completos e precisos para cada um desses registros, mas tais empecilhos não impediram os demógrafos de construir tábuas por caminhos alternativos, que buscam – cada um a sua maneira – contornar as falhas das estatísticas referentes a dada população. Embora tais detalhes da "cozinha" da demografia sejam amplamente conhecidos por demógrafos de formação, cremos importante explicitar alguns de seus pressupostos, em benefício de historiadores e outros usuários dos resultados obtidos pela aplicação dessas tábuas de mortalidade.

2. Objetivos

Com o objetivo de esclarecer o leitor mais distante da demografia formal sobre os múltiplos embasamentos, empíricos e teóricos, das tábuas de mortalidade, propomos uma tipologia dessas tábuas a partir de seus critérios de construção. Nosso objetivo aqui não é reproduzir as explicações dos manuais e, sim, buscar esclarecer sobre a representatividade de cada tipo de tábua para a análise da mortalidade de populações atuais ou passadas.

Complementando e exemplificando essa proposta de tipologia de tábuas de mortalidade segundo suas fontes e método de construção, analisamos comparativamente algumas tábuas referentes à população brasileira do final do século 19 e primeira metade do século 20, como exemplos da classificação proposta.

3. Tábuas empíricas

Desprezando, por ora, as demais funções que constam de uma tábua de mortalidade, podemos nos fixar no número de pessoas, de um hipotético contingente inicial – 100.000, por exemplo – cuja evolução acompanhamos ano a ano, à medida em que envelhecem, isto é, do seu nascimento até a mais provecta idade que haja.

Podemos supor, como primeira hipótese, que essas 100.000 pessoas tenham sido pessoas de fato existentes e que, portanto, a tábua represente seu histórico de vida (e morte). Tábuas para países que mantêm registros populacionais completos, isto é, que acompanharam cada pessoa através de todas as etapas de sua vida (como, por exemplo, na Suécia a partir de meados do século XVIII) ou que foram calculadas sobre conjuntos fechados de pessoas (como os nobres ingleses ou os contribuintes da previdência social brasileira) se enquadrariam nesta classificação[4]. Note-se que, nestes casos, um único banco de dados indica o estado biológico (se vivo ou morto) e, em geral, também o estado civil de cada habitante, suprindo tanto a informação sobre óbitos ocorridos como sobre a população em risco. Apesar de eventuais discrepâncias (como dupla contagem ou falhas no acompanhamento), esta fonte propicia o único método empírico no qual temos certeza de compatibilidade total entre a população em risco e os óbitos registrados.

É importante distinguir estas tábuas, calculadas sobre populações fechadas e, consequentemente, com informação completa tanto sobre óbitos como sobre a correspondente população em risco a cada época, das pioneiras tábuas calculadas por John Graunt (em 1662), Edmund Halley (em 1693) e outros, construídas somente sobre dados de óbito.

[4] Ver: sobre a Suécia, G.Utterström (1965); sobre os nobres ingleses, T.H.Hollingsworth (1964; 1965); sobre contribuintes da previdência brasileira, K.I.Beltrão & S.Sugahara (2002). T.H.Hollingsworth (1969:197-224) cita um grande número de tais estudos, referentes a populações fechadas definidas por religião, classe social e profissão, entre outros critérios.

Um caso especial de tábuas do tipo "histórico de vida" ocorre em estudos de reconstituição de famílias. Como, nesses estudos, não há usualmente fontes que registrem as emigrações, o tratamento dos dados para apurar a população em risco exige cuidados. Em seu estudo pioneiro sobre a paróquia de Crulai, na Normandia, nos séculos XVII e XVIII, E.Gautier e L.Henry explicitam a condição essencial para o cálculo de uma taxa de mortalidade: conhecer-se não somente os óbitos, mas também, por intermédio de informações desvinculadas dos registros de óbitos, a população correspondente que estaria submetida ao risco, isto é, que estivesse presente no local no mesmo período em que ocorreram os óbitos. Dada a inexistência de estatística de migrações, limitam sua amostra de população sujeita ao risco a uma subpopulação sobre a qual têm forte presunção de ter estado presente: a população infantil até 14 anos, cujos pais se casaram na paróquia e os autores sabem que ainda lá estavam no ano seguinte à morte do filho[5]. Seguindo metodologia semelhante, em seu estudo sobre a paróquia vizinha de Tourouvre-au-Perche, na mesma época, H.Charbonneau calcula a mortalidade para as crianças até 15 anos sobre cujas famílias mantém registros até o ano seguinte a cada óbito. Com este expediente, reúne uma amostra de 4488 casos, agrupados por década do nascimento. Dada a natureza individual dos registros em seu banco de dados, até consegue classificar os óbitos segundo a profissão do pai, gerando tábuas de mortalidade específicas para cada um desses grupos de crianças (Charbonneau, 1970:161-181). Para o cálculo da mortalidade dos adultos, nos dois estudos, os autores não encontraram condições de determinar a população em risco e precisaram recorrer a tábuas-modelo para complementar as informações disponíveis.

[5] Sobre 1615 nascimentos, anotam 578 óbitos e 1037 "perdas de observação". "*L'étude de la mortalité dans une paroisse est particulièrement difficile. On connaît par les registres, le nombre de personnes de tel âge, mettons 50 ans, qui sont mortes au cours de telle période (...) Mais, pour obtenir un taux de mortalité, il faut rapporter ces décès à une population (...) celle des personnes qui ont été exposées à mourir dans la paroisse à 50 ans, c'est-à-dire à l'ensemble des personnes qui (...) ont été présentes dans la paroisse à cet âge. Or, il est difficile de déterminer le nombre d'individus présents dans une paroisse, à un certain âge, au cours d'une certaine période. À défaut nous avons pensé à nous limiter à l'étude de la mortalité d'échantillons de personnes pour lesquelles il existe une forte présomption de présence dans la paroisse; encore faut-il que cete présomption soit tirée de renseignements sans liaison avec la mortalité. Cette condition* impérative *restreint, pratiquemente, l'application du procédé aux enfants." (Gautier e Henry, 1958:159)*

Por extensão desta representação de histórico de vidas, uma vez calculada uma tábua, podemos supor que pessoas não incluídas neste cômputo, mas com características biológicas e sociais semelhantes (habitantes de outros países escandinavos, num dos exemplos acima, ou brasileiros não contribuintes para a seguridade social, no outro), tenham trajetórias biológicas próximas àquela descrita pela tábua à mão. A mesma tábua, mas agora somente "representativa" da mortalidade de outra população, não deve ser confundida (como muitas vezes acontece) com a tábua em sua qualidade de "histórico de vida". Enquanto, num caso, os dados da tábua mostram exatamente a mortalidade ocorrida com o grupo que a gerou, no outro, somente supomos uma proximidade estatística entre uma curva de mortalidade conhecida e outra, desconhecida. Neste último caso, não podemos considerar a tábua como determinante do que aconteceu àquele grupo específico. No entanto, dado um conjunto qualquer de pessoas, vivas e mortas, diferente das que geraram a tábua original, podemos testar se seu padrão de mortalidade se conforma estatisticamente àquele representado pela tábua[6].

Um terceiro tipo de tábua de mortalidade decorre de construção baseada na combinação de duas fontes distintas de informação: levantamentos censitários, que aferem a população em risco, e registros de óbitos (dotados de maior ou menor precisão). Seria o método "registro mais censo", o mais difundido para cômputo de tábuas de mortalidade com abrangência nacional. A tábua é construída a partir da comparação dos óbitos registrados em certa idade, num entorno da data censitária, com a população recenseada naquela faixa etária. Este tipo de tábua, é claro, se aproxima tanto mais da tábua "histórico de vida", quanto mais precisas e completas forem as estatísticas vitais e as do censo. Ocorre que nem um nem outro estão isentos de falhas, principalmente em épocas passadas, dificuldade que marca este método e impõe análises críticas de sua precisão.

No caso brasileiro, apesar da existência de um sistema nacional de registros vitais,

[6] Com o surgimento de técnicas matemáticas e estatísticas apropriadas, desenvolveu-se a análise de sobreviência, que dá tratamento estatístico aos conceitos da tábua de sobrevivência/mortalidade, permitindo aplicações em qualquer área (medicina, ecologia, controle de estoques etc.) onde se possam empregar as noções de probabilidade de permanência versus probabilidade de desaparecimento. O livro *The Life Table and Its Applications*, de Chin Long Chiang (1984), é considerado uma das principais referência no assunto.

o sub registro de óbitos (assim como de nascimentos) era fato já conhecido pelos pioneiros na construção de tábuas de mortalidade. Enquanto Bulhões Carvalho acreditava que as estatísticas vitais do então Distrito Federal – em oposição ao que ocorria noutras regiões do país – teriam representatividade suficiente para o cálculo de tábuas de mortalidade com razoável grau de precisão, Giorgio Mortara insistiu sobre as falhas do nosso sistema de registros de nascimentos e de óbitos[7]. Afirmou, em 1957:

"A falta de estatísticas fidedignas dos nascimentos e dos óbitos coloca o Brasil em condição de penosa inferioridade no domínio da estatística demográfica internacional... em consequência... todos os dados que foram divulgados nos últimos lustros acerca da natalidade e da mortalidade no conjunto do Brasil estão apenas baseados, em parte mais ou menos ampla, em conjeturas e não em levantamentos." (Mortara, 1957:1)

Esta avaliação, todavia, não o impediu de estimar diversas tábuas, referente a períodos desde 1870 até a década de 1950. Lyra Madeira, quase uma década depois, reafirmava essas falhas dos registros vitais: "A evolução demográfica brasileira é conhecida apenas naquilo que os 7 recenseamentos gerais permitem obter. Assim, as duas componentes demográficas mais importantes, isto é, a mortalidade e a natalidade, são em grande parte desconhecidas." Sobre as tábuas calculadas sob a direção de Mortara para as capitais e alguns Estados insistiu: "Nenhuma dessas tábuas, porém, pode ser utilizada para a totalidade do país (Madeira, 1966:86)." Detalhado relatório internacional sobre fecundidade e mortalidade no Brasil, de 1983, também considerou serem pouco confiáveis as estatísticas vitais anteriores a 1970 (Committee on Population and Demography, 1983:2,18).

---

[7] Diz Bulhões Carvalho: "Um número mais avultado [de população] poderia agradar a vaidade nacional, mas não exprimiria a realidade, sendo facilmente comprovada a sua inexactidão pelo confronto com as cifras normaes da natalidade e da mortalidade, – comparação essa que demonstra, ao contrário, de modo evidente, a exactidão approximada dos resultados obtidos no recenseamento effectuado em Setembro de 1920. (...) O confronto da população recenseada em 1920 com o obituário occorido no mesmo anno ... revela o coeficiente de 19,1 obitos por 1000 habitantes, taxa mortuaria ... assás razoavel para uma cidade localizada, como é a do Rio de Janeiro, na zona tropical." (Brasil. Directoria Geral de Estatistica, 1923:XVII) E Mortara: "Cumpre observar que as tábuas para o Distrito federal se referem a uma circunscrição territorial em que o registro dos nascimentos e dos óbitos, em 1920, já estava organizado de maneira bastante satisfatória, e compreendia a máxima parte, se não a totalidade, dos casos que se verificavam. Logo, as tábuas de mortalidade do Distrito Federal fornecem uma representação suficientemente aproximada da realidade do fenômeno por elas descrito." (Mortara, 1940b:675) E sobre o país em geral: "*An almost insuperable obstacle to the studies on Brazilian demography is the incomplete registration of births and deaths and the negligence of registration officers, who often omit to notify the central statistical service.*" (Mortara, 1954)

Na utilização do método "registro mais censo", além dos problemas com a fidedignidade das informações sobre óbitos, devemos também considerar os problemas censitários de subenumeração das populações em certas idades e, mais importante ainda, dos erros sistemáticos nas declarações de idade que distorciam os censos até muito recentemente. As técnicas de construção de tábuas de mortalidade procuram, sem dúvida, minimizar tais problemas (por exemplo, pelo agrupamento em faixas quinquenais, graduação de faixas agrupadas e suavização de curvas por diversos métodos). Independentemente do sucesso dessas técnicas em sobrepor tais dificuldades, convém ter em mente, ao considerarmos as diferenças conceituais entre tábuas, quais foram os procedimentos metodológicos empregados na sua construção. A esses problemas específicos dos dados sobre óbitos e dos dados populacionais, ainda devemos acrescentar a necessidade de compatibilizar os limites temporais, geográficos e de cobertura de registro da população onde sobrevieram os óbitos com os da população recenseada.

Um quarto método de construção, que evita as falhas dos registros vitais, mas que sofre duplamente as falhas censitárias, é baseado em dois "censos sucessivos". Partindo da população existente com idade x no primeiro censo, estima-se a mortalidade desta faixa etária pelos sobreviventes com idade $x+c$, à época do segundo censo, realizado c anos após o primeiro. Este método foi criado e divulgado como uma opção factível e suficientemente aproximada para o cálculo da mortalidade em países sem registros de óbitos ou com registros não confiáveis (Coale & Demeny, 1966:6). Depende, no entanto, de pressupostos fortes: que a população seja absolutamente fechada à migração e – requisito talvez até mais problemático – que o nível de sub- ou super-enumeração, em cada faixa etária, tenha sido igual nos dois censos. Além disso, o método não permite estimar a mortalidade para idades inferiores ao número de anos do intervalo inter censitário c. Para uma região sem qualquer informação sobre eventos vitais, este procedimento pode constituir, sem dúvida, uma boa aproximação para o cálculo de sua curva de mortalidade, mas exige uma avaliação apurada da precisão dos censos empregados, em especial quanto a questão de erros nas declarações de idade[8].

[8] Sobre a técnica de estimação de tábua a partir de dois censos, ver A.Coale (1984).

Quadro 1
Tipologia das tábuas de mortalidade – as tábuas empíricas

Categoria geral	Tipo	Método	Exemplos / Referências
Tábuas empíricas	histórico de vida	Registro universal para todo um país	• Suécia a partir de 1750 (Utterström, 1965) • Bélgica a partir de meados do século XIX
		população fechada	• nobreza inglesa (Hollinsworth, 1964; 1965) • contribuintes do sistema previdenciário brasileiro (Beltrão & Sugahara, 2002)
		reconstituição de famílias	• Crulai (Gautier & Henry, 1958) • Tourouvre-au-Perche (Charbonneau, 1970)
	tábua representativa		• cálculos atuariais referentes à projeção de mortalidade de populações em vida, baseados em tábuas-padrão • estudos de demografia histórica sobre populações cujas características demográficas associamos a populações com tábuas calculadas
	registro de óbitos mais censo		• Brasil-Capitais; Brasil-DF 1920 Bulhões Carvalho (Brasil. Directoria Geral de Estatistica, 1928) (Mortara, 1940b) • Brasil-DF; capitais 1940 G.Mortara • Brasil IBGE, tábuas nos anos censitários desde 1970
	censos sucessivos		• Brasil 1870-1890 G.Mortara (1940a; 1941a; b) • Brasil 1890-1920 G.Mortara (idem anterior) • Brasil 1940-1950 G.Mortara (1952; 1957)
	técnica para calcular mortalidade infantil	filhos tidos, filhos Mortos	• Mortara (1956) • W.Brass (1968:88-139)

G.Mortara, também neste campo, mostrou-se pioneiro ao realizar, na década de 1950, avaliações de mortalidade a partir de dois recenseamentos. Nas suas próprias palavras:

> Persistindo a falta de estatísticas fidedignas dos nascimentos e dos óbitos para a maior parte do Brasil, torna-se indispensável aproveitar todos os elementos apropriados para trazer informações sobre esses assuntos, que podem ser obtidos com o auxílio dos dados dos censos demográficos. O presente estudo visa justamente a obter estimativas dos níveis da mortalidade pela comparação entre os resultados dos dois últimos censos.
>
> (Mortara 1953)

O Quadro 1 resume estes comentários de forma esquemática.

Até aqui, tratamos de métodos que podemos agrupar num grande grupo de tábuas "empíricas", já que, de uma forma ou de outra, derivam simultaneamente de óbitos registrados e de populações existentes. Tais dados são compilados em registros estatísticos, ou seja buscam na própria população as referências para construir as curvas de sua mortalidade. Técnicas matemáticas são, sem dúvida, aplicadas na sua construção, mas unicamente para corrigir distorções percebidas (falhas na informações de idade, por exemplo) e "suavizar" a curva obtida com a informação empírica. O nível de mortalidade não é definido através de parâmetros ou formulações matemáticas.

Outra distinção conceitual inerente a tábuas de mortalidade torna-se importante ao analisarmos épocas de mudanças demográficas mais acentuadas, isto é, épocas em que as taxas de fecundidade total e de mortalidade total não se mostram estáveis. Trata-se do entendimento de uma tábua como indicativa da mortalidade, em certo momento, de um conjunto de pessoas de várias idades, isto é, a tábua como uma "visão corrente" das características de toda uma população, em oposição à tábua como indicativa do que aconteceu ou acontecerá a um grupo etário específico, à medida em que envelhece, isto é, a tábua como uma "visão de coorte". No primeiro caso, as características de mortalidade valeriam para toda a população considerada, mas somente para um período restrito e, no segundo, valeriam por toda a vida, mas somente para aquela coorte. Em épocas de alterações importantes na dinâmica demográfica, a conjugação dessas duas perspectivas sobre a curva de mortalidade ajuda a entender melhor que subgrupos estão sofrendo as mudanças e como se comportam sua mortalidade específica e a da população como um todo.

4. Tábuas Matemáticas

As tábuas do segundo tipo, que chamamos de "matemáticas", podem ser derivadas exclusivamente de pressupostos teóricos, da teoria de populações estáveis ou quase-estáveis, ou podem combinar estas hipóteses com tábuas-modelo, empregadas como base para determinar as "formas gerais" da curva de mortalidade. A teoria das populações estáveis ou quase-estáveis demonstra a existência de relações lógicas entre

combinações de taxas de natalidade e de mortalidade e a composição etária que atingirá uma população submetida permanentemente àquelas taxas[9].

Ao contrário das tábuas empíricas, as matemática são parametrizadas, usualmente por uma e às vezes por mais variáveis, de modo a definir um conjunto de tábuas que, dado um formato geral ou um critério teórico, correspondam exatamente a certos níveis de mortalidade ou de esperança de vida ao nascer. Entretanto, têm necessariamente um embasamento empírico na sua construção, pois buscam sempre um formato de curva de mortalidade biologicamente plausível, isto é, que corresponda ao fenômeno empírico verificado em qualquer população humana. Este formato tem por características principais uma mortalidade infantil mais alta do que a dos adolescentes, um recrudecimento da mortalidade para jovens adultos e um aumento gradativo da taxa específica de mortalidade dos adultos, à medida do seu envelhecimento. Além disso, as curvas para mulheres mostram-se quase sempre abaixo das curvas para homens. Praticamente todas as curvas de mortalidade conhecidas seguem estas grandes linhas.

Além dessa conformação geral, a construção de tábuas-modelo se restringe a combinações plausíveis de mortalidade em cada faixa etária, ou seja, adota limites mínimos e máximos históricos para q_x, a probabilidade de óbito no ano seguinte ao momento em que cada pessoa completa a idade x. Ademais, nem toda combinação de valores de q_x (ao variar x) é possível, pois sabe-se empiricamente que curvas de mortalidade, por exemplo mais próximas do limite historicamente máximo, dificilmente apresentariam, para algum x, q_x muito distante desse limite e, portanto, com valor mais próximo do limite historicamente inferior[10].

---

[9] "As Lotka has demonstrated, a population with constant birth rates and constant mortality rates will exhibit constant proportions of numbers in the different sex-age groups, which are determined by the birth and mortality rates. It has been further demonstrated that changes in mortality, unless they are very radical, have only slight effects on this stable age structure. Hence, an estimate of the birth rate, combined with a very rough estimate of the death rate, suffices to construct an approximate age distribution if it can be assumed that the birth rate has been nearly constant for a long period. (United Nations, 1956:17)"

[10] "So far as is known, at each age x there are upper and lower limits such that, in the past (and the foreseeable future), the q_x at that age for human populations has never exceeded the upper limit nor fallen below the lower. It might be supposed then that a set of q values such that, at each age, the q_x was within the permissible limits for that age, might be the basis of a life-table for some population. Observed sets of q's suggest otherwise. Such factors as the social, economic and cultural conditions of a population to some extent tend to influence the mortality level at every age so that if a country has a value of q_x near the lower limit at some age, it is likely to be near this limit at other ages also, and it is extremely improbable that at any age it might be near the upper limit." (Carrier e Hobcraft, 1971:7)

Assim como ocorreu com algumas das técnicas de construção de tábuas empíricas, o desenvolvimento deste tipo de tábuas surgiu como um recurso para a análise da mortalidade em países com registros vitais e processos censitários deficientes. Desde a década de 1940, um grupo na Universidade de Princeton se preocupou com a modelagem de tais tábuas, desaguando estes estudos nos trabalhos que Ansley Coale e Paul Demeny publicaram em 1966, no seu Regional Model Life Tables and Stable Populations (Coale & Demeny, 1966), a mais citada referência sobre tábuas-modelo e também sobre populações estáveis.

As primeiras tábuas-modelo, no entanto, foram publicadas pelas Nações Unidas. A partir de 1952 até a década de 1980, as Nações Unidas publicaram e revisaram tábuas parametrizadas, representativas da mortalidade em diversos países, dentro de seu programa de divulgação e padronização de técnicas demográficas (United Nations, 1952; 1955; 1956; 1967; 1968; 1983). O fundamento empírico dessas tábuas das Nações Unidas (bem como das de Coale e Demeny) provém de um conjunto de numerosas tábuas, representativas de vários países, que serviu de base para regressões entre as probabilidades de morte no ano seguinte, a cada idade (q_x) e uma variável-âncora (por exemplo, q_{10} ou e_0, a esperança de vida ao nascer). Os resultados dessas regressões foram em seguida usados para definir formatos médios das curvas de mortalidade, isto é, proporções entre as mortalidades específicas em certas idades, que, em conjunto, definem "fôrmas" às quais uma curva de mortalidade deveria aderir para "não fugir à experiência geral". Cada uma das curvas de um formato análogo apresenta certo nível de mortalidade e, portanto, certa esperança de vida. Ao inverter o sentido dessa informação, partindo de um nível de mortalidade dado (ou equivalentemente de um nível de esperança de vida) para se chegar à curva que cumpra este requisito, criou-se o sistema de tábuas-modelo de mortalidade.

Dentre os possíveis formatos gerais de curvas de mortalidade que estudaram, Coale e Demeny selecionaram 4 famílias de tábuas, nomeadas como Norte, Sul, Leste e Oeste, que agrupavam curvas de mortalidade assemelhadas. Estas famílias de tábuas são aplicadas até hoje como representativas de padrões distintos de mortalidade. A validade geral dessas 4 famílias, no entanto, sofre algumas críticas. Uma primeira diz respeito às fontes de dados empregadas para chegarem às curvas "médias", já que

a maioria dessas tábuas se referem a países europeus, devido à pequena disponibilidade, na época, de tábuas de outras regiões. Uma segunda, mais geral, argumenta que é improvável que somente esses 4 formatos representem todas as experiências de mortalidade existentes. Essas dificuldades motivaram William Brass a propor um sistema de tábuas com dois parâmetros definidores do formato das curvas, permitindo ajustes mais flexíveis a situações empíricas específicas, em especial com relação a diferentes combinações de mortalidade infantil e mortalidade adulta (Brass & Coale, 1968:122-135). Esta questão, aliás, é um dos problemas que mais comumente ocorrem na análise de contextos individuais. O estudo de 1983 do Panel on Brazil, do Committee on *Population and Demography* (mencionado adiante), após analisar os dados sobre a mortalidade brasileira a partir de diversos métodos de estimação, opta por combinar uma modelagem em que a mortalidade infantil segue a familia "Sul" de Coale e Demeny, enquanto a mortalidade adulta segue a família "Oeste" (Committee on Population and Demography, 1983:89).

Sully Ledermann, do *Institut National d'Études Démographiques* (INED) francês, foi pioneiro em buscar uma formalização genérica, aplicável a qualquer padrão de mortalidade, através de um indicador único (Ledermann e Bréas, 1959). Seu estudo, baseado em técnicas de análise fatorial, mostra que as 157 tábuas que compilou, de países em diversos estágios de desenvolvimento demográfico e em diversas épocas, poderiam ser adequadamente representadas (com altos graus de explicação) por uma só dentre as seguintes variáveis: a esperança de vida ao nascer, a mortalidade de ambos os sexos entre 0 e 5 anos ou ainda a mortalidade de mulheres adultas entre 20 e 50 anos. Sua técnica foi postumamente traduzida em um conjunto de tábuas-modelo por Hervé le Bras (Le Bras 1968; Ledermann 1969). Estas tábuas, embora citadas em trabalhos francófonos, não obtiveram a notoriedade das tábuas publicadas nos Estados Unidos, em mais um exemplo das barreiras "extra-acadêmicas" impostas à produção intelectual fora da comunidade anglo-saxônica. Seguindo esta trilha, demógrafos do INED produziram novo sistema de tábuas-modelo, com parametrização tripla, de nupcialidade, fecundidade e mortalidade (Demonet, Dupâquier et al., 1977), que permite um ajuste "inteligente", isto é, explicável metodologicamente, a uma grande variedade de situações empíricas.

A partir dos anos 1970, demógrafos brasileiros passaram a buscar um sistema de tábuas-modelo adaptado às condições brasileiras. Em que pese a existência de poucas tábuas brasileiras calculadas pelo método "registro mais censo" (22), desigualmente distribuídas entre 1920 e 1970 e entre as diversas regiões do país (cobrindo algumas capitais e o Estado de São Paulo), Luiz A.M.Frias e Paulo Rodrigues calcularam uma família de tábuas, parametrizadas por uma variável – a esperança de vida ao nascer – e com representatividade estatística para toda a gama de níveis de mortalidade (Frias e Rodrigues, 1980).

A grande investida em estudos das populações dos países menos desenvolvidos, principalmente em centros norte-americanos, durante as duas décadas posteriores a 1950, época do medo da "explosão demográfica", gerou outro grupo de técnicas, de estimação de fecundidade e de mortalidade a partir de pesquisas sobre o número de filhos tidos, filhos ainda vivos e filhos tidos no último ano. Tais dados, combinados com tábuas-modelo e técnicas de populações estáveis propiciaram o cálculo de tábuas de mortalidade. Apesar de ter sido Mortara (1956) pioneiro em "utilizar proporções de crianças mortas, entre todos filhos tidos, conforme relatado por mulheres de várias idades, como um indicador de mortalidade infantil"[11], foi William Brass quem primeiro formalizou a técnica e a aplicou com sucesso, num grande projeto de estudo da população da África tropical (Brass, Coale et al., 1968).

[11] Segundo o *Panel on Brazil*, do *Committee on Population and Demography*, "Mortara foi o primeiro a utilizar proporções de crianças mortas, entre todos filhos tidos, conforme relatado por mulheres de várias idades, como um indicador de mortalidade infantil. Brass desenvolveu esta idéia, propondo um procedimento que permitia converter as proporções observadas de crianças mortas em probabilidades de morrer entre o nascimento e certa idade exata x, numa tábua de sobrevivência. (Committee on Populatoin and Demography, 1983:26) Este *Panel on Brazil* estava composto por Carmen Arretx (Celade), William Brass (London School of Tropical Medicine), José Alberto Magno de Carvalho (Cedeplar-UFMG), Valéria da Motta Leite (IBGE), Thomas W. Merrick (Georgetown University) e Axel I. Mundigo (México).

Quadro 2
Tipologia das tábuas de mortalidade – algumas tábuas matemáticas

Categoria geral	Tipo	Método	Exemplos / Referências
Tábuas matemáticas	Modelo empírico	sistema de parâmetro único (e_0, $_5q_0$, $_{20}q^F_{30}$), calculado por análise fatorial	• S.Lederman e J.Bréas (1959) • S.Lederman (1969)
		q_x calculadas como regressão sobre taxas de tábuas empiricas	• tábuas-modelo genérica, Nações Unidas • 4 famílias de tábuas-modelo, Coale e Demeny (1966) • tábuas-modelo para a experiência brasileira, 1920-70. Frias e Rodrigues (1980)
		modelo logito, de 2 parâmetros, sobre curva empírica	• W.Brass (1968:88-139)
	População estável ou quase-estável	sistema de tábuas parametrizadas por uma variável	• populações estáveis de Coale e Demeny (1966)
		sistema de tábuas parametrizadas por 3 variáveis (nupcialidade, fecundidade e mortalidade)	• M.Demonet, J.Dupâquier, H.Le Bras (1977)
	Aplicação de população estável	tábua teórica aferida para parâmetros censitários de países latino-americanos	• E.Arriaga (1976)

Em termos metodológicos, vimos importantes diferenças entre tábuas empíricas, construídas sobre dados estatísticos, e tábuas teóricas, aplicações de regras matemáticas que buscam parametrizar variadas condições de mortalidade. Quanto ao resultado numérico derivado de uma ou de outra, as diferenças podem mostrar-se pequenas ou mesmo desaparecer. Assim, embora as questões epistemológicas permaneçam como um pano de fundo a nos lembrar a origem daquela tabela numérica, o valor instrumental precípuo de uma tábua de mortalidade consiste em sua precisão na representação da mortalidade de certo grupo, em certo período. Esta precisão poderá decorrer tanto de uma contagem apurada de eventos reais, quanto da adequação de uma fórmula matemática aplicada a dados empíricos. A metodologia empregada torna-se, neste ponto, mero recurso heurístico ou argumentativo, pois o importante é representar o mais precisamente possível a mortalidade desconhecida.

5. Algumas tábuas de mortalidade brasileiras

A primeira tábua de mortalidade para a população brasileira foi estimada sobre dados censitários apurados para 1920 e os registros de óbitos das capitais referentes ao mesmo ano, por José Luiz Bulhões Carvalho, que então dirigia a Diretoria Geral de Estatística. Ele calculou três tábuas (para homens, para mulheres e para a população total) em duas regiões geográficas (no Distrito Federal e num conjunto de capitais, incluindo o próprio DF), totalizando 6 tábuas.

Duas décadas depois, G.Mortara estimou tábuas (com homens e mulheres em conjunto) para 1870-1890 e para 1890-1920, com base nas populações apuradas nos censos e em estimativas de natalidade e mortalidade médias nesses períodos intercensitários. Ou seja, fixou uma natalidade geral média ao longo de todo o período, estimando a mortalidade pelo número de sobreviventes computados nos censos de 1890 e 1920. Mortara estimava que o censo de 1920 havia sido superenumerado, possivelmente a partir de "ajustes" arbitrários, não explicitados nas publicações. No entanto, quanto aos censos de 1872 e de 1890, que empregou na estimativa da natalidade e da mortalidade média do Brasil, considerou que, na opinião "dos competentes", "[estes censos] são efetivamente bastante próximos da verdade..."(Mortara, 1941b:511), o que sabemos ser um otimismo oficialista algo exagerado.

Ao final da década de 1960, dentro do programa de longo prazo de estudos sobre demografia latino-americana mantido pelo International Population and Urban Research– Institute of International Studies da Universidade da California em Berkeley, Eduardo Arriaga calculou novas estimativas para homens e para mulheres, referentes aos anos de 1872, 1890, 1900, 1920, 1940, 1950 e 1960, com base nas distribuições etárias dos censos, depois de ajustadas para retificar os erros de declaração de idade e de subenumeração que identificou.

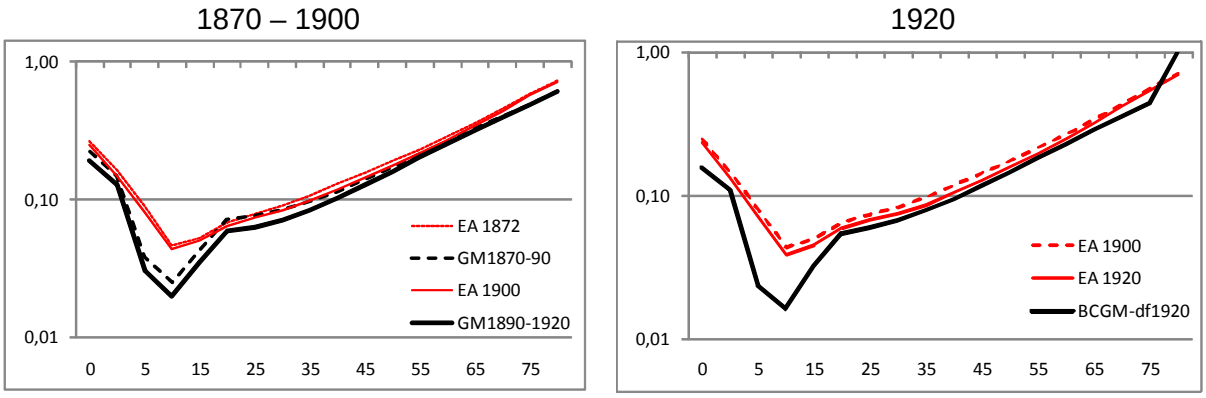
Quadro 3
Brasil. Tábuas de mortalidade referentes ao período de 1870 a 1920

Período de referência	Autor	Sexos	Fontes, metodologia e referências bibliográficas
1920	Bulhões Carvalho	H, M, H+M	Registro de óbitos das capitais e população em risco segundo o Recenseamento de 1920 (Brasil. Directoria Geral de Estatistica, 1928)
	Bulhões Carvalho, com ajuste de G.Mortara	H+M	idem, suavizadas por ajuste matemático (Mortara, 1940b)
1870-1890 e 1890-1920	Giorgio Mortara	H+M	Populações censitárias e óbitos reconstituídos a partir de hipótese de natalidade geral (Mortara, 1941b)
anos censitários de 1872 a 1960	Eduardo Arriaga	H, M	Distribuições etárias por grupos decenais a partir dos dados censitários corrigidos para subenumeração. l_x calculados pela equação de populações estáveis. (Arriaga, 1976)

Para construir suas tábuas, E. Arriaga se vale de resultado da teoria de populações estáveis, utilizando as distribuições relativas da população por grupos etários decenais, entre 10 e 59 anos, e a taxa de crescimento natural da população. As notórias dificuldades de subenumeração de crianças e de velhos seriam assim contornadas. Além disso, os cálculos não seriam afetados pela existência de subenumeração, contanto que o percentual desta seja igual em todos os grupos etários decenais. Há outros dois condicionantes: a constância da fecundidade no período anterior e a insignificância de migrações. No caso do Brasil até 1920, a primeira condição estaria naturalmente satisfeita. Para contornar a segunda dificuldade, Arriaga utilizou os dados censitários referentes a brasileiros natos.

As curvas de probabilidades de morte, q_x, construídas por Arriaga não apresentam a característica "barriga" de baixas probabilidades de morte nas idades entre 1 e 15 anos. A partir dos 20 anos, no entanto, as três curvas se aproximam. Esta discrepância com relação às tábuas de Mortara deve-se, entre outros fatores, ao método de cálculo agregado (em faixas decenais) que empregou. Lembramos que a tábua de Mortara referente ao período setembro de 1870-setembro 1890 corresponderia, numa média pontual, a uma base em setembro de 1880. Da mesma forma, a tábua referente ao período setembro de 1890-setembro de 1920, corresponderia a uma base em setembro de 1905.

Figura 1
Tábuas referentes ao período 1870-1920:
probabilidades de morte por faixas etárias (q_0, $_4q_1$ e $_5q_x$, para x = 5,10,...80)



Fontes: As mencionadas no Quadro 3.

A principal diferença entre os dois grupos de tábuas reside no intervalo de 1 a 15 anos, o que podemos evidenciar pela comparação entre esperanças de vida, calculadas por nós nas idades 0 e 20 (V. Quadro 5). Enquanto as esperanças de vida ao nascer mostram discrepâncias mais importantes entre as tábuas de cada estudioso (diferença de 3,2 anos ou 8,5% da e_{20} média), tais discrepâncias se reduzem significativamente aos 20 anos (diferença de 1,8 anos ou 2,5% da e 20 média).

Em 1940 foi realizado censo considerado bastante preciso, especialmente com relação ao histórico brasileiro até então. O de 1950, no entanto, teria sofrido mais com problemas de subenumeração. Quando ao de 1960, seus resultados detalhados só foram publicados duas décadas depois, prejudicando o cômputo de tábuas na época. Além destas dificuldades censitárias, permaneciam as falhas nos registros de óbitos e de nascimentos, mantendo o cálculo de tábuas de mortalidade como um exercício envolvendo algum malabarismo demográfico. Sobre este período, mencionamos as tábuas listadas no Quadro 4, construídas por diversos métodos.

Quadro 4
Brasil. Algumas tábuas de mortalidade referentes ao período de 1936 a 1950

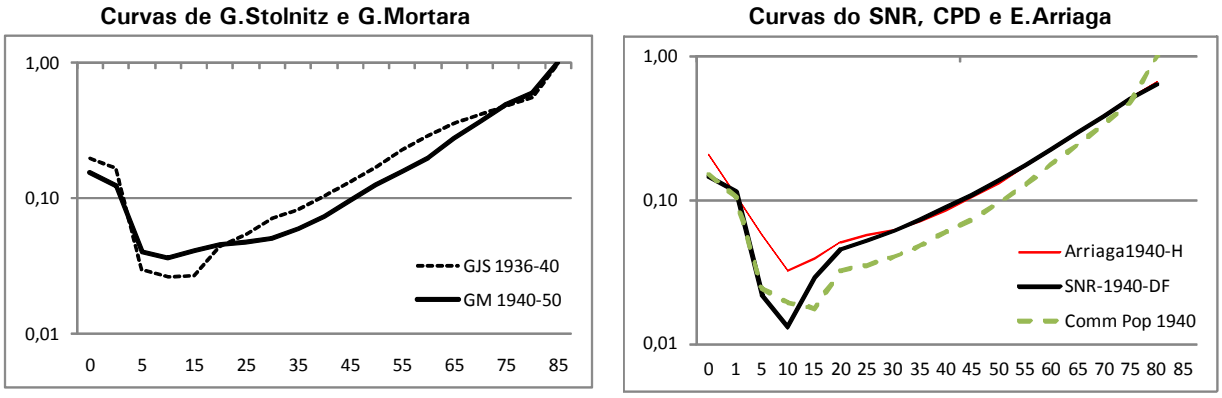
Período de referência	Autor	Sexos	Fontes, metodologia e referências bibliográficas
1940	Eduardo Arriaga	H, M	Idem Arriaga Quadro 3
	SNR DF 1939-41	H, M	Referente à população do Distrito Federal, com óbitos do período 1939-41, elaborada pelo Serviço Nacional de Recenseamento; tábua T13 em Mortara (1946) ou tábua TS-V-1 bis H em Mortara (1945).
	Committee on Population and Demography (CPD)	H, M	Nível de mortalidade infantil estimado pela técnica de filhos tidos (United Nations, 1983); curva distribuída conforme o modelo "Sul" de Coale; nível de mortalidade adulta estimada pela análise comparativa de níveis de mortalidade obtidos segundo diversas técnicas, distribuída conforme o modelo "Oeste" (Committee on Population and Demography, 1983)
1936-40	George J. Stolnitz	H, M	Tábua calculada por comparações de grupos etários quinquenais em 2 censos sucessivos. Transcrita em Mortara (1957:3)
1940-50	Giorgio Mortara	H	Tábua calculada pela comparação entre os censos de 1940 e de 1950. Mortara (1957:8)
1950	Eduardo Arriaga	H, M	Idem Quadro 3 (Arriaga 1976)
	G.Mortara 1949-51	H, M	Mortara (1952)
	Celade 1950-55	H, M	Estimativas por componentes demográficos sobre dados fornecidos pelo IBGE (censo, registro vital, PNAD e DHS).
	CPD	H, M	Idem CPD acima

Nota: Com exceção da tábua SNR DF 1939-41, todas as demais são apresentadas como representando a mortalidade da totalidade da população brasileira.

Na figura 2, vemos algumas semelhanças de formato entre as curvas de G.Stolnitz e G.Mortara, construídas por comparações entre as populações em faixas quinquenais, num mesmo ano no caso de Stolnitz e entre 1940 e 1950, no caso de Mortara. Ficam ambas, contudo, em níveis mais altos de mortalidade do que as demais curvas. Mortara identificou uma mortalidade excessiva em cerca de 21% às estimativas oficiais na tábua produzida por Stolnitz.

As tábuas de E. Arriaga para este período foram obtidas pelo mesmo método já mencionado, baseado na distribuição da população nos momentos censitários, em faixas decenais. As do Serviço Nacional do Recenseamento foram construídas pelo método tradicional de "registro mais censo". As curvas do Celade e do CPD buscaram fontes variadas para cada componente demográfico ou para aferir níveis diferenciados de mortalidade por grupo etário, sendo que esta última chega aos mais baixos níveis de mortalidade entre todas as curvas. É de se notar que as curvas do SNR e do CPD, ao contrário da de E. Arriaga, apresentam a reduzida mortalidade na faixa dos adolescentes, sendo que a do CPD se descola das outras duas a partir dos 20 anos, permanecendo bem abaixo delas.

Figura 2
Brasil. Tábuas referentes a 1940:
probabilidades de morte por faixas etárias ($_nq_0$, $_4q_1$ e $_5q_x$, para x = 5,10,...80)



Fontes: As mencionadas no Quadro 4.

Estas diferenças se refletem nas seguintes esperanças de vida ao nascer, aos 20 e aos 60 anos (calculadas por nós):

Quadro 5
Esperanças de vida ao nascer, aos 20 e aos 60 anos de idade, para diversas tábuas de mortalidade

Tábua	e_0	e_{20}	e_{60}
CPD	44,3	40,7	12,6
SNR	40,8	36,2	12,0
G.Mortara	39,3	37,6	12,2
E.Arriaga	36,1	35,8	11,7
G.Stolnitz	34,2	33,7	10,3

Fonte: Dados nas fontes mencionadas no Quadro 4.

Trazemos estes indicadores como ilustração das marcadas diferenças entre as tábuas, sem buscar as razões para tais diferenças, o que nos levaria além dos limites deste trabalho .

Conclusões

Cremos ter mostrado como cada metodologia para cálculo de tábuas de mortalidade busca apoiar-se, nessa construção, em dados empíricos e balizamentos teóricos bastante distintos. Entender como cada tábua foi construída é pré-requisito importante para uma análise informada dos resultados obtidos. Com estes comentários, esperamos ter motivado os usuários das tábuas de mortalidade, em especial daquelas relativas ao passado mais distante (quando havia menor padronização censitária e, até, ausência de registros demográficos) para o interesse em conhecer-se as fontes de dados e os procedimentos metodológicos com os quais foram construídas.

Referências bibliográficas

ARRIAGA, E. New Life Tables for Latin American Populations in the Nineteenth and Twentieth Centuries. Westport, Conn: Greenwood Press-University of California. 1976 [1968]

BELTRÃO, K.I. e S.SUGAHARA, Tábua de mortalidade para os funcionários públicos civis federais do poder executivo por sexo e escolaridade: comparação com tábuas do mercado. ENCE-Escola Nacional de Ciências Estatísticas, Textos para discussão, Rio de Janeiro, vol.3, n.52, 2002.

BRASIL. Directoria Geral de Estatistica. Recenseamento do Brazil realizado em 1 de setembro de 1920. V.II (1a.parte) População do Rio de Janeiro (Districto Federal). Rio de Janeiro, Ministério da Agricultura, Industria e Comércio, 1923. Disponível em: http://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/Censode1920 /RecenGeraldoBrasil1920_v2_Parte1_Populacao_do_RJ.pdf

Recenseamento do Brazil realizado em 1 de setembro de 1920. Vol.IV (2a.parte, Tomo I) População do Rio de Janeiro (Districto Federal). Rio de Janeiro, Ministério da Agricultura, Industria e Comércio, 1928. Disponível em: http://biblioteca.ibge.gov.br/visualizacao /monografias/GEBIS%20-%20RJ/Censode1920/RecenGeraldoBrasil1920_v4_Parte2_ tomo1_Populacao.pdf

BRASS, W. e A.J.COALE, Methods of Analysis and Estimation. In A.J. COALE. The Demography of Tropical Africa. Princeton NJ, Office of Population Research. Princeton University, 1968.

BRASS, W., COALE, A.J., et al. The Demography of Tropical Africa. Princeton NJ: Office of Population Research. Princeton University, 1968.

CARRIER, N. e J.HOBCRAFT, Demographic Estimation for Developing Societies. A manual of techniques for the detection and reduction of errors in demographic data. London, Population Investigation Committee. London School of Economics, 1971.

CHARBONNEAU, H., Tourouvre-au-Perche aux XVII e et XVIIIe siècles. Paris, Presses Universitaires de France, 1970.

CHIANG, C.L., The Life Table and Its Applications. Malabar, Fla., Robert E. Krieger Publishing Company, 1984.

COALE, A.J., The Growth and Structure of Human Populations. A Mathematical Investigation. Princeton NJ, Princeton University Press, 1972.

Life Table Construction on the Basis of Two Enumerations of a Closed Population. Population Index, Princeton, NJ, vol.50, n.2, p.193-213, 1984.

COALE, A.J. e P.DEMENY, Regional Model Life Tables and Stable Populations. Princeton NJ, Princeton University Press, 1966.

COMMITTEE ON POPULATION AND DEMOGRAPHY, Levels and Recent Trends in Fertility and Mortality in Brazil. Washington DC, National Research Council, 1983. Disponível em: http://books.google.com.br/books?id=zzwrAAAAYAAJ&pg=PP18&dq=%22committee+on+population+and+demography%22&lr=&num=50&cd=127#v=onepage&q=&f=false

DEMONET, M., DUPÂQUIER, J. et al., A Repertory of Stable Populations. In R.D. LEE. Population Patterns in the Past. New York, Academic Press. p.297-309, 1977.

FRIAS, L.A.d.M. e P.RODRIGUES, Brasil: Tábuas modelo de mortalidade e populações estáveis. Anais ABEP, p.420-530, 1980. Disponível em: http://www.abep.nepo.unicamp.br/docs/ anais/pdf /1980/T80V01A10.pdf

GAUTIER, E. e L.HENRY, La population de Crulai, paroisse normande. Paris, INED, 1958.

GRAUNT, J. Natural and Political Observations Mentioned in a Following Index, and Made Upon the Bills of Mortality. In D. SMITH e N.KEYFITZ, Mathematical Demography. Selected Papers. Berlin, Springer-Verlag, 1977.

HALLI, S.S. e K.V.RAO, Advanced Techniques of Population Analysis. New York, Plenum Press, 1992.

HOLLINGSWORTH, T.H. The demography of the British peerage. Population Studies, v.XVIII, p.1-168, nov 1964.

A Demographic Study of the British Ducal Families. In D.V. GLASS e D.E.C. EVERSLEY, Population in History. London, Edward Arnold, p. 354-378, 1965.

Historical Demography. Ithaca-NY, Cornell University Press, 1969.

KEYFITZ, N. Applied Mathematical Demography. New York, John Wiley & Sons, 1977.

KEYFITZ, N. e J.A.BEEKMAN, Demography Through Problems. New York, Springer-Verlag, 1984.

LE BRAS, H. Nouvelles tables-types de mortalité. Population, Paris, v.23, n.4, p.739-744, 1968.

Élements pour une théorie des populations instables. Population, Paris, v.26, n.3, p.525-572, 1971.

LEDERMANN, S. Nouvelles Tables-Types de Mortalité. Paris, Presses Universitaires de France, 1969.

LEDERMANN, S. e J.BRÉAS, Les dimensions de la mortalité. Population, Paris, v.14, n.4, p.637-682, 1959. Disponível em: http://www.persee.fr/web/revues/home/prescript/issue/ pop_0032-4663_1959_num_14_4

MADEIRA, J.L. Aplicação de um modelo teórico na reconstituição da demografia brasileira. Revista Brasileira de Estatística, Rio de Janeiro, v.27, n.106, p.86-92, 1966. Disponível em: http://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/RBE/RBE%201966%20v27%20n106.pdf

MORTARA, G. Estudos sobre a utilização do censo demográfico para a reconstrução das estatísticas do movimento da população do Brasil. II. Conjecturas sobre os níveis da natalidade e da mortalidade no Brasil no período 1870-1920, Revista Brasileira de Estatística, Rio de Janeiro, v.I, n.2, p.229-242, 1940a. Disponível em: http://biblioteca.ibge.gov.br/ (Coleção digital)

Estudos sobre a utilização do censo demográfico para a reconstrução das estatísticas do movimento da população do Brasil. IV. Ensaio de ajustamento das tábuas de mortalidade brasileiras calculadas por Bulhões de Carvalho. Revista Brasileira de Estatística, Rio de Janeiro, v.I, n.4, p.673-693, 1940b. Disponível em: http://biblioteca.ibge.gov.br/ (Coleção digital)

Estudos sobre a utilização do censo demográfico para a reconstrução das estatísticas do movimento da população do Brasil. V. Retificação da distribuição por idade da população natural do Brasil, contante dos censos, e cálculo dos óbitos, dos nascimentos e das variações dessa população no período 1870-1920. Revista Brasileira de Estatística, Rio de Janeiro, v.II, n.5, p.39-89, 1941a. Disponível em: http://biblioteca.ibge.gov.br/ (Coleção digital).

Estudos sobre a utilização do censo demográfico para a reconstrução das estatísticas do movimento da população do Brasil. VII. Tábuas de mortalidade e de sobrevivência para os períodos 1870-1890 e 1890-1920. Cálculo, exame e comparações internacionais. Revista Brasileira de Estatística, Rio de Janeiro, v.II, n.7, p.493-538, 1941b. Disponível em: http://biblioteca.ibge.gov.br/ (Coleção digital).

Tábuas de sobrevivência ajustadas para o Distrito Federal, de 1939-41, retificadas conforme a correção da mortalidade no primeiro ano de idade. Revista Brasileira de Estatística, Rio de Janeiro, v.VI, n.24, p.659-667, 1945. Disponível em: http://biblioteca.ibge.gov.br/ (Coleção digital)

Tábuas brasileiras de mortalidade e sobrevivência. Rio de Janeiro, Fundação Getúlio Vargas, 1946.

Tábuas de sobrevivência para o Distrito Federal, segundo a mortalidade do triênio 1949-51. Revista Brasileira de Estatística, Rio de Janeiro, v.13, n.51, p.306-317, 1952. Disponível em: http://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/RBE/RBE%201952%20v13%20n51.pdf

A mortalidade da população natural do Brasil. Revista Brasileira de Estatística, Rio de Janeiro, v.14, n.56, p.313-324, 1953. Disponível em: http://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/RBE /RBE%201953%20v14%20n56.pdf

The Development and Structure of Brazil's Population. Population Studies, Londres, v.8, n.2, p.121-139, 1954. Disponível em: http://www.jstor.org.

A fecundidade das mulheres e a sobrevivência dos filhos o Brasil, segundo o censo de 1950." Revista Brasileira de Estatística, Rio de Janeiro, v.17, n.67, p.177-186, 1956. Disponível em : http://biblioteca.ibge.gov.br/visualizacao/monografias/GEBIS%20-%20RJ/RBE/RBE%201956%20v17%20n67.pdf

Incógnitas da mortalidade no Brasil. Revista Brasileira de Estatística, Rio de Janeiro, v.18, n.69-70, p.1-17, 1957. Disponível em: http://biblioteca.ibge.gov.br/visualizacao/monografias /GEBIS%20-%20RJ/RBE/RBE%201957%20v18%20n69_70.pdf.

SHRYOCK, H.S. e J.S.SIEGEL, The Methods and Materials of demography. Condensed Edition by Edward G. Stockwella, New York, Academic Press, 1976.

SMITH, D. e N.KEYFITZ (Eds.) Mathematical Demography. Biomathematics. Berlin, Springer-Verlag, 1977.

UNITED NATIONS Manual I: Methods of Estimating Total Population for Current Dates. New York, 1952. Disponível em: http://www.un.org/esa/population/techcoop/DemEst/manual1/ manual1.html

Age and Sex Pattern of Mortality. New York, 1955. Disponível em: http://www.un.org /esa/population /techcoop/DemEst/manual1/manual1.html

Manual III: Manual for Population Projections by Sex and Age. New York, 1956. Disponível em: http://www.un.org/esa/population/techcoop/PopProj/manual3/manual3.html

Manual IV: Methods of Estimating Basic Demographic Measures from Incomplete Data. New York, 1967. Disponível em: http://www.un.org/esa/population/techcoop/DemEst/manual4/ manual4.html

The Concept of Stable Population. Application to the Study of Populations of Countries with Incomplete Demographic Statistics. New York, 1968. Disponível em: http://www.un.org/esa/population/techcoop/DemMod/concept_stablepop/concept_stablepop.html

Manual X: Indirect Techniques for Demographic Estimation. New York, United Nations, 1983. Disponível em: http://www.un.org/esa/population/publications/Manual_X/Manual_X.htm

UTTERSTRÖM, G. Two Essays on Population in Eighteenth-Century Scandinavia. In D.V. GLASS e D.E.C. EVERSLEY Population in History. London, Edward Arnold, 1965.

Agradecimentos

Dedico este trabalho à memória dos professores Oscar Protocarrero e Rio Nogueira, os primeiros a me incentivar a pensar sobre o que estaria atrás da matemática.

Abstract

Mortality tables are fundamental instruments in representing demographic mechanisms. We propose a typology of these tables based on their construction criteria. A first distinction is made between "empirical" tables, which refer to existing or past populations, and "mathematical" tables, modeled on empirical characteristics, but parametrized as function of demographic variables, such as the natural growth of the population or its life expectancy at birth. Demographic studies increasingly use combinations of both these types, benefiting from the regularities proper to theoretical tables to complete information unknown in the data under analysis. The mathematical tables are divided into compositions of empirical tables (model tables), tables derived from the theory of stable or quasi-stable populations, or applications of both these forms. In conclusion, summary comments are made on some tables built to represent Brazilian mortality from 1870 to 1950.

Key-words: demographic history; mortality; mortality table.

VaR, Teste de Estresse e MaxLoss na Presença de Heteroscedasticidade e Longa Dependência na Volatilidade

Taiane S. Prass[1]
Sílvia R.C. Lopes[2]

Resumo

O objetivo deste trabalho é analisar o desempenho dos principais métodos utilizados no cálculo de riscos em um portfólio, sendo eles o teste de estresse *(stress test),* a perda máxima *(Maximum Loss* ou *MaxLoss*) e o valor em risco (*Value-at-Risk* ou VaR), sob a presença de heteroscedasticidade e longa dependência na volatilidade. Consideramos um portfólio formado por ações de quatro empresas e utilizamos o modelo CAPM para estimar os pesos dos ativos no portfólio. Para o cálculo do VaR consideramos a abordagem econométrica onde a variância condicional é modelada através de modelos FIEGARCH. Para esta análise consideramos dados observados entre janeiro de 1995 e dezembro de 2001, período em que a volatilidade apresenta longa dependência e estacionariedade.

MSC (2000). 60G10, 62M10, 62M20, 97M30, 91B84.

Palavras-chave. Longa Dependência; Volatilidade; Medidas de Risco; Processos FIEGARCH; Modelo CAPM; Teste de Estresse; Perda Máxima.

[1] Endereço para correspondência. E-mail: taianeprass@gmail.com
[2] E-mail: silvia.lopes@ufrgs.br
Instituto de Matemática Universidade Federal do Rio Grande do Sul – Porto Alegre – RS

1. Introdução

Em termos financeiros, *risco* é a possibilidade de que um investimento tenha um retorno diferente do esperado, incluindo a possibilidade de perda de uma parte ou até mesmo de todo o investimento original. Formalmente, dado um espaço de probabilidade $(\Omega, \mathcal{F}, P)$ e o conjunto $L^0(\Omega, \mathcal{F}, P)$ de todas as variáveis aleatórias sobre $(\Omega, \mathcal{F})$ que são finitas em quase toda parte, riscos financeiros são representados por um conjunto $\mathcal{M} \subset L^0(\Omega, \mathcal{F}, P)$ de variáveis aleatórias que interpretamos como perdas do portfólio sobre o horizonte de tempo h. Nesse contexto, uma *medida de risco* com domínio $\mathcal{M}$ é uma função real $\rho: \mathcal{M} \to R$.

Medidas de risco são utilizadas, por exemplo, na determinação do capital em risco ou ainda, quando deseja-se limitar os riscos aos quais a instituição financeira está exposta (McNeil et al., 2005). A medida mais utilizada (mas não a única) para medir o risco ou a *volatilidade* de um determinado investimento é a variância (variância condicional ou desvio padrão) da rentabilidade histórica. Em termos estatísticos, a volatilidade pode ser vista como a possibilidade do valor de um ativo cair ou subir muitas vezes, de forma significativa, em um determinado período de tempo. Quando definimos a volatilidade como sendo a variância da rentabilidade histórica de um investimento, obtemos uma medida de volatilidade absoluta. Essa medida varia com o período de tempo determinado, de forma que na hora do cálculo, o período escolhido é de vital importância. Além de medidas de volatilidade absoluta, existem também formas de avaliar a volatilidade de forma relativa, ou seja, em relação à volatilidade do mercado, por exemplo. A medida mais usada é o coeficiente β do modelo CAPM, que mede a volatilidade de um ativo frente a um índice de mercado.

Os métodos mais utilizados no mercado para calcular medidas de risco baseiam-se na suposição de distribuição normal para os processos de inovação. Um dos problemas apresentados neste método é que essa função de distribuição possui caudas mais leves que as observadas em séries temporais financeiras, subestimando assim as perdas. Dentre as medidas de risco mais utilizadas estão o *valor em risco* (*Value-at-Risk* ou VaR) e o *valor em risco condicional* (CVaR ou *Expected Shortfall*). Até pouco tempo atrás, os

modelos para VaR e CVaR assumiam que os retornos de ações eram normalmente distribuídos. Muitas distribuições alternativas vêm sendo propostas na literatura. Mesmo utilizando distribuições de caudas mais pesadas para modelar os retornos financeiros, sabe-se que as medidas de risco ficam subestimadas para probabilidades de ocorrência pequenas (eventos extremos).

Outra ferramenta amplamente utilizada para medir os riscos de um portfólio são os denominados *testes de estresse* (*stress test*). Os testes de estresse não são influenciados por caudas pesadas pois não baseiam-se em hipóteses sobre a função de distribuição das mudanças de fatores de risco. Como tais testes não quantificam a probabilidade de ocorrência de cenários individuais, eles são utilizados como ferramenta auxiliar para verificar e complementar as estatísticas do tipo medidas de risco como, por exemplo, o VaR. Mesmo assim, os cenários precisam ser minimamente plausíveis e para isso é necessária uma idéia, ainda que vaga, das probabilidades de ocorrência de cada cenário.

A medida *perda máxima* (*Maximum Los*s ou *MaxLoss*), introduzida por Studer (1997) pode ser vista como uma maneira sistemática de realizar um teste de estresse. Esta é uma medida de risco que pode ser interpretada como a pior perda possível que pode ocorrer em um portfólio em um determinado instante de tempo. Entretanto, em alguns casos, o pior cenário pode não existir, dado que a função para calcular o valor do portfólio pode ser não-limitada inferiormente. Como a probabilidade de ocorrência de um cenário no qual o mercado a muito tempo não se encontra é muito baixa, para o cálculo dessa medida restringe-se a atenção à cenários em um certo domínio de admissibilidade, também denominado região de confiança.

Neste trabalho comparamos o desempenho da medida valor em risco, do teste de estresse e da medida perda máxima na análise de um portfólio sob a presença de heteroscedasticidade e longa dependência na volatilidade. Para o cálculo do VaR consideramos a abordagem econométrica onde a variância condicional é modelada através de modelos FIEGARCH. Na Seção 2 relembramos o método da média-variância para seleção de portfólios eficientes (Markowitz, 1952) e o modelo de precificação de ativos (CAPM), utilizado como ferramenta para a análise da correlação dos ativos com o mercado financeiro. Além disso, nessa seção apresentamos as principais medidas de risco utilizadas na literatura e a definição de processo FIEGARCH, utilizados na abordagem econométria para a modelagem da variância condicional da série

temporal. A Seção 3 é dedicada à análise de dados reais utilizando as ferramentas descritas na Seção 2. Na Seção 4 apresentamos as conclusões obtidas a partir da análise dos dados.

2. Medidas de Risco

Segundo McNeil et al. (2005), existem várias abordagens para o cálculo de medidas de risco, mas de modo geral, podemos classificá-las em quatro diferentes categorias. Neste trabalho consideramos duas dessas categorias. A primeira delas são as *medidas de risco baseadas em distribuição de perdas*. Tais medidas de risco são estatísticas que descrevem a distribuição condicional e não-condicional das perdas do portfólio sobre um horizonte de tempo pré-determinado, como por exemplo a variância, o Valor em Risco (VaR) e o Valor em Risco Condicional (CVaR ou *Expected Shortfall*). A segunda abordagem são as *medidas de risco baseadas em cenários*. Nesta abordagem, para calcular as medidas de risco em um portfólio, consideramos um número de possíveis mudanças de fatores de risco (cenários), tais como 10% de variação nas taxas de câmbio ou queda de 20% no preço das ações. O risco do portfólio é então medido como a perda máxima do portfólio sobre todos os possíveis cenários.

Lembramos que, dado o portfólio $\mathcal{P} = \{A_1, \cdots, A_N\}$, dizemos que $\mathcal{P}$ é *eficiente* se ele contém uma combinação de ativos que proporciona o máximo retorno para uma dada classe de risco, ou de forma equivalente, o mínimo risco para uma dada classe de retornos. Sendo assim, a seleção dos pesos $\mathbf{a} = (a_1, \cdots, a_N)'$, para os ativos em $\mathcal{P}$, é de fundamental importância. Portanto, antes de discutirmos o cálculo das medidas de risco, relembramos o método da média-variância para seleção de portfólios eficientes (Markowitz, 1952).

2.1. Cálculo do Portfólio Eficiente e o Modelo CAPM

A maioria dos investidores espera que um portfólio eficiente satisfaça ambas as condições, máximo retorno e mínima variância. Segundo Markowitz (1952), o portfólio com máximo valor esperado não é, necessariamente, o de mínima variância. Existe uma proporção tal que o investidor pode aumentar os retornos esperados aumentando a

variância ou reduzir a variância diminuindo os retornos esperados. Em seu trabalho, o autor apresenta um estudo sobre a seleção de portfólios eficientes baseado no valor esperado e na variância dos retornos (E - V *rule*). De acordo com os conceitos desenvolvidos, o investidor pode reduzir o risco total de seus investimentos através da combinação de ativos que não apresentam correlação positiva e a redução dos riscos será tão maior quanto menor for esta correlação.

Sharpe (1964) e Lintner (1965) demonstram que o conjunto de portfólios eficientes disponíveis para o investidor que emprega a análise de média-variância, na ausência de um ativo *livre de risco*, isto é, um ativo não-correlacionado com o mercado e com retorno esperado igual ao retorno observado, é inferior ao disponível quando existe ativo desse tipo. Segundo Fabozzi et al. (2006) os pesos ótimos $\mathbf{a} = (a_1, \cdots, a_N)'$, para o portfólio eficiente $\mathcal{P} = \{A_1, \cdots, A_N, F\}$, na presença de um ativo livre de risco F, são dados por

$$\mathbf{a} = C\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R_F 1), \quad com \quad C = \frac{\mu_0 - R_F}{(\boldsymbol{\mu} - R_F 1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R_F 1)'} \qquad (1)$$

onde μ_0 é o nível de retorno esperado, $\boldsymbol{\Sigma}$ é a matriz de variância-covariância dos retornos, $\boldsymbol{\mu} = (E(R_1), \cdots, E(R_N))'$, R_F é o retorno ativo livre de riscos e $\mathbf{1}$ é o vetor coluna com N entradas iguais a 1. Obviamente, o peso para o ativo livre de risco é $a_F = 1 - \sum_{i=1}^{N} a_i$.

A expressão (1) mostra que todos os pesos para ativos de risco para um portfólio de mínima variância são proporcionais ao vetor $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - R_F 1)$, com constante de proporção C. Mais especificamente (veja Fabozzi et al., 2006), quando introduzimos um ativo livre de risco no portfólio, todos os portfólios de mínima variância são combinações lineares de um ativo livre de risco com um portfólio de risco. Este cenário é representado por uma reta denominada *Linha de Mercado de Capitais* (CML) e corresponde a situação do mercado em equilíbrio.

O modelo de precificação de ativos de capital (*Capital Asset Pricing Model* ou CAPM), desenvolvido por Sharpe (1964) e Lintner (1965), mostra como o retorno esperado de um ativo está relacionado ao seu risco sistemático. Tal modelo baseia-se nas seguintes hipóteses:

• inexistência de custos de transação (mercado perfeito);

• ausência de informação confidencial, o que faz com que não hajam ativos subavaliados ou superavaliados no mercado;

• a decisão do investidor baseia-se unicamente no retorno esperado e no risco; os investidores estimam o risco em função da variabilidade das taxas de retorno estimadas;

• existência de um ativo livre de risco, denotado por F, onde os investidores podem emprestar e tomar emprestado a uma única taxa, denotada por R_F, visando obter alocações ótimas;

• os investidores ajustam a decisão de alocação às preferências de risco decidindo dessa forma, quando investirão em ativos livres de risco ou ativos arriscados.

Se o mercado está em equilíbrio, situação representada pela linha de capital de mercado (CML), o retorno esperado de um portfólio é uma função linear do valor esperado do portfólio do mercado e o retorno esperado de ativos individuais deve estar sobre a *Linha de Mercado de Títulos (Security Market Line* ou SML) e não sobre a linha CML. A SML é definida por

$$E(R_i) = R_F + \left[\frac{E(R_M)-R_F}{\mathrm{Var}(R_M)}\right]\mathrm{Cov}(R_i,R_M) = R_F + \beta_i(E(R_M) - R_F), \qquad (2)$$

para todo $i \in \{1,\cdots,N\}$, onde $\beta_i = \frac{\mathrm{Cov}(R_i,R_M)}{\mathrm{Var}(R_M)}$.

O modelo (2) é também denominado CAPM e sua versão empírica é a regressão linear, denominada *linha característica*, dada por

$$R_{i,t} - R_{F,t} = \alpha_i + \beta_i(R_{M,t} - R_{F,t}) + \varepsilon_{i,t}, \quad i \in \{1,\cdots,N\} \ e \ t \in \{1,\cdots,n\}, \qquad (3)$$

onde n é o tamanho amostral, $\varepsilon_{i,t}$ é o erro e representa o risco não-sistemático, ou seja, a parcela do risco que pode ser eliminada através da diversificação (compra ou venda de ações), β_i representa o risco sistemático, isto é, a parcela do risco que não pode ser eliminada pela diversificação e está relacionada com o comportamento do mercado como um todo e α_i é o coeficiente α da ação, definido por $\alpha_i = \bar{R}_i - E(R_i)$, e determina se o ativo está subavaliado ($\alpha_i > 0$) ou superavaliado ($\alpha_i < 0$).

Observação 2.1 - O coeficiente β *do portfólio* $\mathcal{P} = \{A_1, \cdots, A_N\}$, *denotado por* $\beta_{\mathcal{P}}$, é dado por $\beta_{\mathcal{P}} = \sum_{i=1}^{N} \alpha_i \beta_i$, onde β_i, *para* $i \in \{1, \cdots, N\}$, é o coeficiente β do ativo A_i.

O coeficiente angular β_i, dado na expressão (2), pode ser visto como uma medida de volatilidade das taxas de retorno de um ativo qualquer com relação às taxas de retorno do mercado como um todo. Esse pressuposto parte do princípio de que todos os ativos tendem a ter os seus preços alterados com maior ou menor proporção às alterações do mercado como um todo. Lembramos que o retorno esperado depende apenas do risco sistemático. Como ativos com valores de β maiores têm riscos sistemáticos mais altos, têm também retornos esperados maiores. Dessa forma, conhecendo-se as características de risco (seu respectivo β_i) de uma ação, é possível estimar-se o preço justo (ou valor intrínseco), tendo-se a indicação se o ativo é, ou não, uma boa opção de compra.

2.2. Medidas de Risco Baseadas na Distribuição de Perdas e Ganhos

Dado um portfólio fixo $\mathcal{P} = \{A_1, \cdots, A_N\}$, denotamos por V_t o *valor do portfólio* no tempo $t \in N$. Assumimos que V_t, para todo $t \in N$, são variáveis aleatórias observáveis e denotamos por $L_{t+h} := -(V_{t+h} - V_t)$ a *perda do portfólio* $\mathcal{P}$ sobre o período h. A distribuição da variável aleatória L_{t+h} é denominada *distribuição de perdas*. Em geral, a variável aleatória V_t é definida como uma função de um vetor aleatório observável $\mathbf{Z}_t = (Z_{1,t}, \cdots, Z_{m,t})'$, de *fatores de risco*, ou seja, $V_t = f(t, \mathbf{Z}_t)$, para todo $t \in N$, para alguma função mensurável $f: R \times R^m \to R$, denominada *função dos riscos*.

Observação 2.2 - Em alguns casos, é conveniente utilizar, ao invés dos fatores de risco em si, a série de mudança de fatores de risco $\{\boldsymbol{X}_t\}_{t \in N}$, definida *por* $\boldsymbol{X}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}$, para todo $t \in N$. Nesse caso, a perda do portfólio pode ser reescrita como $L_{t+1} = -(f(t+1, \boldsymbol{Z}_t + \boldsymbol{X}_{t+1}) - f(t, \boldsymbol{Z}_t))$, para todo $t \in N$. *Como* $\boldsymbol{Z}_t$ é observável no tempo t, a distribuição das perdas é determinada pela distribuição da mudança de fator de risco $\boldsymbol{X}_{t+1}$.

A escolha dos fatores de risco e da função $f(\cdot)$ depende tanto do portfólio quanto do nível de precisão desejados. Por exemplo, se P_t ó o preço de um ativo qualquer no instante t, a variação relativa de preços ou retorno líquido simples (ou ainda taxa de retorno) é $R_t = (P_t - P_{t-1})(P_{t-1})^{-1}$ e o retorno composto continuamente (ou simplesmente log-retorno) é dado por $r_t = \ln(P_t) - \ln(P_{t-1})$. Dessa forma, para um

portfólio $\mathcal{P} = \{A_1, \cdots, A_N\}$, com pesos $\mathbf{a} = (a_1, \cdots, a_N)'$, os fatores de risco são os logaritmos dos preços e a série de mudança de fatores de risco são os log-retornos e ainda, $L_{t+1} = -V_t \sum_{i=1}^{N} a_i R_{i,t+1}$, onde $R_{i,t}$ é o retorno do ativo A_i, no tempo $t+1$.

Observação 2.3 - Na prática R_t *e* r_t são valores muito próximos entre si e podemos assumir que $L_{t+1} = -V_t \sum_{i=1}^{N} a_i r_{i,t+1} = -V_t r_{\mathcal{P},t}$. Além disso, como a distribuição das perdas fica totalmente determinada pela distribuição das mudanças de fator de risco, sem perda de generalidade, podemos supor $V_t = 1$ e considerar apenas a série temporal $\{-r_{\mathcal{P},t}\}_{t \in N}$ ou, equivalentemente, $\{r_{\mathcal{P},t}\}_{t \in N}$.

Uma das medidas de risco mais utilizadas é a *variância* da distribuição de perdas e ganhos. Este fato deve-se ao grande impacto que a teoria de portfólios de Markowitz, que utiliza variância como uma medida de risco, apresenta no estudo teórico e prático da área de finanças. Entretanto, como uma medida de risco, a variância apresenta dois problemas. O primeiro deles é que necessitamos assumir que a função de distribuição de perdas possui segundo momento finito. Além disso, essa medida não faz distinção entre desvios positivos e negativos da média. Segundo McNeil et al. (2005), a variância é uma boa medida de risco somente para distribuições que são (aproximadamente) simétricas, tais como a normal ou a t-Student (com variância finita).

Outra abordagem amplamente utilizada atualmente é a análise dos quantis da distribuição de perdas. Considere um portfólio $\mathcal{P} = \{A_1, \cdots, A_N\}$ de ativos de risco e um horizonte h fixos. Seja $F_L(\ell) = P(L \leq \ell)$ a função de distribuição de perdas correspondente. Nosso objetivo é definir uma estatística baseada em $F_L(\cdot)$ que seja capaz de medir os riscos do portfólio sobre o período h. Um candidato natural é a perda máxima possível dada por $\inf\{\ell \in R: F_L(\ell) = 1\}$. Entretanto, em muitos modelos o suporte de $F_L(\cdot)$ é não-limitado. Logo, a perda máxima é infinito. A idéia então, é repassar "perda máxima" por "perda máxima que não é ultrapassada com uma certa probabilidade", com esta probabilidade denominada *nível de confiança*.

Definição 2.1 - Seja $\mathcal{P} = \{A_1, \cdots, A_N\}$ um portfólio fixo. Dado um nível de confiança $p \in (0,1)$, o *valor em risco (VaR)* do portfólio é definido como

$$\mathrm{VaR}_{\mathrm{p},\mathcal{P}} := \inf\{\ell \in R : P(L \geq \ell) \leq 1 - \mathrm{p}\} = \inf\{\ell \in R : F_L(\ell) \geq \mathrm{p}\}.$$

Em termos probabilísticos, VaR_p é o p-quantil da função de distribuição de perdas. Na prática, os valores freqüentemente utilizados para p são 0,95 ou 0,99 e para h são 1 ou 10 dias. Assim como na análise da medida de risco pela variância, o VaR_p apresenta certas desvantagens. Artzner et al. (1999) mostram que o VaR_p não é uma medida de risco coerente pois não satisfaz o axioma da subaditividade. Isto é, o VaR_p de portfólios somados pode não ser limitado pela soma do VaR_p dos portfólios individuais. Isso contradiz a idéia de que os riscos podem ser diminuídos através da diversificação, ou seja, através da compra ou venda de ativos. A medida de risco *expected shortfall*, também referenciada na literatura como *valor em risco condicional* é uma medida de risco coerente e está diretamente relacionada ao VaR_p.

Definição 2.2 - Considere uma perda L com função de distribuição $F_L(\cdot)$, tal que $E(|L|) < \infty$. *O expected shortfall (ES) ao nível de confiança* $p \in (0,1)$, é definido como $ES_p := \frac{1}{1-p}\int_p^1 q_u(F_L)du$, *onde* $q_u(\cdot)$ *é a função quantil definida por* $q_u(F_L) = inf\{\ell \in R : F_L(\ell) \geq u\}$.

As medidas de risco *Expected Shortfall* e VaR estão relacionadas através da expressão $\mathrm{ES}_\mathrm{p} = \frac{1}{1-\mathrm{p}}\int_\mathrm{p}^1 \mathrm{VaR}_u du$. Além disso, mostra-se (veja McNeil et al., 2005) que se L for integrável com função de distribuição $F_L(\cdot)$ contínua então $\mathrm{ES}_\mathrm{p} = E(L|L \geq \mathrm{VaR}_\mathrm{p})$. Se L possui função de distribuição normal com média μ e variância σ^2 então, para todo $\mathrm{p} \in (0,1)$, $\mathrm{ES}_\mathrm{p} = \mu + \sigma[\phi(\Phi^{-1}(\mathrm{p}))](1-\mathrm{p})^{-1}$, onde $\phi(\cdot)$ e $\Phi(\cdot)$ são a função densidade de probabilidade e a função de distribuição normal padrão, respectivamente. Além disso, como a função *Expected Shortfall* (ES) é uma medida de risco coerente, segue que, dado um portfólio $\mathcal{P} = \{A_1, \cdots, A_N\}$, formado por N ativos, com pesos $\mathbf{a} = (a_1, \cdots, a_N)'$, a seguinte desigualdade é sempre válida,

$$\mathrm{ES}_{\mathrm{p},\mathcal{P},t+1} \leq \sum_{i=1}^{N} a_i \mathrm{ES}_{\mathrm{p},i,t+1}, \quad para\ todo \quad t \in Z,$$

onde $\mathrm{ES}_{\mathrm{p},\mathcal{P},t+1}$ e $\mathrm{ES}_{\mathrm{p},i,t+1}$ denotam, respectivamente, o valor da medida *Expected Shortfall* para o portfólio $\mathcal{P}$ e para o ativo A_i, para qualquer $i \in \{1, \cdots, N\}$, no tempo $t+1$.

É imediato que diferentes formas de estimar $F_L(\cdot)$ fornecerão diferentes valores para o VaR e o ES. A abordagem econométrica baseia-se no fato que séries de retornos financeiros (ou log-retornos) raramente apresentam tendência ou sazonalidades (com exceção eventualmente de retornos intra-diários) e possuem ainda algumas características que não são comuns a outras séries temporais tais como: os retornos (ou log-retornos) são, em geral, não-autocorrelacionados; os quadrados dos retornos são autocorrelacionados, apresentando uma correlação para $h = 1$ pequena e depois uma queda lenta para os demais valores de h; séries de retornos apresentam agrupamentos (*clusters*) de volatilidades ao longo do tempo; a função de distribuição (não-condicional) dos retornos apresenta caudas mais pesadas do que uma função de distribuição normal. Além disso, a distribuição é em geral leptocúrtica, embora aproximadamente simétrica; algumas séries de retornos apresentam a característica de não-linearidade.

Para modelar séries temporais na presença de grupos (*clusters*) de volatilidade é necessário recorrer a modelos heteroscedásticos condicionais. Estes modelos consideram a variância de um retorno em um dado instante de tempo como uma função que depende de retornos passados e de outras informações disponíveis até aquele momento, obtendo assim, uma *variância condicional*. Essa variância condicional é diferente da variância global (não-condicional) da série observada pois varia com o tempo. Dentre os modelos não-lineares mais conhecidos destacamos os modelos ARCH (Engle, 1982), GARCH (Bollerslev, 1986), EGARCH (Nelson, 1991), FIGARCH (Baillie et al., 1996) e FIEGARCH (Bollerslev e Mikkelsen, 1996). Entre esses, o FIEGARCH é o modelo mais conveniente pois leva em conta a volatilidade variando com o tempo e a presença de *clusters* de volatilidade (assim como os modelos ARCH e GARCH), a assimetria dos retornos (assim como o modelo EGARCH) e ainda, a longa dependência

da volatilidade (assim como o modelo FIGARCH). A vantagem dos processos FIEGARCH em relação aos processos FIGARCH é a sua estacionariedade fraca.

Definição 2.3 - *Seja* $\{X_t\}_{t\in Z}$ um processo estocástico. Dizemos que $\{X_t\}_{t\in Z}$ é um processo EGARCH *Fracionariamente Integrado*, e denotamos por FIEGARCH(p,d,q), se

$$X_t = \sigma_t Z_t, \qquad (4)$$

$$\ln(\sigma_t^2) = \omega + \frac{\alpha(\mathcal{B})}{\beta(\mathcal{B})}(1-\mathcal{B})^{-d} g(Z_{t-1}), \quad para\ todo\ \ t \in Z, \qquad (5)$$

onde $\omega \in R$, $\{Z_t\}_{t\in Z}$ é uma sequência de variáveis aleatórias i.i.d., com média zero e variância um, $\alpha(\cdot)$ e $\beta(\cdot)$ são polinômios definidos por $\alpha(z) := \sum_{i=0}^{p}(-\alpha_i)z^i$ e $\beta(z) := \sum_{j=0}^{q}(-\beta_j)z^j$, com $\alpha_0 = -1 = \beta_0$ e $\beta(z) \neq 0$ no disco fechado $\{z: |z| \leq 1\}$. O operador $(1-\mathcal{B})^d$ é definido em termos de sua expansão através de séries de Maclaurin, como $(1-\mathcal{B})^d = \sum_{k=0}^{\infty}\frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}\mathcal{B}^k := \sum_{k=0}^{\infty}\delta_{d,k}\mathcal{B}^k$ e $g(Z_t) = \theta Z_t + \gamma[|Z_t| - E(|Z_t|)]$, para todo $t \in Z$, com $\theta, \gamma \in R$.

Uma descrição mais detalhada dos processos FIEGARCH pode ser encontrada em *Lopes e Prass* (2012). Nesse trabalho os autores apresentam as condições necessárias e suficientes para estacionariedade, ergodicidade e invertibilidade dos processos FIEGARCH. Utilizando o fato que $\{g(Z_t)\}_{t\in Z}$ é um processo ruído branco e que $\{\ln(\sigma_t^2)\}_{t\in Z}$ é um processo ARFIMA(q,d,p), os autores mostram que, sob certas condições, $\{\ln(X_t^2)\}_{t\in Z}$ é um ARFIMA$(q,d,0)$. Lopes e Prass (2012) também apresentam as expressões para as medidas de assimetria e curtose do processo $\{X_t\}_{t\in Z}$, para as funções de autocorrelação e densidade espectral dos processos $\{\ln(\sigma_t^2)\}_{t\in Z}$ e $\{\ln(X_t^2)\}_{t\in Z}$ e discutem suas propriedades assintóticas. Resultados sobre estimação e previsão em processos FIEGARCH são também abordados pelos autores e uma fórmula de recorrência para o cálculo dos coeficientes na representação por séries do operador dado em (5) é apresentada e suas propriedades assintóticas são estudadas.

Na literatura encontramos diferentes maneiras de definir um processo FIEGARCH. Na próxima proposição mostramos que, sob certas restrições, a expressão (5) pode ser reescrita como na equação (6), apresentada em Zivot e Wang (2005) e utilizada pelo *software* S-Plus.

Proposição 2.1 - Seja $\{X_t\}_{t\in Z}$ um processo FIEGARCH(p,d,q), dado na Definição *2.3*. Se $d > 0$, *a* expressão de $ln(\sigma_t^2)$, dada em *(5)*, pode ser reescrita como

$$\beta(B)(1-B)^d \ln(\sigma_t^2) = a + \sum_{i=0}^{p} (\psi_i |Z_{t-1-i}| + \gamma_i Z_{t-1-i}), \qquad (6)$$

$$onde \quad a = -\gamma\alpha(1)E(|Z_0|),\ \psi_i = -\gamma\alpha_i \quad e \quad \gamma_i = -\theta\alpha_i,\ para\ todo \quad 0 \le i \le p. \quad (7)$$

Prova: Veja Prass (2008).

Observação 2.4 - É fácil ver que a definição dada por Zivot e Wang (2005) é mais geral que a apresentada por Bollerslev e Mikkelsen (1996) já que os coeficientes ψ_i *e* γ_i, para $i \in \{0,\cdots,p\}$ não precisam, necessariamente, satisfazer as condições da expressão *(7)*.

Em alguns casos as séries temporais dos retornos apresentam autocorrelação significante entre as variáveis aleatórias. Quando isso ocorre, a correlação entre os retornos é modelada com o auxilio de um modelo linear, em geral, um ARMA. Para um modelo ARMA(p_1,q_1)-FIEGARCH(p_2,d,q_2) temos

$$r_t = \phi_0 + \sum_{i=1}^{p_1} \phi_i r_{t-i} + X_t - \sum_{k=1}^{q_1} \theta_k X_{t-k} = \mu_t + X_t$$

e $X_t := \sigma_t Z_t$ é um FIEGARCH(p_2,d,q_2) definido pelas expressões (4) e (5).

Supondo que $Z_t \sim \mathcal{N}(0,1)$, segue que $r_t|\mathcal{F}_{t-1} \sim \mathcal{N}(\mu_t,\sigma_t^2)$. Como $L_t = -V_{t-1}r_t$, assumindo $V_t = 1$, segue que

$$\mathrm{VaR}_{p,t+1} = -\mu_{t+1} + \Phi^{-1}(p)\sigma_{t+1}.$$

Para o cálculo do VaR_p para um horizonte h, note que, assim como para $h = 1$, podemos utilizar a aproximação $L_{t+h} \approx -V_t r_t[h]$. Portanto, precisamos estimar a média $E(r_t[h]|\mathcal{F}_t)$ e a variância condicional $\mathrm{Var}(r_t[h]|\mathcal{F}_t)$ do retorno h dias à frente, dada a informação até o instante t.

Proposição 2.2 - Seja $\{r_t\}_{t\in Z}$ um processo ARMA(p_1, q_1)-FIEGARCH(p_2, d, q_2). Então,

i. A previsão da média do retorno de período h é dada por $\hat{r}_t[h] = \sum_{j=1}^{h} \hat{r}_{t+j}$, onde $\hat{r}_{t+j}$ é a previsão j passos à frente para o modelo ARMA.
ii. O erro de previsão é dado por $e_t[h] = \sum_{j=0}^{h-1} \left[(\sum_{i=0}^{j} \psi_i) X_{t+h-j}\right]$, onde ψ_i, para todo $i > 0$, são os coeficientes da representação $MA(\infty)$ do processo estocástico $\{r_t\}_{t\in Z}$.
iii.
iv.
v.
vi. A previsão da volatilidade do retorno de período h é dada pela expressão

$$\hat{\sigma}_t^2[h] = \sum_{j=0}^{h-1} \left[\left(\sum_{i=0}^{j} \psi_i \right)^2 \hat{\sigma}_{t+h-j}^2 \right],$$

onde ψ_i, *para todo* $i > 0$, são os coeficientes da representação $MA(\infty)$ do processo estocástico $\{r_t\}_{t\in Z}$ *e* $\hat{\sigma}_{t+j}^2$ é a previsão j passos à frente para a variânica condicional descrita pelas equações do modelo FIEGARCH.

Prova: Para provar (i), note que $r_t[h] = \sum_{j=1}^{h} r_{t+j}$. Portanto,

$$\hat{r}_t[h] = E(r_t[h]|\mathcal{F}_{t-1}) = \sum_{j=1}^{h} E(r_{t+j}|\mathcal{F}_{t-1}) = \sum_{j=1}^{h} \hat{r}_{t+j}$$

e, pelas propriedades do modelo ARMA

$$\hat{r}_{t+j} = \phi_0 + \sum_{i=1}^{p_1} \phi_i \hat{r}_{t+j-i} - \sum_{j=0}^{q_1} \theta_j \hat{X}_{t+j-j},$$

com $\hat{r}_{t+j} = r_{t+j}$, se $j \leq 0$, $\hat{X}_{t+j} = X_{t+j}$, se $j \leq 0$ e $\hat{X}_{t+j} = 0$, se $j > 0$, para todo $1 \leq j \leq h$.

Para provar (ii) note que, utilizando-se a representação MA(∞) do modelo ARMA, temos $\hat{r}_{t+j} = \sum_{i=0}^{\infty} \psi_i \hat{X}_{t+j-i} = \sum_{i=j}^{\infty} \psi_i X_{t+j-i}$, onde ψ_i, para todo $i > 0$, são os coeficientes da representação MA(∞) do processo estocástico $\{r_t\}_{t\in Z}$. Logo, o erro de previsão, para cada $1 \leq j \leq h$, é dado por

$$e_t(j) = r_{t+j} - \hat{r}_{t+j} = \sum_{i=0}^{\infty} \psi_i X_{t+j-i} - \sum_{i=j}^{\infty} \psi_i X_{t+j-i} = \sum_{i=0}^{j-1} \psi_i X_{t+j-i}$$

e assim, $e_t[h] = \sum_{j=0}^{h-1} \left[(\sum_{i=0}^{j} \psi_i) X_{t+h-j} \right]$.

Para provar (iii) note que,

$$\sigma_t^2[h] = \mathrm{Var}(r_t[h]|\mathcal{F}_t) = \mathrm{Var}(r_t[h] - \hat{\mu}_t[h]|\mathcal{F}_t) \text{ e } \hat{\mu}_t[h] = \hat{r}_t[h].$$

Além disso, $E(X_{t+j}|\mathcal{F}_t) = E(\sigma_{t+j}|\mathcal{F}_t)E(Z_{t+j}) = 0$. Logo, $\mathrm{Cov}(X_{t+j}, X_{t+k}|\mathcal{F}_t) = 0$, para todo $0 < j < k$. Portanto,

$$\hat{\sigma}_t^2[h] = \mathrm{Var}(r_t[h] - \hat{r}_t[h]|\mathcal{F}_t) = \mathrm{Var}(e_t[h]|\mathcal{F}_t) = \sum_{j=0}^{h-1} \left[\left(\sum_{i=0}^{j} \psi_i \right)^2 \hat{\sigma}_{t+h-j}^2 \right].$$

2.3. Análise de Cenários - Teste de Estresse

O teste de estresse baseia-se na idéia de que o valor do portfólio depende dos fatores de risco. Sejam $\mathbf{Z} = (Z_1, \cdots, Z_m)'$ o vetor dos m fatores de risco que influenciam no valor do portfólio e $f(\cdot)$ a função que determina o valor deste portfólio. Então, dizemos que o vetor $\mathbf{Z}$ descreve a situação do mercado e que $f(\mathbf{Z})$ é o valor do portfólio para esta situação (ou cenário). No que segue, nos referimos a $\mathbf{Z}_{AM}$ como a *situação atual do mercado* e conseqüentemente $f(\mathbf{Z}_{AM})$ representa o *valor atual do portfólio*.

Para o teste de estresse, k diferentes cenários $\mathbf{Z_1}, \cdots, \mathbf{Z_k}$ são selecionados de acordo com algum critério específico e os valores do portfólio $f(\mathbf{Z_1}), \cdots, f(\mathbf{Z_k})$ são calculados sobre esses cenários. Comparando com o valor atual do portfólio $f(\mathbf{Z}_{AM})$, podemos determinar as perdas que ocorreriam se o mercado passasse repentinamente da situação $\mathbf{Z}_{AM}$ para alguma das k situações $\mathbf{Z_1}, \cdots, \mathbf{Z_k}$. Ou seja, o teste de estresse nos diz o que aconteceria se uma dada situação de mercado $\mathbf{Z}$ repentinamente ocorresse.

2.3.1. Identificação dos Máximos e Mínimos para Fatores de Risco Individuais

Sejam $\mathcal{P}$ um portfólio fixo e $\mathbf{Z} = (Z_1, \cdots, Z_m)'$ o vetor dos m fatores de risco que influenciam no valor deste portfólio. O método tradicional de realizar um teste de estresse consiste em construir os cenários baseando-se em dados históricos dos fatores

de risco. Define-se *o período de observação histórica* como o período no qual a série temporal é considerada, por exemplo, 1 ou 10 anos. O período de observação histórica é então sobreposto por janelas de igual duração, por exemplo, 1 ou 10 dias. Para cada janela temporal determinamos as mudanças de fatores de risco ΔZ_i, para cada $i \in \{1, \cdots, m\}$. O máximo ou mínimo das mudanças de fatores de risco entre todas as janelas temporais é então fixado como a mudança de fator de risco ΔZ_i, para cada $i \in \{1, \cdots, m\}$.

A mudança dos fatores de risco é geralmente definida como a variação entre o primeiro e o último dia da janela temporal e é denominada *Start to End* (StE). Note que, se assumimos que os fatores de risco são os logaritmos dos preços dos ativos, então as mudanças de fatores de risco são os log-retornos de período igual a duração da janela. Alternativamente, a mudança dos fatores de risco pode ser definida como o máximo das variações entre dois pontos quaisquer da janela temporal. Essa variação é denominada *drawdown* (DD).

É fácil ver que, se escolhermos a variação mínima dentro de cada janela, então ΔZ_i, para cada $i \in \{1, \cdots, m\}$, representa a redução máxima dos fatores de risco. Se o máximo é selecionado, então estamos considerando o aumento máximo nos fatores de risco. Podemos também considerar ΔZ_i, para cada $i \in \{1, \cdots, m\}$, como máximo para o valor absoluto dessas variações. Neste caso, obtemos a variação máxima dos fatores de risco não levando em consideração se a mudança é positiva ou negativa.

2.3.2. Perda Máxima - *MaxLoss*

Em contraste com o VaR_p, a perda máxima *(MaxLoss)* é uma medida de risco coerente (veja Studer, 1997). Portanto, a propriedade de subaditividade é sempre válida. Além disso, o *MaxLoss* informa, além da dimensão da perda, o cenário em que a pior perda ocorre. A definição formal dessa medida de risco é apresentada a seguir.

Definição 2.4 - Dado um domínio de admissibilidade A, a *perda máxima* de um portfólio contido em A é dado por $MaxLoss_A(f) := f(\mathbf{Z}_{AM}) - min_{\mathbf{Z}\in A}\{f(\mathbf{Z})\}$, onde $f(\cdot)$ é a função que determina o preço do portfólio e $\mathbf{Z}_{AM} = (Z_{AM,1}, \cdots Z_{AM,m})'$ representa o vetor de m fatores de risco para a situação atual do mercado.

Note que, para calcular a perda máxima deve-se escolher uma região de confiança fechada A, com uma dada probabilidade p de ocorrência. Então, uma definição equivalente é

$$MaxLoss_A(f) = \max\{f(\mathbf{Z}_{AM}) - f(\mathbf{Z}): \mathbf{Z} \in A \quad e \quad P(A) = \mathrm{p}\}.$$

Note ainda que a expressão $f(\mathbf{Z}_{AM}) - f(\mathbf{Z})$ representa a perda L (ou $-L$, se $\mathbf{Z}$ é medido em um instante de tempo anterior a AM) do portfólio. Dizemos que um portfólio *é linear* se L_t é uma função linear em relação a cada uma das mudanças de fatores de risco. O teorema que segue fornece a expressão da medida *MaxLoss* para um portfólio linear.

Teorema 2.1 - Sejam $\mathcal{P}$ um *portfólio linear* e $f(\cdot)$ a função que determina o valor do portfólio, então $f(\boldsymbol{X}) = \boldsymbol{a}'\boldsymbol{X}$, *onde* $\boldsymbol{a} \in R^m$ é um vetor de constantes reais e $\boldsymbol{X} \in R^m$ é o vetor de mudanças de fatores de risco. Segue que, dado um nível de confiança p, a perda máxima do portfólio é dada por

$$MaxLoss = -\sqrt{c_{\mathrm{p}}}\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}, \qquad (8)$$

onde $\boldsymbol{\Sigma}$ é a matriz de variância-covariância das mudanças de fatores de risco e c_{p} é o p-quantil da função de distribuição $\mathcal{X}^2$ com m graus de liberdade. O pior cenário é dado por

$$\mathbf{Z}^* = -\frac{\sqrt{c_{\mathrm{p}}}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}}}\boldsymbol{\Sigma}\mathbf{a}. \qquad (9)$$

Prova: Veja Studer (1997), Teorema 3.15.

3. Análise de Séries Temporais Reais

Nesta seção apresentamos a estimação e análise de medidas de risco para um portfólio real de ações. Este portfólio é formado por ações de quatro empresas brasileiras (também denominadas ativos do portfólio). Denotamos esses ativos por A_i, $i \in \{1, \cdots, 4\}$, de forma que A_1 representa as ações do Bradesco; A_2 representa as ações da Brasil Telecom; A_3 representa as ações da Gerdau e A_4 representa as ações da Petrobrás. Tais ações são negociadas na bolsa de valores de São Paulo (Bovespa). Utilizamos a notação A_M (ou, equivalentemente, A_5) para denotar o mercado financeiro. Os valores do portfólio do mercado são representados pelo índice Bovespa.

Nesta análise, os fatores de risco para este portfólio são os logaritmos dos preços das ações. Sendo assim, o vetor de fatores de risco é $\mathbf{Z}_t = (\ln(P_{1,t}), \ln(P_{2,t}), \ln(P_{3,t}), \ln(P_{4,t}))'$ e o vetor de mudanças de fatores de risco $\mathbf{X}_t$ é o vetor dos log-retornos $\mathbf{X}_t = (r_{1,t}, r_{2,t}, r_{3,t}, r_{4,t})'$. Assumimos que o instante inicial $t = 0$ é igual ao primeiro dia de observação das séries temporais e consideramos $V_0 = 1$. Então, o valor porfólio V_t pode ser escrito como

$$V_t = V_0 \left(\frac{a_1}{P_{1,0}} P_{1,t} + \frac{a_2}{P_{2,0}} P_{2,t} + \frac{a_3}{P_{3,0}} P_{3,t} + \frac{a_4}{P_{4,0}} P_{4,t} \right),$$

e depende apenas do preço dos ativos no tempo t. Para o cálculo das medidas de risco consideramos, sem perda de generalidade, a série temporal $\{-r_{P,t}\}_{t=1}^{n}$, que representa a perda percentual, ao invés da série das perdas, definida por $L_t = -V_{t-1} r_{P,t}$, para todo $t \in N$, que representa a perda em unidades monetárias. A função de distribuição considerada para o cálculo das medidas de risco foi a Gaussiana.

3.1. Características das Séries Temporais

No que segue, apresentamos a análise das características das séries temporais dos log-retornos dos preços das ações dos ativos considerados e dos log-retornos do índice de mercado. Tal análise é importante na seleção de um modelo para a série temporal dos log-retornos do portfólio e para identificar a relação entre os ativos e o índice do mercado.

No que segue, apresentamos a análise das características das séries temporais dos log-retornos dos preços das ações dos ativos considerados e dos log-retornos do índice de mercado. Tal análise é importante na seleção de um modelo para a série temporal dos log-retornos do portfólio e para identificar a relação entre os ativos e o índice do mercado.

A Figura 1 apresenta as séries temporais compostas por $n = 1728$ observações que representam, respectivamente, os preços das ações do Bradesco, Brasil Telecom, Gerdau e Petrobrás, bem como a série temporal composta por $n = 1729$ observações que representa o índice diário da Bolsa de Valores de São Paulo no período de janeiro de 1995 a dezembro de 2001. Ainda na Figura 1, apresentamos os respectivos log-retornos e os valores quadráticos dos log-retornos para as séries temporais consideradas. Observa-se que tanto o índice do mercado como os preços dos ativos sofreram uma forte queda em seus valores no período em torno de $t = 1000$ (15 de janeiro de 1999). Esse período corresponde à janeiro de 1999 e foi marcado pela desvalorização do real. Observamos, nestas séries temporais, as características comuns àquelas de retornos financeiros, tais como média aproximadamente em torno de zero e agrupamentos de volatilidade.

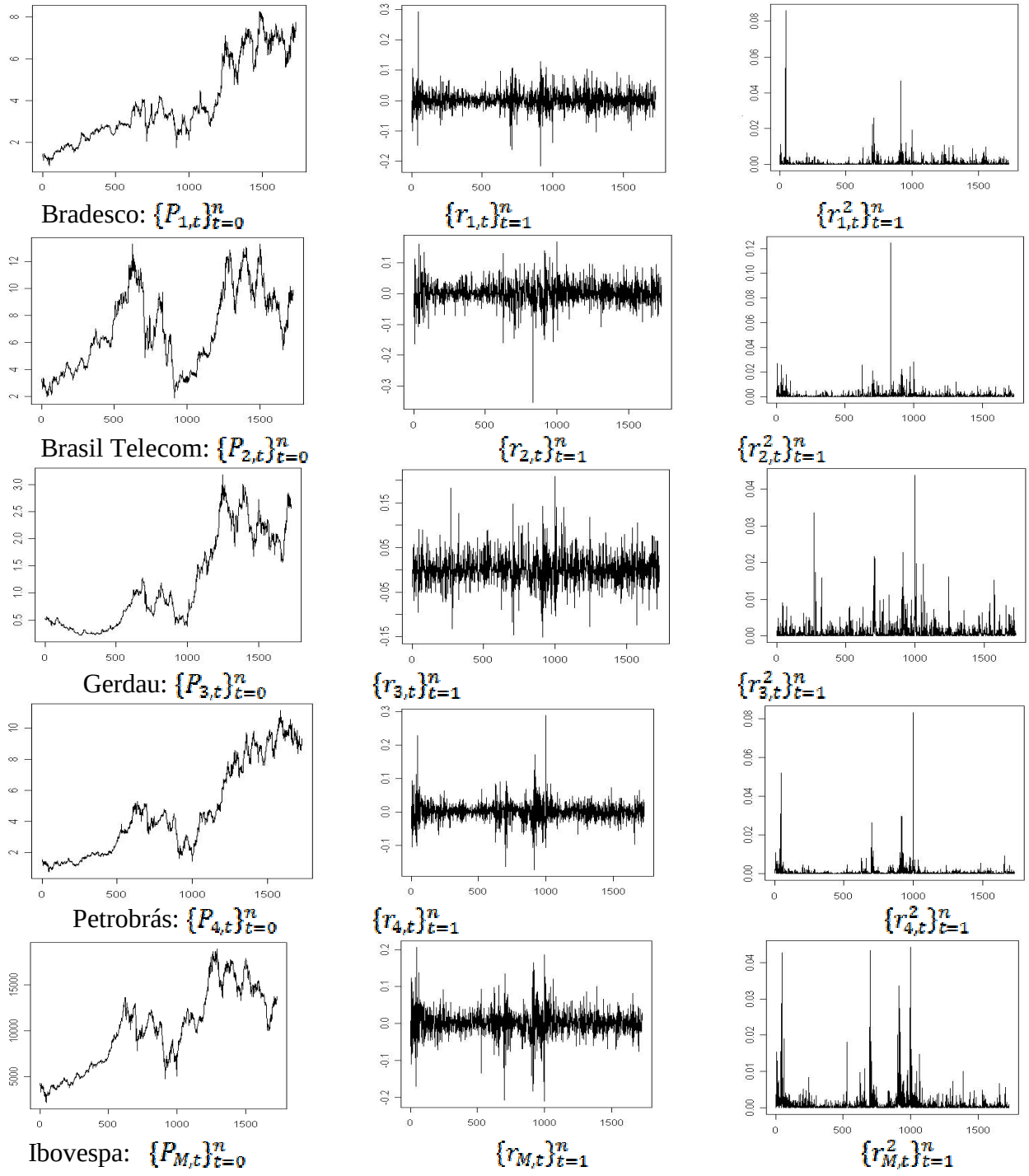


Figura 1: Séries temporais do preços dos ativos, do índice Bovespa e dos respectivos log-retornos $\{r_{i,t}\}_{t=1}^{n}$ e valores quadráticos dos log-retornos $\{r_{i,t}^2\}_{t=1}^{n}$, com $n = 1728$.

3.2. Estatísticas Descritivas dos Ativos Financeiros

Na Tabela 1 apresentamos as estatísticas descritivas das séries temporais analisadas. Observamos que todos os ativos apresentam desempenho semelhante, quantificado pelo valor da mediana dos retornos diários igual a zero em todos os casos. Entretanto, esse valor está abaixo daquele apresentado pelo mercado (0,0014) no mesmo período. O ativo que apresentou maior variabilidade foi o A_2, com variância igual a 0,0013, sendo esse valor maior do que a variância do mercado no período, que foi de 0,0008. O ativo A_1 é o menos volátil, com variância igual a 0,0009. Entre todos os ativos, o que apresentou o menor retorno no período foi o ativo A_2 (-0,3530), e este foi menor do que o menor retorno apresentado pelo mercado no mesmo período (-0,1721). O ativo que apresentou o maior retorno no período foi o A_1 (0,2933) e foi superior ao maior retorno do mercado (0,2883) no período. Percebemos também que, para todas as séries temporais de log-retornos, a função de distribuição dos dados é aproximadamente simétrica, dado que a medida de assimetria assume valores pequenos. Entretanto, todas as séries temporais podem ser consideradas como de caudas pesadas. A série temporal dos log-retornos do ativo A_1 apresenta o maior valor da medida de curtose (11,9990) e esse valor é menor do que o da medida de curtose do mercado (15,5581).

Tabela 1: Estatísticas Descritivas para as Séries Temporais dos Log-retornos dos Preços das Ações do Bradesco (A_1), Brasil Telecom (A_2), Gerdau (A_3) e Petrobrás (A_4) e do Índice do Mercado (A_M).

Ativo	A_1	A_2	A_3	A_4	A_M
Média	0,0010	0,0006	0,0009	0,0010	0,0007
Variância	0,0009	0,0013	0,0011	0,0011	0,0008
Mediana	0,0000	0,0000	0,0000	0,0000	0,0014
Máximo	0,2933	0,1676	0,2092	0,2068	0,2883
Mínimo	-0,2162	-0,3530	-0,1507	-0,2103	-0,1721
Assimetria	0,2390	-0,5773	0,3675	-0,0738	0,6920
Curtose	11,9990	10,6346	6,5431	9,1426	15,5581

3.3. Cálculo dos Pesos do Portfólio

Utilizamos a taxa SELIC, cujo log-retorno médio no período é igual à $R_F = 0{,}02$, como ativo livre de risco. No cálculo dos pesos $\mathbf{a} = (a_1, a_2, a_3, a_4)'$, para o portfólio $\mathcal{P}$ em questão obtivemos

$$(a_1, a_2, a_3, a_4) = (0{,}3381; 0{,}1813; 0{,}3087; 0{,}1719),$$

isto é, para os ativos considerados, o portfólio será eficiente se 33,81% do valor total a ser investido for aplicado no ativo A_1; 18,13% no ativo A_2; 30,87% no ativo A_3 e 17,19% no ativo A_4.

A Figura 2 apresenta a série temporal dos valores do portfólio $\{V_t\}_{t=1}^{n}$ e a dos log-retornos do portfólio (representa os ganhos percentuais), bem como série temporal das perdas do portfólio (calculada em termos de unidades monetárias) e sua função densidade de probabilidade. Comparando a série temporal dos valores do portfólio com a série temporal do índice de mercado, apresentada na Figura 1, observamos que ambas apresentam um comportamento muito semelhante. Observamos ainda que a maior perda ocorre em $t = 698$ com $L_t = 0{,}3729$ (ou seja, o maior ganho, que é igual a aproximadamente R$ 0,37 por cada real investido) e a menor perda ocorre em $t = 1244$, com $L_t = -0{,}3402$ (ou seja, aproximadamente R$ 0,34 por cada real investido). A maior perda corresponde a mudança no valor do portfólio, ocorrida entre 24/10/1997 (sexta-feira) e 27/10/1997 (segunda-feira). Nesse período o valor do índice Bovespa passou de 11.545,20 para 9.816,80 pontos, o que representa uma queda de 14,97%. Nessa data também ocorreram grandes quedas nos índices Dow Jones (7,18%) e S&P 500 (6,87%). Este período é marcado pela forte crise na Ásia. O maior ganho no portfólio corresponde a variação ocorrida no período de 13/01/2001 à 14/01/2001. O Ibovespa apresenta uma alta de 2,08% neste período. Enquanto que, para o mercado mundial temos que em 14/01/2000 o valor do índice Dow Jones foi de 11.722,98 pontos. Esse valor só foi superado em 03/10/2006, quando o índice atingiu 11.727,34 pontos.

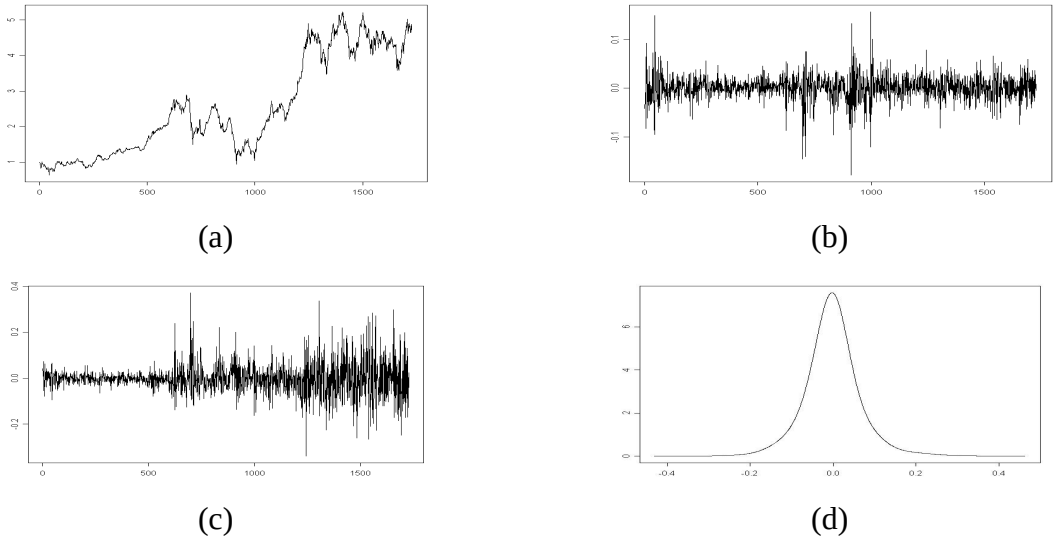


Figura 2: Séries temporais dos (a) valores do portfólio $\mathcal{P}$; (b) log-retornos do portfólio; (c) perdas do portfólio, no período de janeiro de 1995 à dezembro de 2001. (d) Função densidade de probabilidade das perdas do portfólio.

Essa análise nos fornece uma idéia do que aconteceria se um investidor mantivesse seu portfólio constante durante o período considerado. Na prática, o investidor compra e vende ações a fim de obter melhores rendimentos. A fim de decidir se as ações estão subavaliadas ou superavaliadas o investidor pode utilizar como ferramenta o coeficiente α das ações obtido através do modelo CAPM.

3.4. Modelo CAPM

Na Tabela 2 apresentamos a correlação amostral entre as séries temporais analisadas. Observamos que os ativos possuem correlação significante, tanto entre si, como com o mercado financeiro.

Tabela 2: Matriz de correlação amostral para os log-retornos dos preços das ações do Bradesco, Brasil Telecom, Gerdau, Petrobás e do Ibovespa, no período de janeiro de 1995 à dezembro de 2001.

Ativo	Bradesco	Brasil Telecom	Gerdau	Petrobrás	Ibovespa
Bradesco	1,0000	0,6887	0,4517	0,5982	0,7024
Brasil Telecom	0,6887	1,0000	0,5895	0,6435	0,7817
Gerdau	0,4517	0,5895	1,0000	0,5927	0,7379
Petrobrás	0,5982	0,6435	0,5927	1,0000	0,8373
Ibovespa	0,7024	0,7817	0,7379	0,8373	1,0000

Os valores estimados dos parâmetros β_i, que representa o risco sistemático, e α_i, que indica se o ativo está subavaliado ou superavaliado, obtidos através da regressão (3), são apresentados na Tabela 3, para cada $i \in \{1,\cdots,4\}$.

Tabela 3: Valores dos coeficientes β_i e α_i do modelo CAPM para o ativo A_i, $i \in \{1,\cdots,4\}$.

Ativo	A_1	A_2	A_3	A_4
β_i	0,9525	1,0508	1,1267	1,0153
α_i	-0,0006	0,0009	0,0026	0,0006

Utilizando-se os valores apresentados na Tabela 3, conclui-se que o coeficiente β_P do portfólio é igual a $\beta_P = \sum_{i=1}^{4} a_i\beta_i = 1{,}0349$. Sendo assim, se o índice do mercado sofrer um aumento de 10%, o valor do portfólio sofrerá um aumento de 10,35%. Pelos valores apresentados na Tabela 3 conclui-se ainda que, na visão do investidor, as ações do ativo A_1 estão superavaliadas pelo mercado ($\alpha < 0$) o que deverá proporcionar uma rentabilidade inferior àquela calculada pelo modelo de equilíbrio (CAPM). Portanto, devem ser vendidas. Enquanto que, para os outros ativos, o valor positivo do coeficiente α indica que, as ações estão subavaliadas, isto é, seu preço é menor do que seu valor intrínseco. Sendo assim, é conveniente comprá-las.

A Tabela 4 apresenta os valores (observados) dos retornos de períodos 1 e 10 dias, denotados, respectivamente, por $R_i = r_{i,1729}$ e $R_i[10] = r_{i,1728}[10]$, para cada $i \in \{1,\cdots,4\}$. Na Tabela 4 também são dados os valores para a média amostral $\bar{r}_i = 1/1728 \sum_{t=1}^{1728} r_{i,t}$ dos retornos observados e os retornos estimados pelo modelo CAPM, dados por $E(R_i) = R_F + \beta_i(E(R_M) - R_F)$, com $E(R_M)$ estimado através da média amostral dos retornos do mercado, $R_F = 0{,}02$ e β_i dado na Tabela 3, para todo $i \in \{1,\cdots,4\}$.

Tabela 4: Valores observados dos retornos de períodos 1 e 10 dias (R_i e $R_i[10]$, respectivamente), média amostral dos retornos observados ($\bar{R}_i$) e os retornos estimados pelo modelo CAPM ($E(R_i)$) para o ativo A_i, $i \in \{1,\cdots,4\}$.

Ativo	R_i	$R_i[10]$	$\bar{R}_i$	$E(R_i)$
A_1	-0,0026	-0,0156	-0,013	0,0016
A_2	0,0301	-0,0343	-0,017	-0,0003
A_3	0,0680	0,2456	-0,040	-0,0017
A_4	0,0021	-0,0551	0,003	0,0004

Comparando com os valores obtidos pelo modelo CAPM notamos que, para o ativo A_1 (Bradesco) o log-retorno observado foi menor que o retorno estimado, enquanto que, para os outros ativos os resultados obtidos através do modelo CAPM foram inferiores aos observados. Para um período $h=10$ dias o retorno previsto pelo modelo CAPM só foi inferior aos log-retornos observados no caso do ativo A_3. Para o portfólio temos $R_{\mathcal{P}} = 0{,}02593$, $R_{\mathcal{P}}[10] = 0{,}05485$ e $E(R_{\mathcal{P}}) = \sum_{i=1}^{4} a_i E(R_i) = 0{,}00003$. Logo o log-retorno observado foi maior que o estimado, tanto para $h = 1$ quando para $h=10$ dias.

Na Figura 3 apresentamos os valores observados (círculos vazios) e valores estimados (círculos sólidos) dos log-retornos $r_{\mathcal{P},n+h}$ do portfólio, para $n = 1728$ e $h \in \{1,\cdots,10\}$. Os valores estimados foram obtidos através da expressão $\hat{r}_{\mathcal{P},n+h} = R_F + \beta_{\mathcal{P}}(r_{M,n+h} - R_F)$, para todo $h \in \{1,\cdots,10\}$, onde $n = 1728$, $R_F = 0{,}02$, $\beta_{\mathcal{P}} = 1{,}0349$ e $r_{M,n+h}$ é o valor observado do índice da Bovespa, no tempo $n+h$. Note que os valores estimados foram próximos aos valores observados, em particular, para $h = 1$, o log-retorno observado foi igual a 0,02593, enquanto que o estimado foi 0,02217. Em todos os casos, o retorno real foi maior do que o predito.

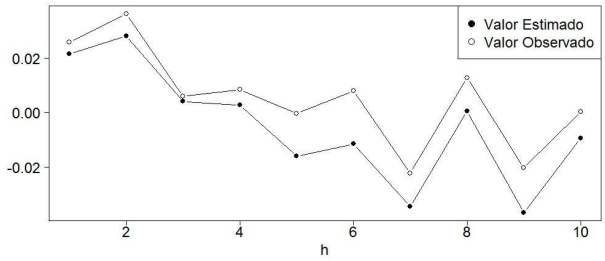
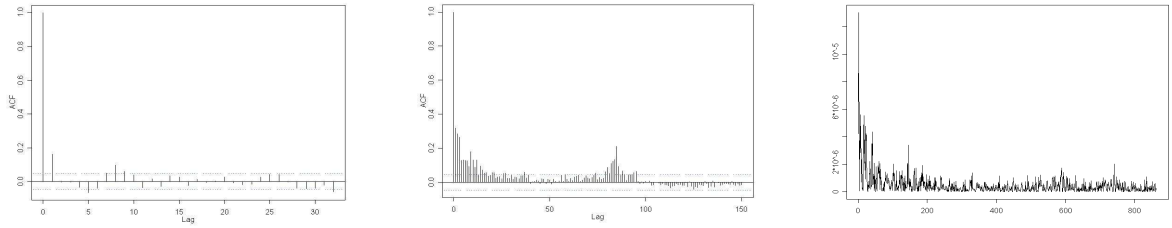


Figura 3: Valores observados (círculos vazios) $r_{\mathcal{P},n+h}$ e valores estimados $\hat{r}_{\mathcal{P},n+h}$ (círculos sólidos) dos log-retornos do portfólio, para $n = 1728$ e $h \in \{1,\cdots,10\}$.

3.5. Cálculo das Medidas de Risco VaR e ES

No que segue apresentamos o cálculo das medidas de risco baseadas na distribuição das perdas do portfólio. Para o cálculo das medidas de risco VaR e ES utilizou-se modelos ARMA(p_1, q_1)-FIEGARCH(p_2, d, q_2) e ARMA(p_1, q_1)-EGARCH(p_2, q_2) (abordagem econométrica) e o horizonte considerado é $h = 1$ dia.



(a) (b) (c)

Figura 4: Função de autocorrelação amostral da série temporal (a) $\{-r_{\mathcal{P},t}\}_{t=1}^{1728}$; (b) $\{r_{\mathcal{P},t}^2\}_{t=1}^{1728}$. (c) Função periodograma da série temporal $\{r_{\mathcal{P},t}^2\}_{t=1}^{1728}$.

A função de autocorrelação amostral da série temporal $\{-r_{\mathcal{P},t}\}_{t=1}^{n}$ e as funções de autocorrelação amostral e periodograma da série dos valores quadráticos dos log-retornos do portfólio são apresentadas na Figura 4. Observamos que, embora a função de autocorrelação amostral da série temporal $\{-r_{\mathcal{P},t}\}_{t=1}^{n}$ apresente um decaimento muito rápido, ela apresenta valores significativos para $h > 1$, o que justifica a escolha de um modelo ARMA(p_1, q_1). Pela Figura 4(b) conclui-se que a série temporal $\{r_{\mathcal{P},t}^2\}_{t=1}^{1728}$ é correlacionada, o que indica a presença de heteroscedasticidade. Na Figura 4(b) observa-se ainda que a função de autocorrelação amostral apresenta valores significativos para diversos valores de h distantes de 1, o que indica longa dependência. Na Figura 4(c) observamos um pico na função periodograma na vizinhança da frequência zero, ou seja, outro indicativo de longa dependência. Tais características justificam a escolha de um modelo FIEGARCH(p_2, d, q_2).

Para a seleção do modelo mais adequado, considerando a longa dependência,

ajustamos modelos ARMA(p_1, q_1)-FIEGARCH(p_2, d, q_2) à série temporal $\{-r_{\mathcal{P},t}\}_{t=1}^{n}$, com valores de $p_1, q_1 \in \{0,1,2,3\}$ e $p_2, q_2 \in \{0,1\}$. Levando em conta a hipótese de que os resíduos devem ser uma sequência i.i.d. e que estamos interessados em modelos estacionários, descartamos os casos para os quais os resíduos ainda apresentavam correlação, heteroscedasticidade, ou tais que o valor estimado do parâmetro d foi maior do que 0,5. Dentre os modelos que restaram, selecionamos os que apresentavam o menor valor dos critérios AIC e BIC e\ou o maior valor da log-verossimilhança. Os valores desses critérios para os três modelos selecionados são apresentados na Tabela 5. Nesta tabela, o valor em negrito indica o melhor modelo em relação a cada critério. Tendo em vista que os valores de um mesmo critério para diferentes modelos são muito próximos, e que a análise dos resíduos indica que os três modelos são adequados para descrever a série temporal dos log-retornos do portfólio, optamos pelo mais parcimonioso. Os coeficientes do modelo escolhido são apresentados na Tabela 6. A título de comparação, utilizamos um modelo ARMA(p_1, q_1)-EGARCH(p_2, q_2) (veja Tabela 6) com os mesmos valores de p_1, p_2, q_1 e q_2 que o modelo ARMA(p_1, q_1)-FIEGARCH(p_2, d, q_2).

Tabela 5: Valores dos critérios AIC, BIC e Log-verossimilhança para três models ARMA(p_1, d_1)-FIEGARCH(p_2, d, q_2) para a série temporal $\{-r_{\mathcal{P},t}\}_{t=1}^{1728}$.

p_1	q_1	p_2	d	q_2	AIC	BIC	Log-Verossimilhança
1	0	0	0,2294	1	-8448,2644	**-8410,0813**	4231,1322
2	1	0	0,2367	1	**-8456,7829**	-8407,6904	4237,3914
2	1	1	0,2835	1	-8456,5281	-8396,5262	**4239,2640**

Tabela 6: Valores estimados para os parâmetros dos models ARMA$(1,0)$-EGARCH$(0,1)$ e ARMA$(1,0)$-FIEGARCH$(0,d,1)$ para a série temporal $\{-r_{\mathcal{P},t}\}_{t=1}^{n}$. O valor em parênteses corresponde ao desvio padrão do estimador.

	Modelo	
Parâmetro	ARMA-EGARCH	ARMA-FIEGARCH
ϕ_0	-0,001 (0,001)	-0,001 (0,001)
ϕ_1	0,172 (0,025)	0,173 (0,026)
α	-1,007 (0,113)	-0,498 (0,093)
β_1	0,899 (0,013)	0,754 (0,060)
ψ_0	0,295 (0,028)	0,285 (0,029)
γ_0	0,448 (0,073)	0,127 (0,017)
d	-	0,233 (0,056)

As previsões para a média e para o desvio padrão condicional da série temporal $\{-r_{\mathcal{P},t}\}_{t=1}^{1728}$, para os horizontes $h \in \{1,\cdots,10\}$ dias, são apresentadas na Tabela 7. Os resultados da estimação das medidas de risco VaR e ES, utilizando a série temporal dos log-retornos do portfólio, são apresentados na Tabela 8. Da Tabela 4 temos que $r_{1,1729} = -0{,}0026$, $r_{2,1729} = 0{,}0301$, $r_{3,1729} = 0{,}0680$ e $r_{4,1729} = 0{,}0021$. Logo, o log-retorno do portfólio, para $n = 1728$ e $h = 1$ dia, é igual a $r_{\mathcal{P},n+1} = 0{,}0259$. Comparando o valor observado do log-retorno do portfólio $r_{\mathcal{P},1729} = 0{,}0259$ com os valores apresentados na Tabela 8, conclui-se que o valor estimado utilizando-se o modelo FIEGARCH foi o que mais se aproximou do valor observado, tanto para $p = 0{,}95$ quanto para $p = 0{,}99$. Além disso, comparando os valores da Tabela 8 com o valor observado $r_{\mathcal{P},1729} = 0{,}0259$ e com o valor estimado utilizando-se o $\beta_{\mathcal{P}}$ e o log-retorno observado do Ibovespa ($\hat{r}_{\mathcal{P},1729} = 0{,}0222$), conclui-se que a abordagem econométrica utilizando modelos FIEGARCH foi mais precisa.

Tabela 7: Previsão da média e do desvio padrão condicional dos log-retornos do portfólio para os horizontes $h \in \{1,\cdots,10\}$ dias.

h	$-\hat{r}_{\mathcal{P},t+h}$	$\hat{\sigma}_{\mathcal{P},t+h}$	h	$-\hat{r}_{\mathcal{P},t+h}$	$\hat{\sigma}_{\mathcal{P},t+h}$
1	0,0001	0,0149	6	-0,0010	0,0161
2	-0,0008	0,0152	7	-0,0010	0,0162
3	-0,0010	0,0155	8	-0,0010	0,0164
4	-0,0010	0,0157	9	-0,0010	0,0165
5	-0,0010	0,0159	10	-0,0010	0,0166

Tabela 8: Valores estimados para as medidas de risco VaR e ES para as perdas do portfólio, ao nível de confiança p = 95% e p = 99% para o horizonte $h = 1$ dia.

Abordagem	p = 0,95		p = 0,99	
	$\mathrm{VaR}_{p,n+1}$	$\mathrm{ES}_{p,n+1}$	$\mathrm{VaR}_{p,n+1}$	$\mathrm{ES}_{p,n+1}$
Empírico	0,0369	0,0588	0,0703	0,0966
Normal	0,0398	0,0712	0,0566	0,0923
EWMA	0,0321	0,0583	0,0461	0,0759
EGARCH	0,0353	0,0625	0,0499	0,0808
FIEGARCH	0,0247	0,0437	0,0349	0,0564

3.6. Análise de Cenários (Teste de Estresse)

Para realizar o teste de estresse, utilizamos duas abordagens diferentes. A primeira delas consiste na escolha dos cenários baseando-se nos dados históricos. Tal procedimento é conhecido na literatura como *teste de estresse tradicional*. A segunda abordagem consiste na escolha dos cenários sob um certo domínio de admissibilidade, obtendo então a medida denominada *perda máxima*, ou *MaxLoss.*

Para a construção de cenários baseados em dados históricos, primeiramente determinamos as variações máximas e mínimas de cada fator de risco (log-retornos) utilizando-se os procedimentos StE (*Start to End*) e DD (*drawdown*), descritos na Seção 2.3.1, com janelas de 1 e 10 dias. Em seguida, utilizando-se essas variações máximas e mínimas, construimos 12 diferentes cenários, denotados por Sj, para $j \in \{1,\cdots,12\}$, como descrito a seguir.

• Denotamos por $r_{i,Sj}$ o valor do log-retorno do ativo A_i sob o cenário Sj, para cada $i \in \{1,\cdots,4\}$ e $j \in \{1,\cdots,12\}$.

• Os cenários S1, S2 e S3 são formados pelos log-retornos máximos dos ativos, obtidos pelo método StE e DD, com janelas de 1 e 10 dias.

• Os cenários S4, S5 e S6 são formados pelos retornos mínimos dos ativos, obtidos pelo método StE e DD, com janelas de 1 e 10 dias.

• Os cenários S7, S8 e S9 são construídos considerando-se os log-retornos máximos do mercado e os coeficientes β_i, para cada $i \in \{1,\cdots,4\}$, fornecidos pelo modelo CAPM, isto é, define-se $r_{i,Sj} := \beta_i \times r_{M,StE}$ ou $r_{i,Sj} := \beta_i \times r_{M,DD}$, onde $r_{M,StE}$ e $r_{M,DD}$ denotam, respectivamente, os log-retornos máximos do mercado obtidos pelo método StE e DD.

• Os cenários S10, S11 e S12 consideram os log-retornos mínimos do mercado e os coeficientes β_i, para cada $i \in \{1, \cdots, 4\}$, fornecidos pelo modelo CAPM. A construção desses cenários é análoga à dos cenários S7,S8 e S9.

Os valores obtidos para as variações máximas e mínimas dos fatores de risco (ou seja, os log-retornos máximos e mínimos) para cada um dos ativos e para o mercado financeiro são apresentados na Tabela 9. Os diferentes cenários considerados são apresentados na Tabela 10. Observamos que S1, S2, S3, S7, S8 e S9 são cenários otimistas (assumem que os log-retornos são iguais aos máximos históricos), enquanto que os demais cenários são todos pessimistas. Os log-retornos do portfólio sob cada cenário são apresentados na Tabela 11.

Tabela 9: Variações máximas e mínimas dos log-retornos dos ativos do portfólio e dos log-retornos do mercado.

	1 dia (StE)		10 dias (StE)		10 dias (DD)	
Ativo	Max	Min	Max	Min	Max	Min
A_M	0,2883	0,1721	0,2360	-0,2853	0,3848	-0,3713
A_1	0,2933	-0,2162	0,3010	-0,4761	0,3669	-0,4761
A_2	0,1676	-0,3530	0,3223	-0,5626	0,4037	-0,5626
A_3	0,2094	-0,1508	0,3390	-0,3111	0,4558	-0,3189
A_4	0,2068	-0,2103	0,3495	-0,2970	0,4242	-0,4937

Tabela 10: Cenários para o Teste de Estresse.

Cenários Baseados em Dados Históricos					Cenários Baseados no β do Modelo CAPM				
	$r_{1,Sj}$	$r_{2,Sj}$	$r_{3,Sj}$	$r_{4,Sj}$		$r_{1,Sj}$	$r_{2,Sj}$	$r_{3,Sj}$	$r_{4,Sj}$
S1	0,2933	0,1676	0,2094	0,2068	S7	0,2746	0,3030	0,3249	0,2927
S2	0,3010	0,3223	0,3390	0,3495	S8	0,2248	0,2480	0,2659	0,2396
S3	0,3669	0,4037	0,4558	0,4242	S9	0,3665	0,4043	0,4335	0,3907
S4	-0,2162	-0,3530	-0,1508	-0,2103	S10	-0,1639	0,1808	-0,1939	-0,1747
S5	-0,4761	-0,5626	-0,3111	-0,2970	S11	-0,2717	-0,2998	-0,3214	0,2896
S6	-0,3713	-0,5626	-0,3189	-0,4937	S12	-0,3537	-0,3902	-0,4184	-0,3770

Tabela 11: Valores dos log-retornos do portfólio sob os diferentes cenários.

Cenário	$r_{\mathcal{P},Sj}$	Cenário	$r_{\mathcal{P},Sj}$	Cenário	$r_{\mathcal{P},Sj}$	Cenário	$r_{\mathcal{P},Sj}$
S1	0,2297	S4	-0,2198	S7	0,2984	S10	-0,1781
S2	0,3249	S5	-0,4100	S8	0,2442	S11	-0,2952
S3	0,4108	S6	-0,4463	S9	0,3982	S12	-0,3843

Os log-retornos reais observados do portfólio para 1 e 10 dias foram, respectivamente, 0,0259 e 0,0548. Comparando esses valores com aqueles apresentados na Tabela 11 conclui-se que, tanto para $h = 1$ quanto para $h = 10$ dias, os log-retornos reais foram sempre superiores aos previstos pelos cenários pessimistas e foram superestimados por todos os cenários otimistas.

Pelo Teorema 2.1, a um nível de confiança p%, a perda máxima do portfólio é dada pela expressão (8) e o cenário onde isso ocorre é dado pela expressão (9). Na Tabela 12 apresentamos a perda máxima do portfólio (*MaxLoss*), para diferentes valores de p (domínios de admissibilidade) e o valor do log-retorno $r_{i,ML}$ do ativo A_i sob o cenário de perda máxima, para cada $i \in \{1, \cdots, 4\}$. Comparando os valores apresentados nas Tabelas 11 e 12, percebemos que as perdas que ocorrem sob os cenários pessimistas construídos sob dados históricos são maiores do que aquelas que ocorrem quando consideramos cenários sob um certo domínio de admissibilidade. Para todos os valores de p, a perda máxima estimada foi maior que a perda real observada no instante $t = 1729$, que foi igual a $-r_{\mathcal{P}}$ = -0,0259.

Tabela 12: Perda máxima (*MaxLoss)* do portfólio para diferentes valores de p (domínio de admissibilidade) e seu respectivo cenário.

		Cenário			
p	$MaxLoss$	$r_{1,ML}$	$r_{2,ML}$	$r_{3,ML}$	$r_{4,ML}$
0,50	-0,0453	-0,0451	-0,0460	-0,0453	-0,0449
0,55	-0,0475	-0,0473	-0,0482	-0,0475	-0,0471
0,65	-0,0521	-0,0519	-0,0529	-0,0521	-0,0517
0,75	-0,0574	-0,0572	-0,0583	-0,0574	-0,0569
0,85	-0,0642	-0,0640	-0,0652	-0,0642	-0,0637
0,95	-0,0762	-0,0759	-0,0773	-0,0762	-0,0756
0,99	-0,0901	-0,0897	-0,0915	-0,0901	-0,0894

4. Conclusão

Neste artigo tratamos da estimação e análise de medidas de risco para um portfólio formado por ações de quatro empresas brasileiras, sendo elas: Bradesco, Brasil Telecom, Gerdau e Petrobrás. Tais ações são negociadas na Bolsa de Valores de São Paulo (Bovespa). Os valores do portfólio do mercado foram representados pelo índice Bovespa. Os pesos do portfólio eficiente foram estimados através do método da média-variância, introduzido por Markowitz (1952). O modelo de precificação de ativos (CAPM) foi utilizado como ferramenta para análise da correlação dos ativos com o mercado financeiro.

Pelo modelo CAPM, obtivemos $\beta_1 < 1$, o que significa que o ativo A_1 (Bradesco) é defensivo, enquanto que $\beta_i > 1$, para $i = 2,3,4$, o que significa que os ativos A_2 (Brasil Telecom), A_3 (Gerdau) e A_4 (Petrobrás), são agressivos. Portanto, se houverem quedas no índice do mercado, os retornos dos ativos sofrerão quedas maiores (em termos de porcentagem) do que a queda do índice do mercado. De forma análoga, se o índice do mercado sofrer um determinado aumento, o aumento sofrido pelos retornos desses ativos será maior.

Os valores reais observados para os log-retornos $r_{i,1729}$, para cada $i \in \{1, \cdots, 4\}$, das séries temporais analisadas foram, respectivamente, -0,0026, 0,0301, 0,0680 e 0,0021. Comparando com os valores obtidos pelo modelo CAPM notamos que, para o ativo A_1 (Bradesco) o log-retorno observado foi menor que o retorno estimado, enquanto que, para os demais ativos os resultados obtidos através do modelo CAPM foram inferiores aos observados. Para um período $h = 10$ dias os log-retornos de período h, dados por $r_{i,t}[h]$, para $i \in \{1, \cdots, 4\}$, são iguais a -0,0156, -0,0343, 0,2456 e -0,0551, respectivamente. Observamos que, neste caso, o retorno previsto pelo modelo CAPM só foi inferior aos log-retornos observados no caso do ativo A_3. Utilizamos o valor do β_P do portfólio para estimar os log-retornos e obtivemos valores muito próximos embora sempre menores do que os valores observados.

Observamos que tanto o índice do mercado como os preços dos ativos sofreram uma forte queda em seus valores no período em torno de $t = 1000$, aproximadamente 15 de janeiro de 1999. Este período foi marcado pela desvalorização do real. Calculamos o valor do portfólio assumindo que o instante inicial de investimento era igual ao primeiro dia de observação das séries temporais. Em seguida calculamos as perdas do portfólio e observamos que a maior perda (ou seja, o maior ganho, que é igual a aproximadamente R$ 0,37 por cada real investido) corresponde à mudança no valor do portfólio, ocorrida entre 24/10/1997 e 27/10/1997. Nesse período o valor do índice Bovespa passou de 11.545,20 para 9.816,80 pontos, o que representa uma queda de 14,97%. Nessa data também ocorreram grandes quedas nos índices Dow Jones (7,18%) e S&P 500 (6,87%). Este período é marcado pela forte crise na Ásia. O maior ganho no portfólio corresponde a variação ocorrida no período de 13/01/2001 à 14/01/2001. O Ibovespa apresenta uma alta de 2,08% neste período. Enquanto que, para o mercado mundial temos que em 14/01/2000 o valor do índice Dow Jones foi de 11.722,98 pontos. Esse valor só foi superado em 03/10/2006, quando o índice atingiu 11.727,34 pontos.

Para estimar as medidas de risco VaR e ES utilizando os log-retornos do portfólio e observamos que o VaR estimado utilizando-se os processos FIEGARCH foi o que mais se aproximou do valor real da perda observada. Comparando com o valor 0.02217, encontrado utilizando-se o β_P e o log-retorno do Ibovespa, observamos que a abordagem econométrica foi mais precisa.

No teste de estresse, observamos que os log-retornos reais foram superiores aos previstos pelos cenários pessimistas. Entretanto, todos os cenários otimistas superestimaram os log-retornos reais. As perdas que ocorrem nos cenários pessimistas construídos a partir dos dados históricos são maiores do que aquelas estimadas quando consideramos cenários com um certo domínio de admissibilidade.

Conclui-se que o teste de estresse, embora seja uma ferramenta indispensável, deve ser realizado em conjunto com uma análise que leva em conta a heteroscedasticidade e longa dependência da variância condicional. Se o investidor optar apenas pelos testes de estresse correrá menos riscos mas também poderá diminuir sua margem de lucro. A abordagem econométrica mostrou-se satisfatória pois forneceu

estimativas próximas dos valores observados.

Referências bibliográficas

Artzner, P.; F. Delbaen, J. Eber e D. Heath (1999). "Coherent Measures of Risk". Math. Finance, Vol. 9, 203 - 228.

Baillie, R.; T. Bollerslev e H. Mikkelsen (1996). "Fractionally Integrated Generalised Autoregressive Conditional Heteroscedasticity". Journal of Econometrics, Vol. 74, 3-30.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity". Journal of Econometrics, Vol. 31, 307-327.

Bollerslev, T. e H.O. Mikkelsen (1996). "Modeling and Pricing Long Memory in Stock Market Volatility". Journal of Econometrics, Vol. 73, 151-184.

Engle, R.F. (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of Variance of U.K. Inflation". Econometrica, Vol. 50, 987-1008.

Fabozzi, F.J.; S.M. Focardi e P.N. Kolm (2006). Financial Modeling of the Equity Market. From CAPM to Cointegration. New York: John Wiley.

Lintner, J. (1965). "The Valuation of Risk Assets and the Selection of Risk Investments in Stock Portfolios and Capital Budgets". Review of Economics and Statistics, Vol. 47, 13-37.

Lopes, S.R.C. e T.S. Prass (2012). "Theoretical Results on FIEGARCH Processes". Submetido.

Markowitz, H. (1952). "Portfolio Selection". Journal of Finance, Vol. 7(1), 77-91.

McNeil, A.J.; R. Frey e P. Embrechts (2005). Quantitative Risk Manangement. New Jersey: Princeton University Press.

Nelson, D.B. (1991). "Conditional Heteroskedasticity in Asset Returns: A New Approach". Econometrica, Vol. 59, 347-370.

Prass, T.S. (2008). Análise e Estimação de Medidas de Risco em Processos FIEGARCH. Dissertação de Mestrado no Programa de Pós Graduação em Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Sharpe, W.F. (1964). "Capital Asset Prices: a Theory of Market Equilibrium Under Conditions of Risk". Journal of Finance. Vol. 19, 425-442.

Studer G. (1997). "Maximum Loss for Measurement of Market Risk". Tese de Doutorado no Programa de Pós-Graduação em Matemática do Swiss Federal Institute of Technology, Zurich.

Zivot, E. e J. Wang (2005). Modeling Financial Time Series with S-PLUS. New York: Springer-Verlag. 2 edição.

Agradecimentos

T.S. Prass recebeu auxilio do CNPq-Brasil. S.R.C. Lopes recebeu auxílio parcial do CNPq-Brasil, da CAPES-Brasil, do INCT *em Matemática* e também do Pronex *Probabilidade e Processos Estocásticos* - E-26/170.008/2008 - APQ1.

Abstract

The aim of this study is to analyze the performance of the main methods considered in the literature to calculate risk in a portfolio, namely, the stress test, the maximum loss (MaxLoss) and the value-at-risk (VaR), under the presence of heteroskedasticity and long-range dependence in volatility. We consider a portfolio consisting of shares of four companies and use the CAPM model to estimate the weights for the assets in the portfolio. To calculate the VaR, we consider the econometric approach where the conditional variance is modeled by FIEGARCH models. For this analysis we consider a data set collected from January 1995 to December 2001, a period for which the volatility presents long-range dependence and stationarity.

MSC (2000). 60G10, 62M10, 62M20, 97M30, 91B84.

Um Método Hierárquico para a Determinação do Número Ideal de Grupos

Gustavo Silva Semaan,[1]
José André de Moura Brito,[2]
Luiz Satoru Ochi [3]

Resumo.

A área de Cluster Analysis agrega diversos métodos de classificação que podem ser aplicados com o objetivo de identificar grupos em um conjunto de dados. O número de grupos pode ser fixado ou determinado mediante avaliação de algum índice (ou coeficiente). O presente trabalho propõe um novo método de agrupamento hierárquico que foi concebido a partir do estudo do algoritmo Bisecting K-Means com o objetivo de identificar a quantidade ideal de grupos. A qualidade das soluções obtidas é indicada pelo coeficiente Silhueta, que combina coesão e separação. Os resultados apresentados neste estudo indicam que o método proposto é de fácil implementação e competitivo em relação à qualidade das soluções quando comparado com os algoritmos mais conhecidos e eficientes da literatura.

Palavras Chave: Problema de Agrupamento Automático, Agrupamento Hierárquico, Índice Silhueta.

[1] Instituto de Computação - Universidade Federal Fluminense (IC-UFF), gsemaan@ic.uff.br,
2 Escola Nacional de Ciências Estatísticas (ENCE - IBGE), jose.m.brito@ibge.gov.br
3 satoru@ic.uff.br

1. Introdução

A análise de agrupamento agrega um conjunto de métodos que são aplicados à determinação de grupos a partir de um conjunto de objetos definidos por certas características (atributos). O objetivo é obter grupos que apresentem padrões (características) semelhantes e que possam refletir a forma como os dados são estruturados. Para isso, deve-se maximizar a similaridade (homogeneidade) entre os objetos de um mesmo grupo e minimizar a similaridade entre objetos de grupos distintos [Han and Kamber, 2006] [Larose, 2005] [Goldschimidt and Passos, 2005].

Formalmente, o problema clássico de agrupamento pode ser definido da seguinte maneira: dado um conjunto formado por n objetos $X = \{x_1, x_2, ..., x_i, ..., x_n\}$, com cada objeto $x_i \in X$ e possui p atributos (dimensões ou características), ou seja, $x_i = \{x_i^1, x_i^2, ..., x_i^p\}$, deve-se construir k grupos C_j $(j = 1, ..., k)$ a partir de X, de forma a garantir que os objetos de cada grupo sejam homogêneos segundo alguma medida de similaridade. Além disso, devem ser respeitadas as restrições concernentes a cada problema particular abordado [Han and Kamber, 2006] [Ester et al., 1995] [Baum, 1986] [Hruschka and Ebecken, 2001] [Dias and Ochi, 2003]. Apresenta-se, a seguir, o conjunto de restrições que definem o problema clássico de agrupamento:

$$\bigcup_{i=1}^{k} C_i = X \tag{1}$$

$$C_i \cap C_j = \varnothing \quad \text{para} \quad i,j=1, .., k \quad \text{e} \quad i \neq j \tag{2}$$

$$C_i \neq \varnothing \quad \text{para} \quad i=1, .., k \tag{3}$$

Estas restrições determinam, respectivamente, que: O conjunto X corresponde à união dos objetos dos grupos, cada objeto pertence a exatamente um grupo e todos os grupos possuem pelo menos um objeto.

Para este problema, o número de soluções possíveis, ou seja, o total de maneiras em que os n objetos podem ser agrupados, considerando um número fixo de k grupos, é dado pelo número de *Stirling* (*NS*) de segundo tipo [Jr, 1968], e podem ser obtidas pela Equação 4 [Liu, 1968]. Para problemas de agrupamento em que o valor de k é desconhecido (agrupamento automático), o número de soluções possíveis aumenta ainda mais. Este número é dado pela Equação 5, que corresponde ao somatório da Equação 4 para o número de grupos variando no intervalo $k = 1, .., k_{max}$, sendo k_{max} o número máximo de grupos. Para que se tenha uma ideia da ordem de grandeza deste número, no caso de n=10 objetos a serem alocados em k=3 grupos, o número de soluções a serem consideradas é de 9.330. Mas considerando apenas dobro de objetos, ou seja, n=20 e k=3, o número de soluções possíveis (Equação 4) sobe para 580.606.446. No problema de agrupamento automático estes valores crescem exponencialmente com o aumento da quantidade de objetos (n). Esta característica torna proibitiva a obtenção da solução ótima mediante a aplicação de um procedimento de enumeração exaustiva. Esta questão é comentada em vários trabalhos da literatura, como por exemplo, no trabalho de Naldi (2011).

$$NS(n,k)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^{n} \qquad (4)$$

$$NS(n)=\sum_{j=1}^{k_{\max}} NS(n,j) \qquad (5)$$

Conforme Kumar et. Al (2009), as últimas décadas, e em particular os últimos anos Conforme Kumar et. Al (2009), as últimas décadas, e em particular os últimos anos, têm sido marcados pelo desenvolvimento de diversos algoritimos de agrupamento. Por sua vez, estes, estes algoritmos encontram aplicação em diversos domínios, quais sejam: inteligência artificial, reconhecimento de padrões, marketing, economia, ecologia, estatística, pesquisas médicas, ciências políticas, etc. Não obstante, nenhum desses algoritmos é apropriado para todos os tipos de dados, formatos de grupos e aplicações. Esta última observação sugere que há espaço para o estudo e o desenvolvimento de novos algoritmos de agrupamento que sejam mais eficientes ou mais apropriados, levando em conta as características específicas de conjuntos de dados. Em muitos casos, inclusive, a análise de *"o que é uma boa solução"* é subjetiva, tendo em vista as especificidades do problema estudado.

2. Revisão da Literatura

Segundo [Kumar et. al., 2009], talvez um dos problemas de seleção de parâmetros mais conhecido seja o de determinar o número ideal de grupos em um problema de agrupamento. Uma dessas técnicas apresentadas na literatura consiste analisar o valor da Soma dos Erros Quadráticos (SEQ, Equação 6) das soluções obtidas em função do número de grupos. O objetivo é encontrar a quantidade natural de grupos, procurando por uma quantidade de grupos em que exista uma inflexão no valor do SEQ. Essa abordagem pode falhar em algumas situações, quais sejam: quando existem grupos entrelaçados, superpostos ou até mesmo aninhados. Na Equação 6, $dist(c_i, x)$ indica a distância (Euclidiana: Equação 7) entre o objeto x e o centroide (ci) que esteja mais próximo deste objeto.

$$SEQ = \min \sum_{i=1}^{k} \sum_{x \in C_i} dist(c_i, x)^2 \qquad (6)$$

Essa primeira abordagem pode ser realizada, por exemplo, com a aplicação de um algoritmo clássico de agrupamento denominado k-Means *(Han and Kamber, 2006)*, considerando que k é um inteiro que assume todos os valores no intervalo $k = 2,..,n$. Dessa forma, aplica-se o algoritmo de agrupamento para cada valor de k e, em seguida, calcula-se o valor de SEQ para cada uma nas $(n-1)$ soluções obtidas. A partir destes valores, torna-se possível construir um gráfico SEQ versus o número de grupos, conforme apresenta a Figura 1.

É importante destacar que o algoritmo *k-Means* é sensível à seleção de protótipos (objetos ou centróides) iniciais. Ou seja, uma seleção aleatória desses protótipos para a formação dos grupos iniciais do algoritmo geralmente resulta em agrupamentos pobres, de baixa qualidade no que concerne à estrutura ou ao valor da similaridade [Kumar et. al., 2009]. Dessa forma, recomenda-se que para cada numero k de grupos no intervalo estabelecido, esse algoritmo seja executado considerando diferentes protótipos iniciais. Em seguida, são consideradas para a análise somente a melhor solução (menor SEQ) para cada número k de grupos.

Outra abordagem concernente à determinação do número ideal de grupos consiste na avaliação da função Silhueta (apresentada na seção 2.1) proposta por Rousseeuw (1987) e utilizada em diversos trabalhos, dentre os quais: [Naldi, 2011], [Cruz, 2010], [Wang et. al., 2007], [Soares, 2004] e [Tseng and Yang, 2001]. Mais especificamente, aplica-se um algoritmo de agrupamento para alguns valores de k no intervalo $k = 2,..,n$, escolhendo-se como o k ideal aquele associado ao maior valor da função Silhueta (Figura 1). Uma vez que esse trabalho utilizou a função Silhueta, há uma seção específica para a descrição detalhada da mesma.

Ainda em relação à análise da função Índice Silhueta, também é possível executar tal abordagem considerando múltiplas execuções do algoritmo *k-Means*. Novamente, aconselha-se a executar esse algoritmo com diferentes inicializações de protótipos para cada número de grupos e, ao final, deve-se utilizar a função Silhueta para avaliar as soluções obtidas para cada k.

Com base no algoritmo *k-Means*, foi proposto por [Pelleg and Moore, 2000] o algoritmo *X-Means* para a resolução do problema de agrupamento automático. Este algoritmo recebe como parâmetros o problema a ser processada e um intervalo com a quantidade de grupos $k = k_{\min},..,k_{\max}$, e utiliza o Índice BIC (*Bayesian Information Criterion*) para identificar e retornar qual o melhor número de grupos. Em [Zalik, 2008] é apresentado um algoritmo que também adapta o *k-Means* para resolver um problema de agrupamento automático.

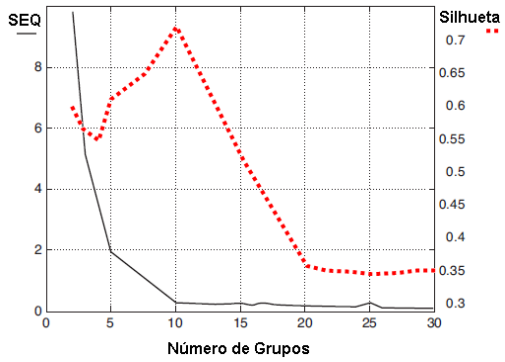


Figura 1: SEQ versus Número de Grupos e Silhueta versus Número de Grupos (adaptação de [Kumar et. al., 2009])

Ainda no contexto do problema agrupamento automáticos vários trabalhos na literatura propõem algoritmos baseados em metaheurísticas que têm por objetivo encontrar um número ideal de grupos e a sua solução correspondente. Dentre estes, destacam-se os seguintes trabalhos: [Soares, 2004] [Cruz, 2010] [Cole, 1998] [Cowgill, 1999] [Bandyopadhyay and Maulik, 2001] [Bandyopadhyay and Maulik, 2002b] [Hruschka and Ebecken, 2003] [Hruschka et. al., 2004a][Hruschka et. al., 2004b] [Hruschka et. al., 2006] [Ma et. al., 2006] [Alves et. al. 2006] [Tseng and Yang, 2001] [Naldi and Carvalho, 2007] [Pan and Cheng, 2007].

Existem, também, as heurísticas que utilizam alguns procedimentos de busca local baseados no algoritmo *k-Means*. Em um primeiro momento, essas heurísticas utilizam algoritmos para construção de grupos, denominados grupos parciais (temporários, componentes conexos) com o objetivo de unir os objetos mais homogêneos. Em seguida, são aplicados algoritmos de busca local e de perturbação sobre esses grupos produzindo soluções de boa qualidade [Cruz, 2010] [Tseng and Yang, 2001] [Hruschka et. al., 2004b] [Alves et. al. 2006] [Hruschka et. al., 2006] [Naldi and Carvalho, 2007] [Naldi, 2011].

Em [Tseng and Yang, 2001] foi apresentado um algoritmo genético denominado CLUSTERING, que também utiliza a função Silhueta para determinar o número ideal de grupos. Para isso, esse algoritmo constrói um grafo, identifica os seus componentes conexos e atua no agrupamento desses componentes com o objetivo de maximizar a função Silhueta.

O trabalho Soares [Soares, 2004] apresenta alguns algoritmos baseados nas metaheurísticas *Simulated Annealing* e Algoritmos Evolutivos para a resolução do problema de agrupamento automático. Este trabalho também propõe algoritmos para construção de soluções, perturbações e refinamentos (buscas locais), incluindo um procedimento de reconexão por caminhos (*path relinking*), que atua na busca de soluções de melhor qualidade. Em seus experimentos foram realizadas algumas comparações com o algoritmo CLUSTERING [Tseng and Yang, 2001].

O algoritmo CLUES (*CLUstEring based on local Shirinking*) [Wang et. al., 2007] também aborda o problema do agrupamento automático, possibilitando a aplicação da função Silhueta ou do Índice CH (Índice de *Calinski-Harabasz*) para a determinação do número ideal de grupos. Trata-se de um algoritmo iterativo que, com a utilização de um procedimento de encolhimento (*Shirinking procedure*) baseado nos k-vizinhos mais próximos, realiza a união dos objetos mais homogêneos segundo os seus atributos.

Após a aplicação do procedimento de encolhimento, o CLUES constrói soluções, avaliando-as mediante o valor da função de Silhueta ou do Índice CH. Ainda em Wang et. al. (2007) é relatado que os resultados obtidos com a utilização da função de Silhueta e do Índice CH foram comparados. A partir dessa comparação, observou-se que mediante a aplicação do Índice Silhueta foram produzidas soluções de melhor qualidade no que concerne ao número de grupos definidos e à formação de soluções denominadas *perfeitas* em tal trabalho. Esse algoritmo foi desenvolvido em R e o seu código fonte está disponível em um pacote do software estatístico R.

O trabalho de Cruz [Cruz, 2010] traz uma proposta de algoritmos heurísticos mais sofisticados no que concerne aos procedimentos de construção, de busca local e de perturbação. Mais especificamente, estes algoritmos foram baseados nas metaheurísticas Algoritmos Genéticos, Busca Local Iterada *(Iterated Local Search)* e GRASP *(Greedy Randomized Adaptive Search Procedure).* O diferencial desses algoritmos está na incorporação de procedimentos para a construção de grupos parciais, conceitos de Memória Adaptativa e Buscas Locais que utilizam conceitos do algoritmo k-means para a união de grupos parciais.

Ainda no trabalho de Cruz (2010) foram propostos também métodos híbridos. Estes métodos utilizam algumas das soluções produzidas pelos algoritmos heurísticos, ou seja, soluções associadas com alguns valores de k e que sejam consideradas promissoras no que concerne ao número ideal de números, porém não necessariamente a melhor solução para tal número. Considerando estes valores específicos de k, são aplicadas duas formulações de programação inteira, quais sejam: para o problema de agrupamento com diâmetro mínimo e dos k-*Medoids*. Nos experimentos apresentados em seu trabalho foram realizadas comparações com o algoritmo da literatura CLUES [Wang et. al., 2007].

O *Bisecting k-Means*, proposto por [Steinbach et. al., 2000], corresponde a uma versão hierárquica do k-*Means*, em que a cada iteração, um grupo é selecionado e dividido em dois novos grupos. Nesse contexto, o presente trabalho propõe algoritmos baseados no *Bisecting K-Means*. Esses algoritmos diferenciam-se, quanto ao critério utilizado para a seleção do grupo a ser particionado. Após a seleção do grupo a ser particionado, é necessário distribuir os objetos desse grupo entre os dois novos grupos, utilizando-se, nessa fase, o algoritmo K-*Means* clássico. Cada solução obtida a partir da aplicação dos algoritmos propostos, para cada quantidade de *grupos* k no intervalo $k = kmin, .., kmax$, é avaliada com a utilização do índice de validação relativo Silhueta. Conforme Naldi [Naldi, 2011], os índices de validação relativos têm sido utilizados e investigados extensivamente, tendo apresentado resultados satisfatórios em diversas aplicações.

Os índices relativos, como próprio nome sugere, têm como finalidade avaliar a qualidade relativa das soluções produzidas por diferentes métodos de agrupamento. Estes índices não têm a propriedade de monotonicidade, ou seja, não são afetados pelo aumento ou pela redução do número de grupos da solução. Dessa forma, podem ser utilizados na avaliação de diversas soluções, provenientes de diversos algoritmos. No presente trabalho, assim como nos algoritmos da literatura considerados nos experimentos, as soluções obtidas são avaliadas pelo Índice de Silhueta.

2.1. Índice Silhueta

O Índice Silhueta foi proposta por Rousseeuw [Rousseeuw, 1987]. Esta medida determina a qualidade das soluções com base na proximidade entre os objetos de determinado grupo e na distância desses objetos ao seu grupo mais próximo. O Índice Silhueta é calculado para cada objeto, sendo possível identificar se o objeto está alocado ao grupo mais adequado. Esse índice combina as ideias de coesão e de separação. Os quatro passos a seguir explicam, brevemente, como calculá-lo:

1. Nesse trabalho dij (Equação 7) corresponde à distância euclidiana entre os objetos i e j, e p é a quantidade de atributos dos objetos. Para cada objeto x_i calcula-se a sua distância média $a(x_i)$ (Equação 8) em relação a todos os demais objetos do mesmo grupo. Na Equação 8, $|C_w|$ representa a quantidade de objetos do grupo C_w , ao qual o objeto x_i pertence.

$$d_{ij}=\sqrt{\sum_{q=1}^{p}(x_i{}^q-x_j{}^q)^2} \tag{7}$$

$$a(x_i)=\frac{1}{|C_w|-1}\sum d_{ij} \qquad \forall x_j\neq x_i\ , \qquad x_j\in C_w \tag{8}$$

2. A Equação 9 apresenta a distância entre o objeto x_i e os objetos do grupo C_t , em que $|C_t|$ é a quantidade de objetos do grupo C_t. Para cada objeto x_i calcula-se a sua distância média em relação a todos os objetos dos demais grupos ($b(x_i)$) (Equação 10).

$$d(x_i,C_t)=\frac{1}{|C_t|}\sum d_{ij} \qquad \forall x_j\in C_t \tag{9}$$

$$b(x_i)=\min d(x_i,C_t) \qquad C_t\neq C_w \qquad C_t\in C \tag{10}$$

3. O coeficiente Silhueta do objeto x_i ($s(x_i)$) pode ser obtido pela Equação 11.

$$s(x_i)=\frac{b(x_i)-a(x_i)}{\max\{b(x_i),a(x_i)\}} \tag{11}$$

O cálculo da Silhueta de uma solução S é a média das Silhuetas de cada objeto, conforme apresenta a Equação 12, em que n é a quantidade de objetos da solução. Essa função deve ser maximizada.

$$Silhueta(S)=\max\frac{1}{n}\sum_{i=1}^{n} s(x_i) \qquad (12)$$

Os valores positivos de Silhueta indicam que o objeto está bem localizado em seu grupo, enquanto valores negativos indicam que o objeto está mais próximo de outro(s) grupo(s). A Figura 2 apresenta um exemplo gráfico de uma solução constituída por dez grupos e objetos com duas dimensões. As cores dos objetos indicam as suas Silhuetas e, quanto mais escuro o tom de cinza, menor o valor da Silhueta (próximo de zero). Observa-se que nesse exemplo nenhum objeto possui a Silhueta negativa.

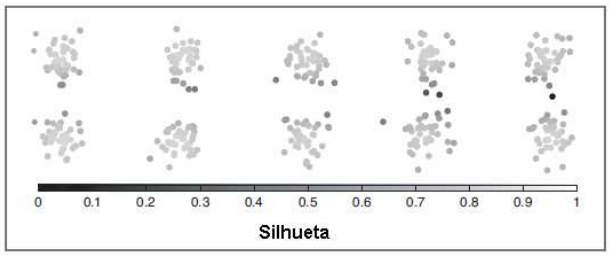


Figura 2: Problema associado ao gráfico da Figura 1 ([Kumar et. al., 2009]).

Conforme Naldi [Naldi, 2011], este índice é mais apropriado para agrupamentos volumétricos, com grupos gerados de acordo com distribuições Gaussianas multidimensionais hiperesféricas ou moderadamente alongadas, porém ele não obteve bons resultados para grupos com formatos arbitrários [Rousseeuw, 1987].

Em [Hruschka et. al., 2004a] é proposta uma versão simplificada do Índice de Silhueta. Nesta versão são efetuadas modificações nos cálculos de $a(x_i)$ e $b(x_i)$ com o objetivo de reduzir a complexidade do algoritmo de $O(n^2)$ para $O(n)$. Segundo os autores desse trabalho, mesmo com a redução da complexidade, esse novo índice mantém a qualidade próxima ao da Silhueta tradicional, o que é confirmado por [Vendramin et. al., 2009] [Vendramin et. al., 2010].

3. Algoritmo Proposto Baseado em Bisecting K-Means

Com o objetivo de identificar o número ideal de grupos em cada problema, o presente trabalho propõe algoritmos baseados no *Bisecting k-Means* [Steinbach et. al. (2000)]. Esse algoritmo corresponde a uma versão hierárquica do algoritmo *k-Means*, em que a cada iteração ocorre a seleção de um dos grupos a ser particionado e, em seguida, a distribuição de seus objetos nos dois novos grupos. A Figura 3 ilustra uma representação de um Agrupamento Hierárquico utilizando um Dendrograma. Nesse sentido, cada linha horizontal tracejada indica a seleção de um grupo, que é dividido em dois novos grupos. A seguir são relacionadas as soluções obtidas com base na Figura 3.

- Seleção 1: o grupo C1 = {a,b,c,d,e} é selecionado e dividido em dois novos grupos. Forma-se a solução com dois grupos C1 = {a,b,c} e C2 = {d,e}.
- Seleção 2: o grupo C1 = {a,b,c} selecionado e dividido em dois novos grupos C1 = {a,b} e C3 = {c}. Forma-se a solução com três grupos C1 = {a,b}, C2 = {d,e} e C3 = {c}.
- Seleção 3: o grupo C1 = {a,b} é selecionado e dividido em dois novos grupos C1 = {a} e C4 = {b}. Forma-se a solução com quatro grupos C1 = {a}, C2 = {d,e}, C3 = {c} e C4 = {b}.
- Seleção 4: o grupo C2 = {d,e} é selecionado e dividido em dois novos grupos C5 = {d} e C2 = {e}. Forma-se a solução com cinco grupos C1 = {a}, C2 = {e}, C3 = {c}, C4 = {b} e C5 = {d}.

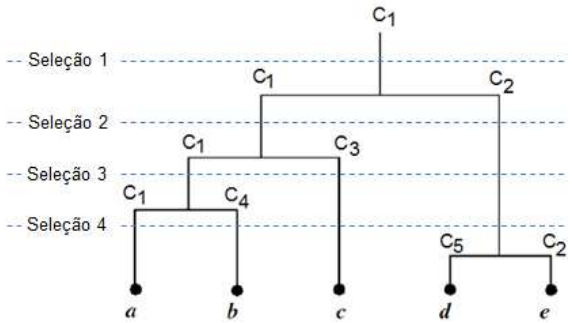


Figura 3: Representação de um Agrupamento Hierárquico utilizando um Dendrograma.

O funcionamento do algoritmo *Bisecting k-Means* pode ser resumido nos seguintes passos:

Inicialmente, todos os objetos pertencem ao mesmo grupo, ou seja, a solução inicial possui apenas um grupo.

Deve ser selecionado um grupo Cs que será particionado. Nesse trabalho são apresentados cinco critérios para seleção desse grupo.

Os objetos do grupo Cs devem ser distribuídos em dois novos grupos por meio da utilização do algoritmo

1. *K-Means*, para $K = 2$. Com o objetivo de obter uma melhor distribuição dos objetos, o algoritmo *K-Means* pode ser executado no grupo C_s várias vezes com diferentes sementes.

2. Os passos 2 e 3 devem ser repetidos até que a quantidade de grupos desejada seja alcançada (k_{max}).

3. Os passos 2 e 3 devem ser repetidos até que a quantidade de grupos desejada seja alcançada (k_{max}).

Com base nos passos do algoritmo *Bisecting K-Means* observa-se que são obtidas soluções para todos os valores de k pertencentes a um intervalo de k pré-estabelecido $k=kmin, .., kmax$. Em geral, os parâmetros *kmin e kmax* são definidos em função da aplicação que será abordada. Mais especificamente, o parâmetro *kmin* poder ser definido a priori pelo pesquisador, considerando a sua necessidade de segmentar os seus dados em um número mínimo de grupos. Já o parâmetro *kmax* é normalmente limitado pelo ganho na qualidade da solução, isto é, pela redução obtida no valor da função objetivo (SEQ). Um exemplo disso é o problema de estratificação estatística [Brito et. al. 2009], em que não há uma redução significativa no valor da função objetivo para o número de grupos superior a oito.

O presente trabalho propõe cinco critérios para a seleção do grupo a ser particionado (passo 2 do algoritmo acima), sejam eles:

1. Menor Silhueta: o grupo que possuir menor valor de Silhueta média (Equação 12).
2. Maior Diâmetro: o diâmetro de um grupo C_j corresponde à maior distância entre dois objetos (dissimilaridade) desse grupo (Equação 13). Dessa forma esse critério consiste em selecionar o grupo que possuir o maior valor de diâmetro.

$$d(Cj) = \underset{k,l : x_k, x_l \in C_j}{Max} d_{kl} \quad (13)$$

3. Maior Clique: o clique de um grupo C_j corresponde à soma das distâncias entre todos os pares de objetos de C_j (Equação 14). Assim, nesse critério deve ser selecionado o grupo que possuir o maior valor de *clique*.

$$cl(Cj) = \sum_{k,l:x_k, x_l \in C_j} d_{kl} \quad (14)$$

4. Maior Clique Médio: nesse critério deve-se utilizar o valor do clique do grupo C_j (critério 3) dividido pelo total de combinações obtidas considerando que os objetos do grupo C_j sejam tomados dois a dois. Esse total é dado pela Equação 15, sendo n_j a quantidade de objetos do grupo C_j.

$$combinacoes(Cj) = \frac{n_j(n_j - 1)}{2} \quad (15)$$

5. Maior Soma dos Erros Quadráticos: o grupo que possuir a maior soma das distâncias entre os seus objetos e o seu respectivo centróide (Equação 6).

A partir da descrição acima, o novo método proposto nesse trabalho pode ser resumido aos seguintes passos:

1. Inicialmente todos os objetos pertencem ao mesmo grupo.
2. Deve ser selecionado um grupo C_s a ser particionado utilizando um dos cinco critérios supracitados.
3. Os objetos do grupo C_s devem ser distribuídos em dois novos grupos por meio da utilização do algoritmo *K-Means*. Uma vez que trata-se de um algoritmo heurístico e a seleção de protótipos iniciais podem determinar a qualidade das soluções obtidas, o *K-Means* é executado dez vezes, com diferentes protótipos iniciais, com o objetivo de obter uma melhor distribuição de objetos do grupo selecionado entre os dois novos grupos. A quantidade de execuções foi identificada com a realização de experimentos preliminares. Após as dez divisões do grupo selecionado, somente a que resulta na melhor solução, ou seja, aquela que possuir o maior Índice Silhueta é armazenada em um vetor de soluções denotado por *VS*.

4. Os passos 2 e 3 devem ser repetidos até que o número máximo de grupos seja alcançado (*kmax*).
5. O vetor *VS* possui a melhor solução obtida pelo algoritmo para cada número *k* de grupos variando no intervalo estabelecido previamente. Para identificar a quantidade ideal de grupos do problema submetido ao método, deve-se verificar a solução existente no vetor *VS* que possua o maior Índice Silhueta. A utilização de vetor *VS* possibilita análises sobre os resultados obtidos bem como a identificação não somente do melhor resultado (considerado número ideal), mas também de outros números de grupos que também indicam soluções de boa qualidade quanto à estrutura dos grupos.

De a forma a identificar as cinco versões associadas aos critérios de seleção do novo método, cada versão inicia com as letras "BK" seguida do número identificador do critério de seleção. Por exemplo, a versão BK1 *(Bisecting K-Means 1)* corresponde ao algoritmo cuja seleção do grupo a ser particionado utiliza o critério 1 (menor Silhueta média).

Após a seleção do grupo a ser dividido, a distribuição de seus elementos é feita com a utilização do clássico algoritmo *K-Means*. Nesse algoritmo, a seleção de protótipos iniciais é determinante para a formação de soluções boa qualidade. Uma vez que o número de grupos *k* no método proposto varia no intervalo *k=kmin*, .., *kmax* e a aplicação das cinco versões do novo método proposto, são produzidas *(kmin - kmax + 1) * 5* soluções. Todas essas soluções são avaliadas mediante a utilização do índice de validação relativo de Silhueta. A solução associada ao maior valor do Índice Silhueta possui a quantidade ideal de grupos.

4. Experimentos Computacionais

A presente seção traz um conjunto de resultados computacionais obtidos a partir da aplicação de alguns algoritmos da literatura e dos algoritmos propostos nesse trabalho. Observa-se que os algoritmos da literatura foram implementados utilizando diferentes linguagens de programação, compiladores e foram executados em diferentes máquinas e sistemas operacionais. Além disso, alguns códigos fonte da literatura não estavam disponíveis até o momento da produção desse trabalho. Em face destas observações, a comparação entre os algoritmos da literatura e das cinco versões utilizados no método proposto, no que concerne à sua performance, ficou restrita à qualidade das soluções com base na função Silhueta e ao número de grupos ideal fornecido pelos algoritmos.

As implementações dos algoritmos propostos foram feitas em Linguagem C++, utilizando o paradigma orientação à objetos, em um ambiente de desenvolvimento Eclipse for C/C++ Developers. Todos os experimentos computacionais foram realizados em um computador dotado de um processador i7 de 3.0 GHz e com 8GB de RAM e o sistema operacional Ubuntu 9.10, kernel 2.6.18. Os problemas considerados nos experimentos do presente trabalho estão disponíveis no endereço *http://labic.ic.uff.br/Instance/index.php*, arquivo CP.zip.

É importante destacar que não foi explorada a capacidade de multiprocessamento do equipamento utilizado e não foi utilizado nenhum conhecimento prévio sobre os problemas ou resultados obtidos por outros trabalhos da literatura. Os algoritmos da literatura para os quais resultados foram apresentados e comparados foram os seguintes:

- **CLUES** *(CLUstEring based on local Shirinking)* [Wang et. al., 2007]: implementado no software estatístico R [Matloff 2011] e disponível no pacote *clues*.
- **CLUSTERING** [Tseng and Yang, 2001]: implementação de um Algoritmo Genético em C++ realizada por [Soares, 2004].
- **SAPCA** *(Simulated Annealing)* e AEC-RC (Algoritmo Evolutivo com Reconexão por Caminhos): proposto e implementado em C++ por [Soares, 2004].
- **AECBL1** (Algoritmo Evolutivo com Busca Local), **GBLITRC1** (GRASP com Reconexão de Caminhos) e **IBLITRC1** (Busca Local Iterada com Reconexão de Caminhos) de [Cruz, 2010]: os melhores resultados obtidos para cada problema considerando os três algoritmos. Desenvolvimento feito em linguagem C++.

Em relação ao intervalo relacionado com o número de grupos, é uma prática comum em abordagens sistemáticas utilizar $k = 2,..,k_{max}$, sendo $k_{max} = n^{1/2}$ ([Pal and Bezdek, 1995][Pakhira et. al., 2005][Campello et. al., 2009]. Esse intervalo foi utilizado nos algoritmos propostos.

Para a realização dos experimentos foram utilizados 83 problemas da literatura que estão distribuídas em três conjuntos (DS - Datasets). Estes problemas possuem um número de objetos variando entre 30 e 2000, o número de dimensões (atributos) entre 2 e 60 e diferentes características relacionadas, por exemplo, com a coesão, à separação e às densidades dos grupos.

O primeiro conjunto (DS1), apresentado pela Tabela 1, possui nove problemas conhecidos da literatura com a quantidade de objetos entre 75 e 1484 e dimensões (quantidade de atributos) entre 2 e 13 [Fisher, 1936][Ruspini, 1970][Maronna and Jacovkis, 1974][Wang et. al., 2007][Hastie et. al., 2001][Naldi, 2011].

Tabela 1: Conjunto de Problemas DS1

Problema	Nº Objetos	Dimensão
200DATA	200	2
gauss9	900	2
iris	150	4
maronna	200	2
ruspini	75	2
spherical_4d3c	400	3
vowel2	528	2
wine	178	13
yeast	1484	7

O segundo conjunto (DS2), apresentado na Tabela 2, possui 51 problemas que foram construídos por [Cruz, 2010] utilizando a ferramenta Dots desenvolvida por [Soares and Ochi, 2004]. Esses problemas possuem quantidades de objetos entre 100 e 2000, sendo todos com duas dimensões e o número de grupos entre 2 e 27. Nesse conjunto os nomes dos problemas foram definidos conforme a quantidade de objetos, de grupos, e se os grupos são bem definidos, coesos e separados (denominados *"comportados"* em Cruz (2010)). A Figura 4 apresenta, respectivamente, os problemas 200p4c e 300p4c1, em que 200p4c indica um problema *"comportado"* com 200 objetos e 4 grupos e o problema 300p4c1 indica um problema *"não comportado"* com 300 objetos e 4 grupos.

Tabela 2: Conjunto de Problemas DS2

Problema	Nº Objetos	Problema	Nº Objetos	Problema	Nº Objetos
100p10c	100	300p3c	300	800p23c	806
100p2c1	100	300p3c1	300	900p12c	900
100p3c	100	300p4c1	300	900p5c	900
100p3c1	100	300p6c1	300	1000p14c	1000
100p7c	100	400p3c	400	1000p5c1	1000
100p8c1	106	400p4c1	400	1000p6c	1000
100p5c1	110	400p17c1	408	1000p27c1	1005
100p7c1	112	500p19c1	500	1100p6c1	1100
200p2c1	200	500p3c	500	1300p17c	1300
200p3c1	200	500p4c1	500	1500p6c1	1500
200p4c	200	500p6c1	500	1800p22c	1800
200p4c1	200	600p15c	600	1900p24c	1901
200p7c1	210	600p3c1	600	2000p11c	2000
200p8c1	212	700p4c	700	2000p26c	2000
200p12c1	222	700p15c1	703	2000p9c1	2000
300p13c1	235	800p10c1	800		
300p10c1	300	800p18c1	800		
300p2c1	300	800p4c1	800		

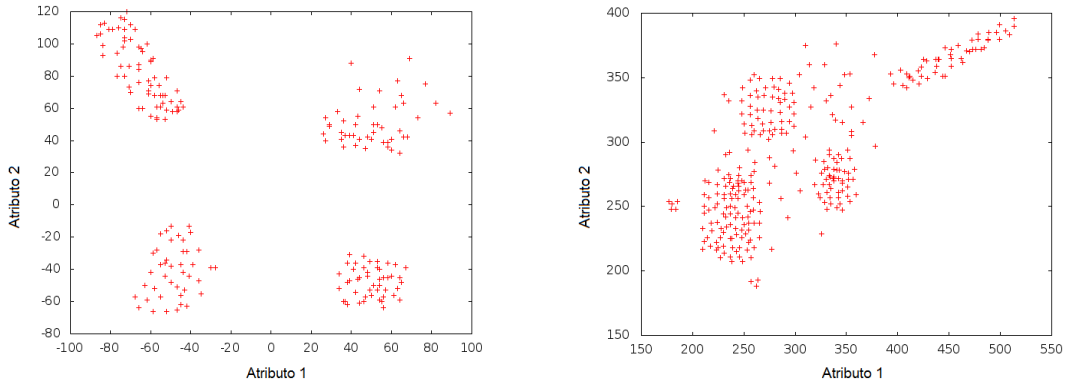


Figura 4: ilustrações dos problemas 200p4c e 300p4c1.

O terceiro conjunto (DS3), apresentado pela Tabela 3, possui 11 problemas que foram construídos e utilizados por [Soares and Ochi, 2004][Soares, 2004]. Esses problemas possuem a quantidade de objetos entre 30 e 2000, sendo todos com duas dimensões.

Tabela 3: Conjunto de Problemas DS3

Problema	Nº Objetos	Problema	Nº Objetos
30p	30	300p4c	300
outliers_ags	80	350p5c	350
97p	97	numbers	437
3dens	128	450p4c	450
Outliers	150	moreshapes	489
157p	157	500p3c	500
convdensity	175	numbers2	540
181p	181	600p3c	600
convexo	199	900p5c	900
2face	200	1000p6c	1000
Face	296	2000p11c	2000

No primeiro experimento são apresentados os resultados obtidos com a execução das cinco versões do novo algoritmo proposto nos três conjuntos de dados. Nesse experimento é apresentado o valor do Índice Silhueta para cada um dos problemas. As Tabelas 4, 5, 6 e 7 apresentam os melhores resultados em relação às cinco versões do algoritmo proposto para, respectivamente, os problemas dos conjuntos DS1, DS2 parte 1, DS2 parte 2 e DS3.

Tabela 4: Melhores Resultados Produzidos pelo Método Proposto Considerando o Conjunto DS1

	Índice Silhueta				
	BK1	**BK2**	**BK3**	**BK4**	**BK5**
200DATA	0,823	0,823	0,763	0,823	0,823
gauss9	0,415	0,356	0,417	0,420	0,419
iris	0,687	0,687	0,687	0,687	0,687
maronna	0,575	0,562	0,575	0,575	0,575
ruspini	0,738	0,641	0,738	0,738	0,738
spherical_4d3c	0,689	0,645	0,689	0,689	0,689
vowel2	0,393	0,391	0,402	0,402	0,402
wine	0,660	0,660	0,660	0,660	0,660
yeast	0,572	0,572	0,572	0,572	0,572

Tabela 5: Resultados Produzidos pelo Método Proposto Considerando o Conjunto DS2 Parte 1

	Índice Silhueta				
	BK1	**BK2**	**BK3**	**BK4**	**BK5**
1000p14c	0,806	0,492	0,806	0,831	0,831
1000p27c1	0,478	0,478	0,478	0,488	0,482
1000p5c1	0,634	0,427	0,633	0,633	0,633
1000p6c	0,736	0,554	0,736	0,736	0,736
100p10c	0,762	0,551	0,762	0,762	0,762
100p2c1	0,743	0,743	0,743	0,743	0,743
100p3c	0,786	0,786	0,786	0,786	0,786
100p3c1	0,574	0,468	0,574	0,574	0,574
100p5c1	0,688	0,528	0,688	0,688	0,688
100p7c	0,834	0,638	0,834	0,834	0,834
100p7c1	0,475	0,475	0,475	0,475	0,475
100p8c1	0,527	0,527	0,527	0,527	0,527
1100p6c1	0,650	0,526	0,645	0,645	0,645
1300p17c	0,804	0,442	0,804	0,823	0,823
1500p6c1	0,636	0,532	0,636	0,636	0,636
1800p22c	0,733	0,457	0,722	0,785	0,747
1900p24c	0,750	0,454	0,738	0,799	0,772
2000p11c	0,713	0,516	0,685	0,713	0,713
2000p26c	0,690	0,431	0,695	0,732	0,734
2000p9c1	0,596	0,502	0,596	0,613	0,616
200p12c1	0,563	0,525	0,564	0,564	0,564
200p2c1	0,764	0,764	0,764	0,764	0,764
200p3c1	0,675	0,675	0,675	0,675	0,675
200p4c	0,773	0,523	0,773	0,773	0,773
200p4c1	0,745	0,745	0,745	0,745	0,745
200p7c1	0,555	0,541	0,555	0,555	0,555

Tabela 6: Resultados Produzidos pelo Método Proposto Considerando o Conjunto DS2 Parte 2

	Índice Silhueta				
	BK1	**BK2**	**BK3**	**BK4**	**BK5**
200p8c1	0,556	0,539	0,556	0,556	0,556
300p10c1	0,580	0,570	0,570	0,583	0,583
300p13c1	0,545	0,545	0,546	0,546	0,546
300p2c1	0,778	0,778	0,778	0,778	0,778
300p3c	0,766	0,766	0,766	0,766	0,766
300p3c1	0,674	0,642	0,674	0,674	0,674
300p4c1	0,607	0,607	0,607	0,607	0,607
300p6c1	0,617	0,548	0,596	0,610	0,610
400p17c1	0,514	0,514	0,514	0,514	0,514
400p3c	0,799	0,673	0,799	0,799	0,799
400p4c1	0,605	0,541	0,605	0,605	0,605
500p19c1	0,474	0,474	0,474	0,474	0,474
500p3c	0,825	0,756	0,825	0,825	0,825
500p4c1	0,663	0,663	0,663	0,663	0,663
500p6c1	0,617	0,460	0,617	0,617	0,617
600p15c	0,742	0,420	0,742	0,775	0,775
600p3c1	0,721	0,687	0,721	0,721	0,721
700p15c1	0,609	0,385	0,607	0,642	0,618
700p4c	0,797	0,543	0,797	0,797	0,797
800p10c1	0,468	0,468	0,468	0,468	0,468
800p18c1	0,661	0,422	0,665	0,688	0,689
800p23c	0,787	0,432	0,752	0,787	0,764
800p4c1	0,696	0,578	0,696	0,696	0,696
900p12c	0,841	0,495	0,841	0,841	0,841
900p5c	0,707	0,470	0,707	0,707	0,707

Tabela 7: Resultados Produzidos pelo Método Proposto Considerando o Conjunto DS3

	Índice Silhueta				
	BK1	**BK2**	**BK3**	**BK4**	**BK5**
1000p6c	0,736	0,554	0,736	0,736	0,736
157p	0,665	0,518	0,665	0,665	0,665
181p	0,714	0,714	0,714	0,714	0,714
2000p11c	0,713	0,516	0,685	0,713	0,713
2face	0,667	0,667	0,667	0,667	0,667
300p4c	0,750	0,643	0,750	0,750	0,750
30p	0,589	0,472	0,527	0,589	0,589
350p5c	0,759	0,600	0,759	0,759	0,759
3dens	0,762	0,762	0,762	0,762	0,762
450p4c	0,766	0,712	0,766	0,766	0,766
500p3c	0,825	0,756	0,825	0,825	0,825
600p3c	0,751	0,565	0,751	0,751	0,751
900p5c	0,707	0,470	0,707	0,707	0,707
97p	0,709	0,541	0,709	0,709	0,709
convdensity	0,854	0,854	0,854	0,854	0,854
convexo	0,668	0,669	0,646	0,668	0,646
face	0,503	0,412	0,493	0,495	0,496
moreshapes	0,728	0,606	0,732	0,713	0,732
numbers	0,562	0,457	0,560	0,562	0,560
numbers2	0,594	0,516	0,600	0,600	0,600
outliers	0,635	0,635	0,635	0,635	0,635
outliers_ags	0,754	0,711	0,754	0,754	0,754

Na Tabela 8 são apresentas estatísticas de cada algoritmo por conjuntos de problemas em relação aos valores de Índice Silhueta alcançados. Com base na Tabela 8 é possível observar que o valor da Silhueta foi positivo e maior ou igual a 0,3 para todos os problemas de todos os conjuntos de dados. Esses resultados indicam que os grupos têm uma boa estrutura [Rousseeuw, 1987]. Nessa tabela, a coluna Todos apresenta estatísticas em relação aos resultados obtidos com execução dos cinco algoritmos hierárquicos propostos.

Tabela 8: Resumo dos Resultados Produzidos pelo Método Proposto por conjunto de dados

		Índice Silhueta					
		Todos	**BK1**	**BK2**	**BK3**	**BK4**	**BK5**
DS1	**Média**	**0,612**	0,617	0,593	0,611	0,618	0,618
	Mediana	**0,660**	0,660	0,641	0,660	0,660	0,660
	Menor	**0,356**	0,393	**0,356**	0,402	0,402	0,402
	Maior	**0,823**	**0,823**	**0,823**	0,763	**0,823**	**0,823**
DS2	**Média**	**0,651**	0,673	0,554	0,671	0,679	0,677
	Mediana	**0,663**	0,688	0,528	0,685	0,688	0,689
	Menor	**0,385**	0,468	**0,385**	0,468	0,468	0,468
	Maior	**0,841**	**0,841**	0,786	**0,841**	**0,841**	**0,841**
DS3	**Média**	**0,680**	0,700	0,607	0,695	0,700	0,699
	Mediana	**0,709**	0,713	0,603	0,712	0,713	0,713
	Menor	**0,412**	0,503	**0,412**	0,493	0,495	0,496
	Maior	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**

A Tabela 9 apresenta uma sumarização dos resultados apresentados pela Tabela 8, sem discriminar o conjunto de dados. É possível observar que a versão BK4 foi superior no que concerne a todas as medidas estatísticas analisadas em relação à qualidade das soluções obtidas. Ou seja, os maiores valores de Índice Silhueta (considerando a média, a mediana, o menor e o maior valor). Além disso, essa versão não obteve o melhor resultado em 9 dos 83 problemas, porém, em relação aos valores de Silhuetas médias obtidos, a maior diferença foi de apenas 0,019 e a média das diferenças entre o melhor resultado e os resultados da versão BK4 foi inferior a 0,001.

Tabela 9: Resumo dos Resultados Produzidos pelo Método Proposto Considerando todos os Conjuntos de dados

	Índice Silhueta				
	BK1	**BK2**	**BK3**	**BK4**	**BK5**
Média	**0,674**	**0,573**	**0,671**	**0,678**	**0,677**
Mediana	**0,690**	**0,544**	**0,687**	**0,693**	**0,693**
Menor	**0,393**	**0,356**	**0,402**	**0,402**	**0,402**
Maior	**0,854**	**0,854**	**0,854**	**0,854**	**0,854**

Já a Tabela 10 apresenta os percentuais relacionados ao alcance à melhor solução obtida por algoritmo. O algoritmo BK1, por exemplo, alcançou os melhores resultados em 77,8% dos problemas do conjunto BK1 (sete dos nove problemas do DS1), enquanto o algoritmo BK4 obteve o melhor resultado para todos os problemas do conjunto DS1 e em 89% dos problemas utilizados no presente trabalho. O algoritmo BK4 foi superior também em cada conjunto de dados individualmente, em que houve empate apenas com os algoritmos BK1 e BK5 em relação ao conjunto DS3.

Tabela 10: Percentuais de soluções equivalentes à melhor solução obtida por algoritmo.

	BK1	BK2	BK3	BK4	BK5
DS1	77,8%	44,4%	77,8%	100,0%	88,9%
DS2	72,5%	27,5%	68,6%	88,2%	84,3%
DS3	86,4%	27,3%	77,3%	86,4%	86,4%
Todos	76,8%	29,3%	72,0%	89,0%	85,4%

As Tabelas 11 e 12 apresentam comparativos entre os melhores resultados obtidos pelo método proposto nesse trabalho em relação aos resultados obtidos por algoritmos da literatura. Mais especificamente, a Tabela 11 traz uma comparação entre os melhores resultados produzidos pelo algoritmo CLUSTERING proposto por [Tseng and Yang, 2001], os dois algoritmos propostos por [Soares, 2004] (SAPCA (*Simulated Annealing*) e AEC-RC (Algoritmo Evolutivo com Reconexão por Caminhos) e novo método proposto nesse trabalho.

Em relação aos comparativos realizados com o CLUSTERING, apresentados na Tabela 11, os resultados obtidos pelos algoritmos desse trabalho foram superiores em todos os problemas. Já nos comparativos com os melhores resultados entre os algoritmos propostos por [Soares, 2004], também relacionados na Tabela 11, os algoritmos do método proposto foram superiores em dez problemas. Além disso, os resultados foram inferiores em três problemas, com uma maior diferença entre as Silhuetas de apenas 0,008. É importante ressaltar que para esses três problemas em que os algoritmos propostos não alcançaram os resultados da literatura, o número de grupos foi o mesmo ou a diferença foi de apenas uma unidade.

Tabela 11: Comparativo de Resultados com algoritmos da literatura.

	TZENG and YANG	SOARES				Novo método Hierárquico	
Problema	(CLUSTERING) Índice Silhueta	(SAPCA) Índice Silhueta	(AEC-RC) Índice Silhueta	(BEST) Índice Silhueta	K	Índice Silhueta	k
200Data	0,541	0,823	0,823	0,823	3	0,823	5
Iris	0,601	0,686	0,686	0,686	3	0,687	2
Ruspini	0,55	0,737	0,737	0,737	4	0,738	6
1000p6c	0,367	0,735	0,727	0,735	6	0,736	8
157p	0,657	0,667	0,667	0,667	4	0,665	3
2000p11c	0,287	0,658	0,611	0,658	11	0,713	12
2face	0,513	0,666	0,666	0,666	2	0,667	2
350p5c	0,568	0,758	0,758	0,758	5	0,759	7
3dens	0,742	0,762	0,762	0,762	2	0,762	2
97p	0,706	0,71	0,71	0,71	3	0,709	3
Convdensity	0,818	0,854	0,854	0,854	3	0,854	3
Convexo	0,618	0,667	0,667	0,667	3	0,669	3
Face	0,402	0,511	0,511	0,511	3	0,503	2
Moreshapes	0,436	0,731	0,725	0,731	6	0,732	9
Numbers	0,417	0,546	0,542	0,546	10	0,562	2
Numbers2	0,513	0,527	0,565	0,565	10	0,600	2

A Tabela 12 apresenta comparativos entre os melhores resultados obtidos pelo método proposto em relação aos resultados obtidos pelos algoritmos da literatura, quais sejam: CLUES *(CLUstEring based on local Shirinking)* [Wang et. al., 2007], AECBL1 (Algoritmo Evolutivo com Busca Local), GBLITRC1 (GRASP com Reconexão de Caminhos) e IBLITRC1 (Busca Local Iterada com Reconexão de Caminhos) [Cruz, 2010].

Com base nos resultados apresentados na Tabela 12, as soluções foram equivalentes ou superiores aquelas relatadas na literatura em cerca de 50% dos problemas. Dessa forma, com o objetivo de identificar as limitações do novo método, bem como características dos problemas, os resultados foram analisados distintamente para os problemas denominados *"comportados"* e *"não comportados".*

Tabela 12: Comparativo de Resultados com algoritmos da literatura

	Cruz [2010] e CLUES		Novo método Hierárquico	
Problema	**Índice Silhueta**	**K**	**Índice Silhueta**	**K**
Ruspini	0,737	4	0,738	6
Iris	0,686	3	0,687	2
Maronna	0,575	4	0,575	2
200data	0,823	3	0,823	5
1000p14c	0,831	14	0,831	15
1000p27c1	0,563	14	0,488	2
1000p5c1	0,639	5	0,634	3
1000p6c	0,736	6	0,736	8
100p10c	0,834	10	0,762	8
100p2c1	0,743	2	0,743	2
100p3c	0,786	3	0,786	3
100p3c1	0,597	3	0,574	2
100p5c1	0,703	6	0,688	6
100p7c	0,834	7	0,834	9
100p7c1	0,551	7	0,475	3
1100p6c1	0,685	6	0,650	3
1300p17c	0,823	17	0,823	20
1500p6c1	0,660	6	0,636	3
1800p22c	0,804	22	0,785	3
2000p11c	0,713	11	0,713	12
2000p9c1	0,623	9	0,616	14
200p12c1	0,575	13	0,564	9
200p2c1	0,764	2	0,764	2
200p3c1	0,680	3	0,676	3
200p4c	0,773	4	0,773	4

	Cruz [2010] e CLUES		Novo método Hierárquico	
Problema	**Índice Silhueta**	**K**	**Índice Silhueta**	**K**
200p4c1	0,754	4	0,745	6
200p7c1	0,576	8	0,555	2
300p13c1	0,594	9	0,546	2
300p2c1	0,776	2	0,778	2
300p3c	0,766	3	0,766	2
300p3c1	0,677	3	0,674	3
300p4c1	0,592	4	0,607	4
300p6c1	0,664	8	0,617	8
400p17c1	0,552	15	0,514	2
400p3c	0,799	3	0,799	2
400p4c1	0,620	4	0,605	2
500p3c	0,825	3	0,825	3
500p4c1	0,660	3	0,663	2
500p6c1	0,668	6	0,617	5
600p15c	0,781	15	0,775	13
600p3c1	0,721	3	0,721	2
700p15c1	0,680	15	0,642	21
700p4c	0,797	4	0,797	2
800p10c1	0,507	8	0,468	2
800p18c1	0,694	19	0,689	2
800p23c	0,787	23	0,787	2
800p4c1	0,714	4	0,696	2
900p12c	0,841	12	0,841	20
900p5c	0,716	5	0,707	6

Os resultados obtidos com os algoritmos propostos considerando apenas os problemas "comportados" foram superiores ou equivalentes em cerca de 80% dos casos. Porém, considerando apenas os problemas "não comportados" esse percentual cai para apenas 21,5%. Esses resultados sugeriram uma análise de o quão distante dos melhores resultados estão às soluções obtidos pelo método proposto.

Nesse novo experimento foram estabelecidas três classes relacionadas às distâncias entre a melhor solução da literatura e a solução obtida pelos algoritmos hierárquicos propostos. Em 91,8% dos experimentos, a diferença entre os resultados da literatura e dos resultados dos algoritmos propostos foram inferiores a 0,05. O percentual é de 95,9% se a diferença entre as Silhuetas for de até 0,075 e, em todos os experimentos, a diferença foi inferior a 0,1.

Com base nos casos em que os algoritmos propostos foram inferiores aos apresentados pela literatura, em 84,6% dos experimentos, a diferença entre os resultados foi inferior a 0,05. O percentual é de 92,3% se a diferença entre as Silhuetas for de até 0,075 e, em todos os experimentos, a diferença foi inferior a 0,1. Outros dados estatísticos concernentes ao conjunto DS2 indicam a proximidade entre os resultados obtidos por algoritmos sofisticados apresentados na literatura e os algoritmos propostos nesse trabalho. A maior diferença entre as Silhuetas da literatura e dos algoritmos propostos foi de 0,077, a menor de 0, a diferença média foi 0,015 e a mediana foi de apenas 0,005.

A Figura 5 apresenta ilustrações de soluções obtidas com a aplicação do BK4 no problema 100p3. Tais soluções possuem quantidade de grupos no intervalo [3,10] e seus Índices Silhueta estão relacionados abaixo de cada solução.

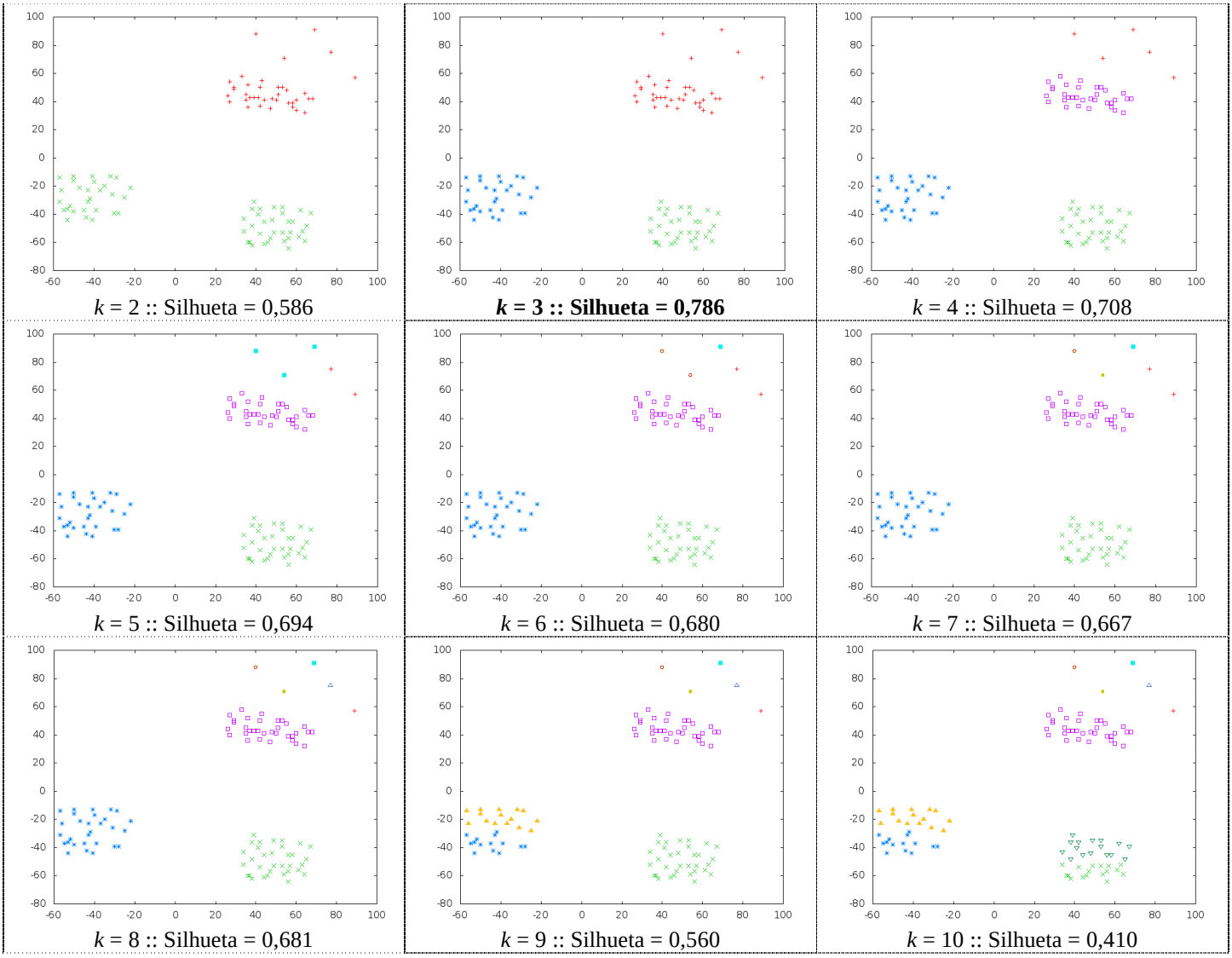


Figura 5: ilustrações das soluções obtidas pelo algoritmo BK4 para o problema 100p3c para k no intervalo $k = 2,..,10$.

5. Conclusões e Trabalhos Futuros

Com o objetivo de identificar o número ideal de grupos em cada problema, o método proposto neste trabalho consiste na aplicação de cinco versões de um novo método hierárquico divisivos baseado no conhecido algoritmo da literatura *Bisecting K-Means* [Steinbach et. al. (2000)]. Os algoritmos hierárquicos propostos possuem cinco diferentes critérios para seleção do grupo a ser dividido, quais sejam: o grupo com a *Menor Silhueta,* o grupo com o *Maior Diâmetro*, o grupo com o *Maior Clique*, o grupo com o *Maior Clique Médio* e o grupo com a *Maior Soma dos Erros Quadráticos*.

Em um primeiro experimento, foram realizados comparativos em relação aos valores do Índice Silhueta entre as cinco versões do novo método. Nesse experimento observou-se que a versão BK4 *(Maior Clique Médio* critério seleção do grupo a ser particionado) obteve os melhores resultados considerando a média, mediana, menor e maior valores do Índice Silhueta. Essa versão se destacou também por ter produzido o melhor resultado obtido em 89% dos problemas utilizados no presente trabalho. Além disso, para os problemas em que o BK4 não produziu o melhor resultado, a maior diferença entre as suas soluções e a as soluções da melhor versão foi de apenas 0,019 e a média das diferenças entre o melhor resultado e os resultados dessa versão foi inferior a 0,001.

Em relação aos comparativos realizados com o algoritmo CLUSTERING, os resultados produzidos pelas cinco versões do método proposto nesse trabalho foram superiores em todos os problemas. Nos comparativos com os melhores resultados entre os algoritmos SAPCA e AEC-RC [Soares, 2004], os algoritmos hierárquicos propostos foram superiores ou equivalentes em treze problemas. Além disso, os resultados foram inferiores em apenas três problemas, com maior diferença entre as Silhuetas de apenas 0,008. É importante ressaltar que para esses três problemas em que os algoritmos propostos não alcançaram os resultados da literatura, o número de grupos foi ou mesmo ou a diferença foi de apenas uma unidade.

A Tabela 12 apresentou comparativos entre os melhores resultados obtidos pelos algoritmos hierárquicos divisivos propostos em relação aos resultados obtidos por quatro algoritmos sofisticados da literatura CLUES [Wang et. al., 2007], AECBL1, GBLITRC1 e IBLITRC1 [Cruz, 2010]. Os resultados obtidos com os algoritmos propostos foram equivalentes ou superiores aos relatados na literatura em apenas 53% dos problemas. Dessa forma, com o objetivo de identificar as dificuldades do algoritmo bem como características dos problemas, os resultados foram analisados distintamente para os problemas considerados *"comportados"*, em que os resultados obtidos foram superiores ou equivalentes em 80,9% dos problemas, e *"não comportados"* cujo percentual cai para apenas 21,5%.

Embora para os problemas *"não comportados"* o percentual de soluções que alcançaram os melhores Índices Silhueta da literatura seja baixo, as diferenças entre as Silhuetas em relação aos valores *best* da literatura são pequenas. Em 84,6% dos experimentos a diferença entre as Silhuetas é inferior a 0,05, para diferenças até 0,075 o percentual é de 92,3% e, em todos os experimentos, a diferença é inferior a 0,1. Com a mesma analise apresentada considerando todos os resultados, em 91,8% dos experimentos a diferença entre as Silhuetas é inferior a 0,05, para diferenças até 0,075 o percentual é de 95,9% e, em todos os experimentos, a diferença é inferior a 0,1.

Os resultados apresentados neste estudo indicam que o método proposto é de fácil implementação e é competitivo em relação à qualidade das soluções quando comparado com os algoritmos mais sofisticados da literatura. A dificuldade do método em obter a quantidade de grupos em problemas considerados *"não comportados"* decorre, principalmente, da ausência de uma busca local para refinar a solução, realizando migrações, união ou divisão de grupos. Além disso, conforme já mencionado, o Índice Silhueta é mais apropriado para agrupamentos volumétricos, com grupos gerados de acordo com distribuições Gaussianas multidimensionais hiperesféricas ou moderadamente alongadas.

Como proposta de trabalhos futuros temos:

- Utilizar a versão simplificada da Silhueta proposta em [Hruschka et. al., 2004a] que reduz o custo computacional de ordem quadrática para ordem linear e que mantém a qualidade próxima ao da Silhueta tradicional [Vendramin et. al., 2009] [Vendramin et. al., 2010].
- Desenvolver heurísticas baseadas em metaheurísticas [Glover, 2003] considerando o método proposto nesse trabalho como uma heurística para a construção de soluções iniciais. Dessa forma, os procedimentos de busca local e as perturbações podem percorrer um novo espaço de busca para formação de novas soluções, que não seriam obtidas apenas com a utilização apenas dos algoritmos hierárquicos propostos.

Referências bibliográficas

Alves et. al. 2006] Alves, V., R. Campello, & E. Hruschka (2006). Towards a fast evolutionary algorithm for clustering. In IEEE Congress on Evolutionary Computation, 2006, Vancouver, Canada, pp. 1776–1783.

Brito et. al. 2009] Brito, J. A. M. ; Montenegro, F. M. T. ; Ochi, L. S (2009). Um Algoritmo ILS para a Melhoria de Eficiência da Estratificação Estatística. In: Simpósio Brasileiro de Pesquisa Operacional, Porto Seguro - BA, 2009.

Bandyopadhyay and Maulik, 2001] Bandyopadhyay, S. & U. Maulik (2001). Nonparametric genetic clustering: Comparison of validity indices. IEEE Transactions on Systems, Man and Cybernetics, Part C : Applications and Reviews. 31 (1), 120–125.

Bandyopadhyay and Maulik, 2002b] Bandyopadhyay, S. & U. Maulik (2002b). Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recognition 35, 1197–1208.

Baum, 1986] Baum, E.B. Iterated descent: A better algorithm for local search in combinatorial optimization problems. Technical report Caltech, Pasadena, CA. Manuscript, 1986.

Calinski and Harabasz, 1974] Calinski, R. B. & J. Harabasz (1974). A dendrite method for cluster analysis. Communications in Statistics 3.

Campello et. al., 2009] Campello, R. J. G. B., E. R. Hruschka, & V. S. Alves (2009). On the efficiency of evolutionary fuzzy clustering. Journal of Heuristics 15 (1), 43–75.

Cole, 1998] Cole, R. M. (1998). Clustering with genetic algorithms. MSc Dissertation, Department of Computer Science, University of Western Australia.

Cowgill, 1999] Cowgill, M. C., R. J. Harvey, & L. T. Watson (1999). A genetic algorithm approach to cluster analysis. Computational Mathematics and its Applications 37, 99–108.

Cruz, 2010] Cruz, M. D. O Problema de Clusterização Automática. Tese de Doutorado, UFRJ, Rio de Janeiro, 2010.

Dias and Ochi, 2003] Dias, C.R.; & Ochi, L.S.. Efficient Evolutionary Algorithms for the Clustering Problems in Directed Graphs. Proc. of the IEEE Congress on Evolutionary Computation (IEEE-CEC), 983-988. Canberra, Austrália, 2003.

Ester et al., 1995] Ester, M., Kriegel, H.-P., and Xu, X., A Database Interface for Clustering in Large Spatial Databases, In: Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining (KDD-95), pp. 94- 99, Montreal, Canada, August, 1995.

Ester et. al., 1996] Ester, M., H.-P. Kriegel, J. Sander, & X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), pp. 226–231.

Fisher, 1936] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annual Eugenics 7,pp. 179-188.

Glover, 2003] Glover, F.. Handbook of Metaheuristics. Kluwer Academic Publishers, 2003.

Goldschimidt and Passos, 2005] Goldschimidt R.; Passos, E. Data Mining: um guia prático. Editora Campus, Rio de Janeiro: Elsevier, 2005.

Han and Kamber, 2006] Han, J., e Kamber, M., Cluster Analysis. In: Morgan Kaufmann. Publishers (eds.), Data Mining: Concepts and Techniques, 2 ed., chapter 8, New York, USA, Academic Press, 2006.

Handl and Knowles, 2007] Handl, J. & J. Knowles (2007). An evolutionary approach to multiobjective clustering.IEEE Trans. on Evolutionary Computation 34, 56–76.

Hastie et. al., 2001] Hastie, t.;Tibshirani, R.; Friedman, J. (2001). The Elements of Statistical Learning. Data Mining, Inference, and prediction. Springer.

Hruschka and Ebecken, 2001] Hruschka, E. R., Ebecken, N. F. F. A Genetic algorithm for cluster analysis. IEEE Transactions on Evolutionary Computation , 2001.

Hruschka and Ebecken, 2003] Hruschka, E. R. & Ebecken, N. F. F. (2003). A genetic algorithm for cluster analysis. Intelligent Data Analysis 7 (1), 15–25.

Hruschka et. al., 2004a] Hruschka, E. R., R. J. G. B. Campello, & L. N. de Castro (2004a). Evolutionary algorithms for clustering gene-expression data. In Proc. IEEE Int. Conf. on Data Mining, Brighton/England, pp. 403–406.

Hruschka et. al., 2004b] Hruschka, E. R., R. J. G. B. Campello, & L. N. de Castro (2004b). Improving the efficiency of a clustering genetic algorithm. In Advances in Artificial Intelligence - IBERAMIA 2004: 9th Ibero-American Conference on AI, Puebla, Mexico, November 22-25. Proceedings, Volume 3315, pp. 861–870. Springer-Verlag GmbH, Lecture Notes in Computer Science.

Hruschka et. al., 2006] Hruschka, E. R., R. J. G. B. Campello, & L. N. de Castro (2006). Evolving clusters in gene-expression data. Information Sciences 176 (13), 1898–1927.

Jain and Dubes, 1988] Jain, A. & R. Dubes (1988). Algorithms for Clustering Data. Prentice Hall.

Jr, 1968] Jr, H. S. (1968). Cardinality of finite topologies. Journal of Combinatorial Theory 5 (1), 82–86.

Kumar et. al., 2009] Kumar, V. ; Steinbach, M. ; Tan, P. N. Introdução ao Data Mining - Mineração De Dados. Ciência Moderna, 2009.

Larose, 2005] Larose, D. T. Discovering Knowledge in Data, An Introduction to Data Mining. John Wiley & Sons, 2005.

Liu, 1968] Liu, G. (1968). Introduction to Combinatorial Mathematics. McGraw Hill.

Ma et. al., 2006] Ma, P. C. H., K. C. C. Chan, X. Yao, & D. K. Y. Chiu (2006). An evolutionary clustering algorithm for gene expression microarray data analysis. IEEE Trans. Evolutionary Computations 10 (3), 296–314.

Maronna and Jacovkis, 1974] Maronna, R.; Jacovkis, P. M. (1974). Multivariate clustering procedures with variable metrics. Biometrics30, pp. 499-505.

Matloff 2011] Matloff, N. The Art of R Programming: A Tour of Statistical Software Design. No Starch. Press, 2011.

Naldi and Carvalho, 2007] Naldi, M. C. & A. C. P. L. F. Carvalho (2007). Clustering using genetic algorithm combining validation criteria. In Proceedings of the 15th European Symposium on Artificial Neural Networks, ESANN 2007, Volume 1, pp. 139–144. Evere.

Naldi, 2011] Naldi, C. N. Técnicas de Combinação para Agrupamento Centralizado e Distribuído de Dados. Tese de Doutorado, USP - São Carlos, 2011.

Pakhira et. al., 2005] Pakhira, M., S. Bandyopadhyay, & U. Maulik (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. Fuzzy Sets Systems 155 (2), 191–214.

Pal and Bezdek, 1995] Pal, N. & J. Bezdek (1995). On cluster validity for the fuzzy c-means model. IEEE Transactions of Fuzzy Systems 3 (3), 370–379.

Pan and Cheng, 2007] Pan, S. & K. Cheng (2007). Evolution-based tabu search approach to automatic clustering. IEEE Transactions on Systems, Man, and Cybernetics, Part C - Applications and Reviews 37 (5), 827–838.

Pelleg and Moore, 2000] Pelleg, D. & A. Moore (2000). X-means: extending k-means with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727–734.

Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65.

Ruspini, 1970] Ruspini, E. H. (1970). Numerical methods for fuzzy clustering. Information Science.pp. pp. 319-350.

Soares and Ochi, 2004] Soares, S. S. R. F., Ochi, L. S. Um Algoritmo Evolutivo com Reconexão de Caminhos para o Problema de Clusterização Automática. in XII Latin Ibero American Congress on Operations Research, 2004, Havana. Proc. of the XII CLAIO (em CD-ROM). ALIO, 2004. v.1, p. 7 -13.

Soares, 2004] Soares, A. S. R. F. Metaheurísticas para o Problema de Clusterização Automática, Dissertação de Mestrado, UFF - Niterói, 2004.

Steinbach et. al., 2000] Steinbach, M., G. Karypis, & V. Kumar (2000). A comparison of document clustering techniques. Technical Report 34, University of Minnesota.

Tseng and Yang, 2001] Tseng, L. & . Yang, S.B. (2001). A genetic approach to the automatic clustering problem. Pattern Recognition 34, 415–424.

Vendramin et. al., 2009] Vendramin, L., R. J. G. B. Campello, & E. R. Hruschka (2009). *On the comparison of relative* clustering validity criteria. In SIAM International Conference on Data Mining, Sparks/USA, pp. 733–744.

Vendramin et. al., 2010] Vendramin, L., R. J. G. B. Campello, & E. R. Hruschka (2010). Relative clustering validity criteria: A comparative overview. Statistical Analysis and Data Mining 3 (4), 209–235.

Wang et. al., 2007] Wang, X., Qiu, W., Zamar, R. H. (2007). CLUES: A non-parametric clustering method based on local shrinking. Computational Statistics & Data Analysis 52, pp. 286-298.

Zalik, 2008] An Efficient K'-Means Clustering Algorithm, Pattern Recognition Letters 29, 2008.

Agradecimentos

Os autores agradecem ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

Abstract

The Cluster Analysis has several classification methods that can be applied in order to identify groups within a data set. The number of groups either may be fixed or must be identified upon review of some index (or coefficient). This way, the solutions quality is indicated by the silhouette coefficient, which combines cohesion and separation. This paper presents a new hierarchical clustering method based on Bisecting K-Means Algorithm, aiming to identify the ideal number of groups. The results indicate that the proposed method is easier to implement and is competitive in the quality of solutions when compared with efficient algorithms of the literature.

Keywords: Automatic Clustering Problem, Hierarchical Clustering, Silhouette Index.

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.

Se o assunto é **Brasil**,
procure o **IBGE**.

www.ibge.gov.br
www.twitter.com/ibgecomunica
www.facebook.com/ibgeoficial

Atendimento
0800 721 8181

ISSN 0034-7175